Reciprocity relations and forward amplitude theorems for elastic waves

V. Varatharajulu

Department of Theoretical and Applied Mechanics, Cornell University, Ithaca, New York 14853
(Received 13 May 1976)

Relations are derived between the amplitudes of scattered elastic waves from obstacles of arbitrary shape when the direction of observation and incidence are reversed and interchanged, similar to the reciprocity theorem for the scalar wave equation. For waves which undergo a polarization change on scattering (longitudinal to shear and vice versa), a new type of reciprocity relation is derived which is peculiar to elastic waves. Expressions for the scattering cross section are obtained for shear and longitudinal incident waves, and it is shown to be proportional to the amplitude of the scattered wave in the forward direction alone. Incident waves are restricted to plane, monochromatic waves.

1. INTRODUCTION

The forward scattering theorem more popularly known as the optical theorem was first derived by Feenberg[1] for atomic collisions in quantum mechanics. The theorem relates the total cross section (absorption plus elastic) to the imaginary part (real part in 2-D) of the asymptotic or far field amplitude of the scattered field in the forward direction (by forward we mean parallel to the incident beam). The theorem is so called because of the analogy of light propagating through an absorbing medium when the imaginary part of the complex index of refraction can be related to the total extinction coefficient. An analogous theorem for the scattering of electromagnetic waves by obstacles of arbitrary shape was derived by Van de Hulst[2] and put on a more rigorous mathematical foundation by Jones.[3] In quantum mechanics the optical theorem was generalized by Glauber and Schomaker[4] to yield the generalized optical theorem and the reciprocity theorem.

The reciprocity theorem is related to symmetry under the interchange of source and receiver in the presence of an obstacle. The reciprocity relation is a statement of the equivalence of the scattered field amplitudes for the two situations. For scattering in a central field (hence possessing inversion symmetry), the reciprocity relation is an identity since in this case the amplitude of the scattered field only depends on the angle between the incident and scattered directions.

In this paper, the forward scattering theorem and the reciprocity theorem are extended to include the scattering of elastic waves by obstacles of arbitrary shape. Elastic waves undergo mode conversion on scattering, i.e., an incident longitudinal wave will in general result in both longitudinal (compressional) and transverse (shear) waves on scattering and vice versa. An interesting consequence of reciprocity is a relation between the amplitudes of waves that have undergone mode conversion. It is shown in the following sections that reciprocity relations are obeyed for the scattering of compressional waves to compressional waves and shear waves to shear waves. But for mode conversion processes, it is shown that the far field amplitude of scattered shear waves due to incident compressional waves is directly proportional to the far field amplitude of scattered compressional waves due to incident shear waves with all directions reversed. From this result, valid in the far field, one can then obtain an interesting relation between the expansion coefficients of the two waves for the case of a circular cylinder. This will eliminate separate calculations for incident compressional and incident shear waves. More importantly, the reciprocity relation can be used to check numerical calculations for obstacles of arbitrary shape since these do not exist in the literature.

The forward scattering theorem for elastic waves states that the total rate at which energy is transmitted across a closed surface by both scattered shear waves and compressional waves is directly proportional to the real part (imaginary part in 2-D) of the far field scattered amplitude of compressional waves for incident compressional waves and of shear waves for incident shear waves in the forward direction.

In Sec. 2 the asymptotic form of the displacement and stress due to scattered waves are presented in both two and three dimension. In Sec. 3 reciprocity relations are derived for various types of scattering processes. In Sec. 4 the forward scattering theorem is derived for both shear and compressional wave incidence. (All derivations are given for both two and three dimensions.) The integrals are evaluated using the method of stationary phase.

2. ASYMPTOTIC FORMS OF THE DISPLACEMENT AND STRESS DUE TO SCATTERED WAVES

Consider an infinite, homogeneous, linear isotropic elastic medium of density ρ and Lamé constants λ and μ in which an inhomogeneity bounded by the closed surface S is embedded. The surface S can be of arbitrary shape with a continuous turning normal. It may enclose an elastic inclusion of different material properties, an inviscid fluid or a cavity. Let a plane monochromatic wave of frequency ω be incident on S. Thus all field quantities

 Copyright © 1977 American Institute of Physics

will have a time dependence exp $(-i\omega t)$ which will be suppressed for notational convenience.

If $\mathbf{u}$ is the displacement vector outside S, then

$$\nabla\cdot\mathbf{T} + \rho\omega^2\mathbf{u} = 0, \tag{2.1}$$

where $\mathbf{T}$ is the stress tensor related to $\mathbf{u}$ by Hooke's law,

$$\mathbf{T} = \lambda\mathbf{I}\,\nabla\cdot\mathbf{u} + \mu(\nabla\mathbf{u} + \mathbf{u}\nabla), \tag{2.2}$$

$\mathbf{I}$ being the idemfactor. The right-hand side of (2.1) is zero since $\mathbf{u}$ has no sources outside S, the sources of the incident plane wave being at infinity. $\mathbf{u}$ can be written as

$$\mathbf{u} = \mathbf{u}^i + \mathbf{u}^s, \tag{2.3}$$

where $\mathbf{u}^i$ is the displacement due to the incident waves and $\mathbf{u}^s$ is due to the scattered waves.

At distances large compared to the dimensions of the obstacle, physical considerations require that $\mathbf{u}^s$ represent outgoing waves. In 3-D, $\mathbf{u}^s$ will consist of outgoing spherical waves and in 2-D, outgoing cylindrical waves. Given below are the asymptotic forms of the elastic field for 2-D and 3-D.

A. 2-D

Problems in 2-D also include scattering from infinitely long cylinders of arbitrary cross section when the incident plane wave propagates perpendicular to the axis of the cylinder; then the displacement vector will be independent of the coordinate parallel to the axis of the cylinder and the problem is reduced effectively to two dimensions. If the incident plane wave is a shear wave with polarization vector parallel to the axis of the cylinder, then the scattered field also consists of shear waves with polarization vector parallel to the axis; the displacement equation (2.1) in this case reduces to the Helmholtz equation for which discussions of the reciprocity theorem and forward scattering theorem abound in the literature (see, for example, Jones,[5] Secs. 1.32, 8.6, 8.24). Here, we confine ourselves to incident longitudinal waves and shear waves polarized in a plane perpendicular to the axis of the cylinder. In elasticity, this is known as plane strain and the scattered field is also comprised of waves with polarization and propagation vectors perpendicular to the axis of the cylinder.

Let the Z axis be perpendicular to the plane containing the boundary S of the scatterer; for a cylinder the Z axis will coincide with the cylinder axis. It is convenient to expand the scalar and vector potentials of the scattered displacement vector in terms of the cylinder eigenfunctions, namely the Bessel functions. The scalar potential ϕ is associated with longitudinal waves and the vector potential ψ with shear waves, and they satisfy the Helmholtz equation. This is known as the Helmholtz decompontion of a vector field; see for example Achenbach[6] and Morse and Feshbach.[7] Let $k_p = \omega(\rho/\lambda+2\mu)^{1/2}$, and $k_s = \omega(\rho/\mu)^{1/2}$ be the wavenumbers for longitudinal waves and shear waves respectively; then

$$\phi(\mathbf{r}) = \sum_{n=0}^{\infty}\varepsilon_n^{1/2}(A_n\cos n\theta + B_n\sin n\theta)H_n(k_p r) \tag{2.4}$$

and

$$\psi(\mathbf{r}) = \sum_{n=0}^{\infty}\varepsilon_n^{1/2}(C_n\cos n\theta + D_n\sin n\theta)H_n(k_s r), \tag{2.5}$$

where $H_n(kr)$ are Hankel functions of the first kind, A_n, B_n, C_n, D_n are unknown constants to be determined from the boundary conditions, and ε_n is the Neumann factor, $\varepsilon_0 = 1$ and $\varepsilon_n = 2$ for $n \neq 0$. r, θ are polar coordinates in a plane perpendicular to the Z axis. The displacement vector constructed from (2.4) and (2.5) is given by

$$\mathbf{u}^s(\mathbf{r}) = \nabla\phi(\hat{\mathbf{r}}) + \nabla\times(\hat{\mathbf{z}}\psi(\hat{\mathbf{r}})). \tag{2.6}$$

At large distances from S (the origin of coordinates being centered in S), we may use the asymptotic form of the Hankel functions in (2.4) and (2.5), substitute in (2.6) and obtain for the far displacement field

$$\mathbf{u}^s(\mathbf{r}) \xrightarrow[r\to\infty]{} \hat{\mathbf{r}}f_p(\hat{\mathbf{r}})\left(\frac{2}{i\pi k_p r}\right)^{1/2}e^{ik_p r} + \hat{\boldsymbol{\theta}}f_s(\hat{\mathbf{r}})\left(\frac{2}{i\pi k_s r}\right)^{1/2}e^{ik_s r}. \tag{2.7}$$

In (2.7), $f_p(\hat{\mathbf{r}})$ and $f_s(\hat{\mathbf{r}})$ are angular distributions of the amplitudes of the scattered compressional and shear waves respectively, and they contain all the angular dependence indicated by the unit vector $\hat{\mathbf{r}}$. As a result of expanding the potentials with Hankel functions of the first kind, the far displacement field is comprised of outgoing cyclindrical compressional waves polarized along $\hat{\mathbf{r}}$ and outgoing cylindrical shear waves polarized along $\hat{\boldsymbol{\theta}}$, $(\hat{\mathbf{r}}\cdot\hat{\boldsymbol{\theta}} = 0)$. From (2.4)–(2.7), the definitions of f_p and f_s are

$$f_p(\hat{\mathbf{r}}) = k_p\sum_n\varepsilon_n^{1/2}(A_n\cos n\theta + B_n\sin n\theta)i^{-(n-1)} \tag{2.8}$$

and

$$f_s(\hat{\mathbf{r}}) = -k_s\sum_n\varepsilon_n^{1/2}(C_n\cos n\theta + D_n\sin n\theta)i^{-(n-1)}; \tag{2.9}$$

by using (2.2) and (2.7), the far field stress tensor is given by

$$\begin{aligned}\mathbf{T}(\mathbf{r}) \xrightarrow[r\to\infty]{} & ik_p(\lambda\mathbf{I} + 2\mu\hat{\mathbf{r}}\hat{\mathbf{r}})f_p(\hat{\mathbf{r}})\left(\frac{2}{i\pi k_p r}\right)^{1/2}e^{ik_p r} \\ & + ik_s\mu(\hat{\mathbf{r}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}\hat{\mathbf{r}})f_s(\hat{\mathbf{r}})\left(\frac{2}{i\pi k_s r}\right)^{1/2}e^{ik_s r}.\end{aligned} \tag{2.10}$$

B. 3-D

This situation is complicated by the fact that in general for a given P (compressional) or S (shear) wave, the scatteredfield comprises of both P waves and S waves, but the polarization of the scattered S wave is undefined. It is convenient to decompose the scattered shear wave in two mutually perpendicular directions denoted by s_1 and s_2. The potentials in this case are expanded in spherical eigenfunctions, the radial part being described by spherical Bessel functions and the angular part by spherical harmonics. There are three potentials ϕ, ψ, x associated with the P wave and the two components of the S wave respectively, and they are given by

$$\phi(\mathbf{r}) = \sum_{n=0}^{\infty}\sum_{m=0}^{n}(\gamma_{n,m})^{1/2}[A_n Y^e_{n,m}(\theta,\phi) + B_n Y^o_{n,m}(\theta,\phi)]\, h_n(k_p r), \tag{2.11}$$

$$\psi(\mathbf{r}) = \sum_{n=0}^{\infty}\sum_{m=0}^{n}(\gamma_{n,m})^{1/2}[C_n Y^e_{n,m}(\theta,\phi) + D_n Y^o_{n,m}(\theta,\phi)]\, h_n(k_s r), \tag{2.12}$$

$$\chi(\mathbf{r}) = \sum_{n=0}^{\infty}\sum_{m=0}^{n}(\gamma_{n,m})^{1/2}[E_n Y^e_{n,m}(\theta,\phi) + F_n Y^o_{n,m}(\theta,\phi)]\, h_n(k_s r), \tag{2.13}$$

where (r, θ, ϕ) are spherical polar coordinates with origin in S, $Y^{e,o}_{n,m}(\theta,\phi)$ are the even and odd spherical harmonics, the $h_n(kr)$ are spherical Bessel functions, and the normalizing constant $\gamma_{n,m}$ is given by

$$\gamma_{n,m} = \varepsilon_n(2n+1)(n-m)!/4\pi(n+m)!. \tag{2.14}$$

The displacement vector is constructed from ϕ, ψ, and χ by (see Ref. 6 or 7)

$$\mathbf{u}^s(\mathbf{r}) = \nabla\phi(\mathbf{r}) + k_s\nabla\times[\mathbf{r}\psi(\mathbf{r})] + \nabla\times\nabla[\mathbf{r}\times\chi(\mathbf{r})]. \tag{2.15}$$

Again using the asymptotic form of the spherical Hankel functions for large values of $\mathbf{r}$, we obtain from (2.11)–(2.15)

$$\mathbf{u}^s(\mathbf{r})_r = \hat{\mathbf{r}} f_p(\hat{\mathbf{r}}) e^{ik_p r}/r + [\hat{\boldsymbol{\theta}} f_{s_1}(\hat{\mathbf{r}}) + \hat{\boldsymbol{\phi}} f_{s_2}(\hat{\mathbf{r}})] e^{ik_s r}/r, \tag{2.16}$$

where f_p is the angular distribution of the amplitude of the scattered P wave and f_{s_1} and f_{s_2} are components of the amplitude of the S wave along two mutually perpendicular vectors $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$. They are defined by

$$f_p(\hat{\mathbf{r}}) = \sum_{n=0}^{\infty}\sum_{m=0}^{n}(\gamma_{n,m})^{1/2} i^{-n}[A_n Y^e_{n,m}(\theta,\phi) + B_n Y^o_{n,m}(\theta,\phi)],$$

$$f_{S_1}(\hat{\mathbf{r}}) = \sum_{n=0}^{\infty}\sum_{m=0}^{n}(\gamma_{m,n})^{1/2} \times\left[\frac{i^{-(n+1)}}{\sin\theta}\left(C_n\frac{\partial}{\partial\phi}Y^e_{n,m}(\theta,\phi) + D_n\frac{\partial}{\partial\phi}Y^o_{n,m}(\theta,\phi)\right) + i^{-n}\left(E_n\frac{\partial}{\partial\theta}Y^e_{n,m}(\theta,\phi) + F_n\frac{\partial}{\partial\phi}Y^o_{n,m}(\theta,\phi)\right)\right] \tag{2.17, 2.18}$$

and

$$f_{s_2}(\hat{\mathbf{r}}) = \sum_{n=0}^{\infty}\sum_{m=0}^{n}(\gamma_{m,n})^{1/2} \times\left[-i^{-(n+1)}\left(C_n\frac{\partial}{\partial\theta}Y^e_{n,m}(\theta,\phi) + D_n\frac{\partial}{\partial\theta}Y^o_{n,m}(\theta,\phi)\right) + \frac{i^{-n}}{\sin\theta}\left(E_n\frac{\partial}{\partial\phi}Y^e_{n,m}(\theta,\phi) + F_n\frac{\partial}{\partial\phi}Y^o_{n,m}(\theta,\phi)\right)\right]. \tag{2.19}$$

The stress corresponding to (2.16) may be written as

$$\mathbf{T}^s(\mathbf{r}) \xrightarrow[r\to\infty]{} ik_p[\lambda\mathbf{I} + 2\mu\hat{\mathbf{r}}\hat{\mathbf{r}}] f_p(\hat{\mathbf{r}}) e^{ik_p r}/r + ik_s\mu[(\hat{\mathbf{r}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}\hat{\mathbf{r}}) f_{s_1}(\hat{\mathbf{r}}) + (\hat{\mathbf{r}}\hat{\boldsymbol{\phi}} + \hat{\boldsymbol{\phi}}\hat{\mathbf{r}}) f_{s_2}(\hat{\mathbf{r}})] e^{ik_s r}/r. \tag{2.20}$$

Expressions for incident plane waves are the same in 2-D and 3-D, the vector $\mathbf{r}$ being the cylindrical polar coordinate in 2-D and the spherical polar coordinate in 3-D. We denote by the unit vector $\hat{\mathbf{u}}$ the polarization of an incident shear wave. In 2-D since we are only concerned with problems of plane strain $\hat{\mathbf{u}} = (\hat{\mathbf{k}}\times\hat{\mathbf{z}})$, where $\hat{\mathbf{k}}$ is a unit vector along the direction of propagation and $\hat{\mathbf{z}}$ is a unit vector along the Z axis. Thus, the displacement and stress due to an incident compressional wave are given by

$$\mathbf{u}^i_p(\mathbf{r}) = ik_p\hat{\mathbf{k}} e^{ik_p\hat{\mathbf{k}}\cdot\mathbf{r}}, \tag{2.21}$$

$$\mathbf{T}^i_p(\mathbf{r}) = -k_p^2(\lambda\mathbf{I} + 2\mu\hat{\mathbf{r}}\hat{\mathbf{r}})\, e^{ik_p\hat{\mathbf{k}}\cdot\mathbf{r}}, \tag{2.22}$$

and that due to an incident shear wave are given by

$$\mathbf{u}^i_s(\mathbf{r}) = ik_s\hat{\mathbf{u}}\, e^{ik_s\hat{\mathbf{k}}\cdot\mathbf{r}} \tag{2.23}$$

and

$$\mathbf{T}^i_s(\mathbf{r}) = -k_s^2(\hat{\mathbf{k}}\hat{\mathbf{u}} + \hat{\mathbf{u}}\hat{\mathbf{k}})\, e^{ik_s\hat{\mathbf{k}}\cdot\mathbf{r}}. \tag{2.24}$$

3. RECIPROCITY RELATIONS

Let $\mathbf{u}(\hat{\mathbf{k}}, \mathbf{r})$ be the total displacement field at $\mathbf{r}$ due to a wave incident along $\hat{\mathbf{k}}$ and $\mathbf{u}(-\hat{\mathbf{k}}', \mathbf{r})$ that due to a wave incident along $-\hat{\mathbf{k}}'$. Both $\mathbf{u}(\hat{\mathbf{k}}, \mathbf{r})$ and $\mathbf{u}(-\hat{\mathbf{k}}', \mathbf{r})$ satisfy the displacement equation (2.1) and the same boundary conditions at S, the boundary of the obstacle. Thus,

$$\mathbf{u}(\hat{\mathbf{k}}, \mathbf{r})\cdot[\nabla\cdot\mathbf{T}(-\hat{\mathbf{k}}', \mathbf{r}) + \rho\omega^2\mathbf{u}(-\hat{\mathbf{k}}', \mathbf{r})] - [\nabla\cdot\mathbf{T}(\hat{\mathbf{k}}, \mathbf{r}) + \rho\omega^2\mathbf{u}(\hat{\mathbf{k}}, \mathbf{r})]\cdot\mathbf{u}(-\hat{\mathbf{k}}', \mathbf{r}) = 0. \tag{3.1}$$

From generalized Hooke's law

$$T_{ij}(\mathbf{r}) = \Gamma_{ijkl}\partial u_l(\mathbf{r})/\partial x_k, \tag{3.2}$$

where Γ_{ijkl} is the elastic stiffness tensor that has the following properties:

$$\Gamma_{ijkl} = \Gamma_{klij} = \Gamma_{jikl}. \tag{3.3}$$

Using (3.3), (3.2) in (3.1), we obtain

$$\nabla\cdot[\mathbf{u}(\hat{\mathbf{k}}, \mathbf{r})\cdot\mathbf{T}(-\hat{\mathbf{k}}', \mathbf{r}) - \mathbf{T}(\hat{\mathbf{k}}, \mathbf{r})\cdot\mathbf{u}(-\hat{\mathbf{k}}', \mathbf{r})] = 0. \tag{3.4}$$

Applying the divergence theorem to (3.4) yields

$$\int_{S'} - \int_{S}\hat{n}(\mathbf{r})\cdot[\mathbf{u}(\hat{\mathbf{k}}, \mathbf{r})\cdot\mathbf{T}(-\hat{\mathbf{k}}', \mathbf{r}) - \mathbf{T}(\hat{\mathbf{k}}, \mathbf{r})\cdot\mathbf{u}(-\hat{\mathbf{k}}', \mathbf{r})]dS(\mathbf{r}) = 0, \tag{3.5}$$

were S' is any closed surface that encloses S.

The integral on S is identically zero due to the boundary conditions satisfied by the elastic field variables $\mathbf{u}$ and $\mathbf{T}$. For, if S encloses an elastic obstacle, then $\mathbf{u}$ and $\mathbf{T}$ must be continuous across S; if S is a cavity, then $\mathbf{T}$ vanishes on S. In either case the integral on S vanishes. For a general discussion of this point refer to Jones, Refs. 5, Sec. 1.32.

For convenience we taken S' to be a circle in 2-D (a sphere in 3-D) of radius large compared to the dimensions of S. Thus, the unit normal $\hat{n}(\mathbf{r})$ on S' is simply $\hat{\mathbf{r}}$, and use the asymptotic form of the elastic field variables derived in Sec. 2 in the integrand of (3.5). We consider below the two- and three-dimensional cases separately.

A. 2-D

P–P incidence

First consider the case when the incident waves along $-\hat{\mathbf{k}}'$ and $\hat{\mathbf{k}}$ are both P waves of equal amplitude and frequency. Substituting (2.7), (2.10), (2.21), (2.22), and (2.3) in (3.5), we obtain

$$
\begin{aligned}
&- ik_p^3 r\int_0^{2\pi}[\lambda(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}) + 2\mu(\hat{\mathbf{k}}\cdot\hat{\mathbf{k}}')(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}}) + \lambda(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}}) \\
&\quad + 2\mu(\hat{\mathbf{k}}'\cdot\hat{\mathbf{k}})\,(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})]\; e^{ik_p\hat{\mathbf{r}}\cdot(\hat{\mathbf{k}}-\hat{\mathbf{k}}')\cdot\mathbf{r}}d\theta - k_p^2 r\left(\frac{2}{i\pi k_p r}\right)^{1/2} e^{ik_p r} \\
&\times\int_0^{2\pi}[f_{pp}(-\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}})\{(\lambda+2\mu)(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}) - \lambda - 2\mu(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})^2\}e^{ik_p\hat{\mathbf{k}}\cdot\mathbf{r}} \\
&\quad + f_{pp}(\hat{\mathbf{k}},\hat{\mathbf{r}})\{\lambda + 2\mu(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}})^2 + (\lambda+2\mu)(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}})\}e^{-ik_p\hat{\mathbf{k}}'\cdot\mathbf{r}}]d\theta \\
&\quad - \mu k_p\left(\frac{2}{i\pi k_s r}\right)^{1/2} e^{ik_s r}\int_0^{2\pi}[f_{sp}(-\hat{\mathbf{k}}',\hat{\mathbf{r}})\}k_s(\hat{\mathbf{k}}\cdot\hat{\boldsymbol{\theta}}) \\
&\quad - 2k_p\mu(\hat{\mathbf{k}}\cdot\hat{\boldsymbol{\theta}})(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})\}e^{ik_p\hat{\mathbf{k}}\cdot\mathbf{r}} + f_{sp}(\hat{\mathbf{k}},\hat{\mathbf{r}})[2\hat{\mathbf{k}}_p(\hat{\mathbf{k}}'\cdot\hat{\boldsymbol{\theta}})(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}}) \\
&\quad + \mathbf{k}_s(\hat{\mathbf{k}}'\cdot\hat{\boldsymbol{\theta}})]e^{-ik_s\hat{\mathbf{k}}'\cdot\mathbf{r}}]d\theta + O\left(\frac{1}{r^{1/2}}\right) = 0. \qquad (3.6)
\end{aligned}
$$

The notation used to describe the various types of scattered wave amplitudes is as follows. In $f_{A,B}(\hat{\mathbf{i}},\hat{\mathbf{J}})$, the second subscript denotes the polarization of the incident wave P or S, the first unit vector $\hat{\mathbf{i}}$ in the argument denotes the direction of incidence. The first subscript of f denotes the polarization of the scattered wave and the second unit vector $\hat{\mathbf{J}}$ denotes the direction of propagation of the scattered wave. The two subscripts on the scattered wave amplitudes indicate the polarization of the incident wave and that of the scattered wave, respectively.

In (3.6), the first integral is identically zero since for every pair of angles $(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})$ and $(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}})$ there is a corresponding pair at $-(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})$, $-(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}})$ so that the contribution from the two pairs cancel exactly. For the remaining integrals in (3.6), we note that since $r\to\infty$ on S', the phase of the exponential function in the inegrand osicllates very rapidly for finite k_p as we integrate θ from 0 to 2π, so that the dominant contribution is from the vicinity of those points where the phase is stationary, i.e., an extremum. Thus the integrals in (3.6) can be evaluated to a very good approximation by the method of stationary phase (see for example Copson[8] or Erdelyi[9]).

Let

$$I = \int_\alpha^B \phi(x)e^{i\nu f(x)}dx. \qquad (3.7)$$

Then, if $\nu f(x)$ and $\phi(x)$ are analyiic functions and if $f(x)$ has points of stationary phase at α, β and $\alpha < \gamma_i < \beta$, where $f'(x) = 0$, then

$$
\begin{aligned}
I &= \left(\frac{\pi}{2\nu|f''(\alpha)|}\right)^{1/2}\phi(\alpha)e^{i\nu f(\alpha)\pm i\pi/4} \\
&\quad + \left(\frac{\pi}{2\nu|f''(\beta)|}\right)^{1/2}\phi(\beta)e^{i\nu f(\beta)\pm i\pi/4} \\
&\quad + \sum_i\left(\frac{2\pi}{\nu|f''(\gamma_i)|}\right)^{1/2}\phi(\gamma_i)e^{if(\gamma_i)\pm i\pi/4} + 0\left(\frac{1}{\nu}\right). \qquad (3.8)
\end{aligned}
$$

In (3.8), the primes on f denote differentiation with respect to the argument, the $\pm$ signs should be used according to $f''(x) \gtrless 0$ at the point of stationary phase.

It is seen that in the first term of the third integral, the points of stationary phase are at $\cos^{-1}(\hat{\mathbf{r}}\cdot\hat{\mathbf{k}}) = 0, \pi, 2\pi$ and those of the second term at $\cos^{-1}(\hat{\mathbf{r}}\cdot\hat{\mathbf{k}}') = 0, \pi, 2\pi$. However, the integrand of the third integral vanishes at all these points since $\hat{\mathbf{k}}\cdot\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{k}}'\cdot\hat{\boldsymbol{\theta}}$ are zero. So the contribution of the third integral is $O(1/\sqrt{r})$. The second integral also has points of staitonary phase at $\cos^{-1}(\hat{\mathbf{r}}\cdot\hat{\mathbf{k}}) = 0, \pi, 2\pi$ and $\cos^{-1}(\hat{\mathbf{r}}\cdot\hat{\mathbf{k}}') = 0, \pi, 2\pi$ which give a contribution proportional to $[f_{pp}(-\hat{\mathbf{k}}', -\hat{\mathbf{k}}) - f_{pp}(\hat{\mathbf{k}},\hat{\mathbf{k}}')$, so that in the limit of $r\to\infty$

$$f_{pp}(-\hat{\mathbf{k}}', -\hat{\mathbf{k}}) = f_{pp}(\hat{\mathbf{k}},\hat{\mathbf{k}}'). \qquad (3.9)$$

S–S incidence

If the incident waves propagating along $\hat{\mathbf{k}}$ and $-\hat{\mathbf{k}}'$ are both shear waves of equal amplitude and frequency polarized along $\hat{\mathbf{u}}$ and $-\hat{\mathbf{u}}'$ respectively, then, substituting (2.3), (2.7), (2.9), (2.23), and (2.24) in (3.5) and evaluating all the integrals by the method of stationary phase, we obtain

$$(\hat{\mathbf{u}}\cdot\hat{\boldsymbol{\theta}}_{-k})f_{ss}(-\hat{\mathbf{k}}', -\hat{\mathbf{k}}) = -(\hat{\mathbf{u}}'\cdot\hat{\boldsymbol{\theta}}_{k'})f_{ss}(\hat{\mathbf{k}},\hat{\mathbf{k}}'). \qquad (3.10)$$

In (3.10), the notation $\hat{\boldsymbol{\theta}}_{-k}$ is used to indicate the unit vector perpendicular to $-\hat{\mathbf{k}}$ in the increasing direction of the polar angle θ. $\hat{\boldsymbol{\theta}}_{k'}$, is the corresponding vector orthogonal to $\hat{\mathbf{k}}'$.

S–P or P–S incidence

If the incident wave propagating along $\hat{\mathbf{k}}$ is a shear wave polarized along $\hat{\mathbf{u}}$ and that along $-\hat{\mathbf{k}}'$ is a longitudinal wave of equal amplitude and frequency, we obtain the interesting result

$$f_{ps}(\hat{\mathbf{k}},\hat{\mathbf{k}}') = -\left(\frac{\mu}{\lambda+2\mu}\right)^{1/2}(\hat{\mathbf{u}}\cdot\hat{\boldsymbol{\theta}}_{-k})f_{sp}(-\hat{\mathbf{k}}',\hat{\mathbf{k}}), \qquad (3.11)$$

where $[\mu/(\lambda+\mu 2)]^{1/2}$ is the ratio of the two wave speeds.

Equations (3.9)–(3.11) are reciprocity relations between the scattered wave amplitudes for elastic waves in two-dimensional scattering geometries. Equations (3.9) and (3.10) do not involve any mode conversion and the relations are identical to that obtained for acoustic and electromagnetic waves respectively. For obstacles possessing inversion symmetry, (3.9) and (3.10) are identities. Equation (3.11) is, however, peculiar to elastic waves and relates the amplitudes of two waves that have undergone mode conversion in the opposite sense, i.e., S to P and P to S respectively, and is not an obvious relation even for obstacles possessing inversion symmetry. The relation will prove to be a useful way of checking numerical calcualtions on obstacles of arbitrary shape, the results for which are not well established in the literature.

For a circular cylinder, the scattered wave amplitude only depends on the magnitude of the angle between the direction of incidence and the direction of observation. Recalling the definition of the far field amplitudes, (2.8)

and (2.9), we note that the expansion coefficients A_n, B_n, C_n, and D_n are independent of the direction of incidence. From symmetry considerations, $C_n = 0$ for incident P waves and $A_n = 0$ for incident S waves. Then, using (3.11), we can obtain the following relation between B_n^S and D_n^P:

$$B_n^S = -D_n^P. \tag{3.12}$$

Equation (3.13) agress with the result obtained by White[10] for scattering at circular cylindrical cavities up to a sign. In the same reference it is remarked that a relation like (3.13) will not hold for scattering at elastic circular cylinders whereas (3.13) is not restricted by the material of which the cylinder is composed. Equation (3.13) has been checked numerically with very good agreement for both hard and soft cylinders embedded in a soft and hard matrix respectively.

In a recent note, Lewis and Kraft[11] have pointed out this sign error and showed that the result is applicable to solid cylinders also. The reciprocity relation, however, arrives at the same result in a more elegant manner and also gives physical meaning to the result.

B. 3-D

We now consider three-dimensional scattering geometries. The procedure is same as before, only now S is a sphere of large radius, and hence we have to integrate over two polar angles θ and ϕ.

Again we consider three cases, but this time we give the details for S–P incidence.

S–P incidence

Let the polarization of the wave incident along $\hat{\mathbf{k}}$ be denoted by $\hat{\mathbf{u}}$ which is any unit vector that lies in a plane perpendicular to $\hat{\mathbf{k}}$, and let the wave incident along $-\hat{\mathbf{k}}'$ be a compressional wave of the same amplitude and frequency. Substituting (2.16), (2.20)–(2.24), and (2.3) in (3.5), we obtain

$$\begin{aligned}
&-ik_sk_pr^2\int_\Omega \{k_p[\lambda(\hat{\mathbf{u}}\cdot\hat{\mathbf{r}}) + 2\mu(\hat{\mathbf{k}}'\cdot\hat{\mathbf{u}})(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}})] \\
&\quad + k_s\mu[(\hat{\mathbf{k}}'\cdot\hat{\mathbf{k}})(\hat{\mathbf{u}}\cdot\hat{\mathbf{r}}) + (\hat{\mathbf{k}}'\cdot\hat{\mathbf{u}})(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})]\}e^{i[k_sr\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}-k_pr\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}}]}d\Omega \\
&\quad - e^{ik_pr}\int_\Omega \{k_sf_{pp}(-\hat{\mathbf{k}}'\cdot\mathbf{r})[k_p(\lambda+2\mu)(\hat{\mathbf{u}}\cdot\hat{\mathbf{r}}) \\
&\quad - 2k_s\mu(\hat{\mathbf{r}}\cdot\hat{\mathbf{k}})(\hat{\mathbf{u}}\cdot\hat{\mathbf{r}})]e^{ik_sr\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}} \\
&\quad + k_p^2f_{ps}(\hat{\mathbf{k}},\hat{\mathbf{r}})[\lambda + 2\mu(\mathbf{k}'\cdot\hat{\mathbf{r}})^2 + (\lambda+2\mu)(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}})] \\
&\quad \times e^{-ik_pr\mathbf{k}'\cdot\hat{\mathbf{r}}}\}d\Omega - \int_\Omega \{\mu k_s^2f_{s1b}(-\hat{\mathbf{k}}',\hat{\mathbf{r}}) \\
&\quad \times [\hat{\mathbf{u}}\cdot\hat{\boldsymbol{\theta}} - (\hat{\mathbf{u}}\cdot\hat{\boldsymbol{\theta}})(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}) - (\hat{\mathbf{u}}\cdot\hat{\mathbf{r}})(\hat{\mathbf{k}}\cdot\hat{\boldsymbol{\theta}})]e^{ik_sr\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}} \\
&\quad + k_pf_{s1s}(\hat{\mathbf{k}},\hat{\mathbf{r}})[2\mu k_p(\hat{\mathbf{k}}'\cdot\hat{\boldsymbol{\theta}})(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}}) \\
&\quad + k_s\mu(\hat{\mathbf{k}}'\cdot\hat{\boldsymbol{\theta}})]e^{-ik_pr\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}}}\}d\Omega \\
&\quad - \int_\Omega \{\mu k_s^2f_{s1p}(-\hat{\mathbf{k}}',\hat{\mathbf{r}})[\hat{\mathbf{u}}\cdot\hat{\boldsymbol{\phi}} - (\hat{\mathbf{u}}\cdot\hat{\boldsymbol{\phi}})(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})]e^{ik_sr\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}} \\
&\quad + k_pf_{s1s}(\hat{\mathbf{k}},\hat{\mathbf{r}})[2\mu k_p(\hat{\mathbf{k}}'\cdot\hat{\boldsymbol{\phi}})(\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}}) + k_s\mu(\hat{\mathbf{k}}'\cdot\hat{\boldsymbol{\phi}})] \\
&\quad \times e^{-ik_pr\hat{\mathbf{k}}'\cdot\hat{\mathbf{r}}}\}d\Omega + O\left(\frac{1}{r}\right) = 0.
\end{aligned} \tag{3.13}$$

In (3.14), $\Omega = 4\pi$ is the solid angle subtended by the sphere S' and $d\Omega = \sin\theta\, d\theta\, d\phi$ is an element of solid angle.

The first integral in (3.14) vanishes by the same argument as that used in (3.6). The remaining integrals can be evaluated by the method of stationary phase as in 2—D, but we now have two variables of integration. However, for large r, there is a lemma due to Jones,[12] which states that

$$\int_\Omega f(\mathbf{r})e^{\pm ikr\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}}d\Omega = \pm\frac{2\pi i}{kr}[f(\mathbf{k})e^{\mp ikr} - f(-\mathbf{k})e^{\pm ikr}] + o\left(\frac{1}{r}\right)_{r\to\infty} \tag{3.14}$$

if $f(\hat{\mathbf{r}})$ is analytic in the range of integration.

Applying (3.15) to the remaining integrals in (3.14), we find that for large r

$$\begin{aligned}
f_{ps}(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}) = &-\left(\frac{\mu}{\lambda+2\mu}\right)\hat{\mathbf{u}}\cdot[\hat{\boldsymbol{\theta}}_{-k}f_{s1p}(-\hat{\mathbf{k}}',-\hat{\mathbf{k}}) \\
&+ \hat{\boldsymbol{\phi}}_{-k}f_{s1p}(-\hat{\mathbf{k}}',-\hat{\mathbf{k}})].
\end{aligned} \tag{3.15}$$

P–P incidence

If the waves incident along both $\hat{\mathbf{k}}$ and $-\hat{\mathbf{k}}'$ are compressional waves of the same frequency and amplitude, then

$$f_{pp}(\hat{\mathbf{k}},\hat{\mathbf{k}}') = f_{pp}(-\hat{\mathbf{k}}',-\hat{\mathbf{k}}). \tag{3.16}$$

S–S incidence

When both incident waves are shear waves polarized along $\hat{\mathbf{u}}$ and $-\hat{\mathbf{u}}'$, then proceeding in the same manner

$$\begin{aligned}
&-\hat{\mathbf{u}}\cdot[\hat{\boldsymbol{\theta}}_{-k}f_{s1s}(-\hat{\mathbf{k}},-\hat{\mathbf{k}}) + \hat{\boldsymbol{\phi}}_{-k}f_{s2s}(-\hat{\mathbf{k}}',-\hat{\mathbf{k}})] \\
&\quad = -\hat{\mathbf{u}}'\cdot[\hat{\boldsymbol{\theta}}_{k'}f_{s1s}(\hat{\mathbf{k}},\hat{\mathbf{k}}) + \hat{\boldsymbol{\phi}}_{k'}f_{s2s}(\hat{\mathbf{k}},\hat{\mathbf{k}}')].
\end{aligned} \tag{3.17}$$

Equations (3.16)–(3.18) are reciprocity relations between the scattered wave amplitudes from three-dimensional obstacles. Equation (3.17) does not involve any mode conversion and is same as that for acoustic waves; (3.16) and (3.18) are slightly more complicated than the corresponding result in 2-D because the polarization of the scattered shear waves cannot be uniquely fixed and will in general have two mutually orthogonal components. The reciprocity relations derived here can be used to prove the symmetry of the scattering matrix for elastic waves.

4. SCATTERING CROSS SECTIONS AND THE FORWARD AMPLITUDE THEOREM

In this section expressions are derived for the average rate at which energy is scattered by an obstacle of arbitrary shape—the scattering cross section. It is shown that the scattering cross section is proportional to the imaginary

part (real part in 3-D) of the amplitude of the scattered wave propagating in the forward direction.

The rate at which energy is radiated across an element of area dS at $\mathbf{r}$ with unit normal $\hat{\mathbf{n}}(\mathbf{r})$ is given by

$$dP = \frac{\partial}{\partial t}\,\mathbf{u}(r, t)\cdot\mathbf{T}(\mathbf{r}, t)\cdot\hat{\mathbf{n}}(\mathbf{r})dS, \tag{4.1}$$

where $\partial\mathbf{u}/\partial t$ and $\mathbf{T}$ are the particle velocity and stress at $\mathbf{r}$. In view of our choice of $e^{-t\omega i}$ for the time dependence of field variables instead of real time functions like $\cos\omega t$ and $\sin\omega t$ it is understood that one takes the real part of the field variables in (4.1). We define a complex energy flux vector similar to the complex Poynting vector in electromagnetism.

Since it is not the energy flow at any instant that is of interest in a harmonic field but the rate of flow of energy or power, a convenient representation is the average rate of energy flow per cycle divided by the period. Thus,

$$d\langle P\rangle = \left(\frac{\omega}{2\pi}\right)\int_0^{2\pi/\omega}[R(-i\omega\mathbf{u}(\mathbf{r})e^{-i\omega t})\cdot R(\mathbf{T}(\mathbf{r})e^{-i\omega t})]\cdot\hat{\mathbf{n}}(\mathbf{r})dS\,dt$$

$$= \frac{\omega}{2}I(\mathbf{u}(\mathbf{r})\cdot\mathbf{T}^*(\mathbf{r}))\cdot\hat{\mathbf{n}}(\mathbf{r})ds. \tag{4.2}$$

In (4.2), $d\langle P\rangle$ is the average energy flux rate vector, R and I denote real and imaginary parts respectively and the asterisk denotes complex conjugate. $\mathbf{u}$ and $\mathbf{T}$ are now complex functions independent of time.

Consider, again, an elastic obstacle of material constants ρ', λ', μ' embedded in an infinite elastic medium of material constants ρ, λ, μ. The shape S of the obstacle may be arbitrary and of two- or three-dimension. Let a plane wave either compressional or shear be incident on S. Since classical elasticity does not take into account damping (which is quite small in practise), the average rate of energy flux through a closed surface S' completely surrounding S is zero. This is a simple statement of the conservation of energy in a closed system, i.e., the rate at which energy flows into a closed surface is the rate at which it flows out. If damping was taken into account, the average power flux through S would give the rate at which energy is absorbed or emitted from within S' depending on whether the net flow is inward or outward.

The average power radiated across the closed surface S' is thus given by

$$\langle P\rangle = \frac{\omega}{2}I\left(\int_{S'}\mathbf{u}(\mathbf{r})\cdot\mathbf{T}^*(\mathbf{r})\cdot\hat{\mathbf{n}}(\mathbf{r})dS.\right) \tag{4.3}$$

For convenience we choose S' to be a circle in 2-D and a sphere in 3-D of very large radius. Conservation of energy in elastic solids implies that the integral in (4.3) is zero. Recalling that $\mathbf{u} = \mathbf{u}^i + \mathbf{u}^s$ and $\mathbf{T} = \mathbf{T}^i + \mathbf{T}^s$, we obtain from $\langle P\rangle = 0$

$$\frac{\omega}{2}I\left(\int_{S'}\mathbf{u}^s(\mathbf{r})\cdot\mathbf{T}^{s*}(\mathbf{r})\cdot\hat{\mathbf{r}}dS\right) = -\frac{\omega}{2}I\left(\int_S(\mathbf{u}^i\cdot\mathbf{T}^{s*} + \mathbf{u}^s\cdot\mathbf{T}^{i*})(\mathbf{r})\cdot\hat{\mathbf{r}}dS\right), \tag{4.4}$$

where we have used the fact that $\int_{S'}\mathbf{u}^i\cdot\mathbf{T}^{s*}\cdot\hat{\mathbf{r}}dS$ is zero for plane wave incidence.

Equation (4.3) suitably normalized is defined as the absorption cross section of the obstacle, and for elastic solids this is zero. By the same token, the left-hand side of (4.4) suitably normalized is defined as the scattering cross section of the obstacle σ. The normalization is usually made with respect to the incident power P_0. We thus have

$$\sigma = \frac{\omega}{2P_0}I\left(\int_{S'}\mathbf{u}^s\cdot\mathbf{T}^{s*}\cdot\mathbf{r}\,\mathbf{dS}\right)$$
$$= -\frac{\omega}{2P_0}I\left(\int_S(\mathbf{u}^i\cdot\mathbf{T}^{s*} + \mathbf{u}^s\cdot\mathbf{T}^{i*})\cdot\hat{\mathbf{r}}dS\right). \tag{4.5}$$

To evaluate the integral in (4.5), we again use the asymptotic form of the scattered field variables and the incident wave may be compressional or shear. All the integrals may be again performed by the method of stationary phase, as in Sec. 3. Only the results are quoted below.

A. 2-D

P-wave incidence

$$\sigma^P = \frac{2(\lambda+2\mu)k_p\omega}{P_0}I[f_{pp}(\hat{\mathbf{k}}, \hat{\mathbf{k}})]. \tag{4.6}$$

S-wave incidence

$$\sigma^s = \frac{2\mu k_s\omega}{P_0}(\hat{\mathbf{u}}\cdot\hat{\boldsymbol{\theta}}_K)I[f_{ss}(\hat{\mathbf{k}}, \hat{\mathbf{k}})], \tag{4.7}$$

where $\hat{\mathbf{u}}$ is the direction of polarization of the incident shear wave.

B.3-D

P-wave incidence

$$\sigma^P = \frac{2\pi(\lambda + 2\mu)k_p\omega}{P_0}R[f_{pp}(\hat{\mathbf{k}}, \hat{\mathbf{k}})], \tag{4.8}$$

$$\sigma^s = \frac{2\pi\mu k_s\omega}{P_0}R\{\hat{\mathbf{u}}\cdot[\hat{\boldsymbol{\phi}}_k f_{s_1s}(\hat{\mathbf{k}}, \hat{\mathbf{k}}) + \hat{\boldsymbol{\phi}}_{k_k} f_{s_2s}(\hat{\mathbf{k}}, \hat{\mathbf{k}})]\}; \tag{4.9}$$

again $\hat{\mathbf{u}}$ is the direction of polarization of the incident shear wave.

The analog of (4.6)–(4.9) for electromagnetic and acoustic waves is usually referred to as the optical or forward amplitude theorem. The relations are very useful from an experimental point of view since the rate at which energy is carried by the scattered field alone can be obtained by measuring the scattered wave amplitude in the forward direction alone. This is because except in the forward direction the scattered field propagates independent of the incident field. Relations (4.6)–(4.9) will be indispensable in proving the unitarity of the scattering matrix for elastic waves.

ACKNOWLEDGMENTS

The author wishes to thank Professors Y.H. Pao, V.K. Varatharajulu, and A.K. Kapila for helpful discussions. This research was supported by the Materials Research Center, Cornell University through a grant from the National Science Foundation.

[1]E. Freenberg, Phys. Rev. **40,** 41 (1932).
[2]H.C. Van de Hulst, Physica **25,** 740 (1949).
[3]D.S. Jones, Phil. Mag. **46,** 957 (1955).
[4]R. Glauber and V. Schomaker, Phys. Rev. **89,** 667 (1953).
[5]D.S. Jones, *The Theory of Electro-magnetism* (Pergamon, Oxford, London, New York, Paris, 1964).
[6]J.D. Achenbach, *Wave Propagation in Elastic Solids* (North-Holland–American Elsevier, Amsterdam, London, New York, 1973), Sec. 3.4.
[7]P.M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, Toronto, London, 1953), Part II, Sec. 13.1.
[8]E.T. Copson, *Asymptotic Expansions* (Cambridge U.P., Cambridge, 1967), Sec. 14.
[9]A. Erdelyi, *Asymptotic Expansions* (Dover, New York, 1956).
[10]R.M. White, J. Acoust. Soc. Amer. **30,** 771 (1958).
[11]T.S. Lewis and D.W. Kraft, J. Acoust. Soc. Amer. **56,** 1899 (1974).
[12]D.S. Jones, Proc. Camb. Phil. Soc. **48,** 736 (1952).

Positive solutions and subcriticality of energy dependent transport systems in slab geometry

C. V. Pao

Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27607
(Received 21 May 1976)

The energy dependent transport system in an anisotropic medium in slab geometry subjecting to internal source and incoming fluxes is investigated. The investigation is based on a corresponding integral equation of the boundary value problem from which a recursion formula for the determination of the solution is obtained. It is shown by using the notion of an upper solution that the convergence or divergence of the sequence of iterations depends solely on the existence or nonexistence of an upper solution. Through the construction of a suitable upper solution one can obtain an explicit estimate for the value of c, which represents the average number of secondary neutrons per collision, in terms of the (optical) slab length $2a$ so that the system is either critical or subcritical. It is shown in particular that if $c < [1 - E_2(a)]^{-1}$, where $E_2(a)$ is the exponential integral of order two, then the integral equation has exactly one nonnegative solution for any nonnegative source and incoming fluxes. This result insures the subcriticality of the system as well as a constructive recursion formula for the determination of the solution. Estimates for a more general system and the energy independent system are also obtained.

1. INTRODUCTION

In the theory of neutron transport in a finite multiplying medium an important and interesting problem is to determine the critical size of the transport region in terms of the physical parameters of the transport medium. This problem has been discussed by Case and Zweifel,[1] Mullikan[2], Keller[3], Nelson[4], and others. Most of these works are devoted to energy independent transport system in isotropic mediums. On the other hand, the energy dependent problem has been treated by Bednarz and Mika[5] and Leonard and Ferziger[6,7] in an infinite or semi-infinite slab geometry and by Pao[8] in a multidimensional region. However, these papers are mainly concerned with the construction of a solution in a nonmultiplying medium. More recently, Pao[9,10] studied the criticality question for a multiplying medium in an m-dimensional bounded region as well as in rod geometry in which some explicit estimates for the critical size of the transport region are obtained. In particular, a sufficient condition for insuring the criticality of the slab problem is given in Ref. 9. This is done by proving the existence of a nonnegative solution for any given nonnegative source, but there is no uniqueness result for the slab problem. In this paper, we study the criticality and the subcriticality question for the energy dependent transport problem in slab geometry. Our purpose is to improve the estimates obtained in Ref. 9 and to show the uniqueness of the solution so that the slab problem under consideration is actually subcritical. Moreover, we also include possible nonnegative incoming fluxes on the slab faces in the present consideration. As is to be expected, the criticality and the subcriticality of the system is independent of the source as well as the incoming fluxes. In addition to the energy dependent system, we give a brief discussion on the energy independent problem. Some conclusions on the subcriticality of this system are included.

2. AN INTERATIVE SCHEME

Consider the energy dependent neutron transport in a slab of (optical) length $2a$. The slab faces are assumed to be perpendicular to the x axis and located at $x = -a$ and $x = a$. Denote by $q(x, \mu, E)$ the internal source and by $\phi_0(\mu, E)$, $\phi_1(\mu, E)$ the incoming fluxes on the slab faces. Then the equations governing the density function $N \equiv N(x, \mu, E)$ are given by

$$\mu \frac{\partial N}{\partial x} + N = \frac{c}{2} \int_{E_0}^{E_1} \int_{-1}^{1} \sigma(\mu, E; \mu', E') N(x, \mu', E') \, d\mu' dE' + q(x, \mu, E) \tag{2.1}$$

$$(-a \le x \le a, \quad -1 \le \mu \le 1, \quad E_0 \le E \le E_1),$$

$$N(-a, \mu, E) = \phi_0(\mu, E) \quad (0 < \mu \le 1, \quad E_0 \le E \le E_1),$$

$$N(a, \mu, E) = \phi_1(\mu, E) \quad (-1 \le \mu < 0, \quad E_0 \le E \le E_1), \tag{2.2}$$

where c is the average number of secondary neutrons per collision, μ is the direction cosine relative to the x axis, E_0 and E_1 are the lower and upper limits of the energy interval, and σ is the scattering cross section satisfying the condition

$$\sup_{-1 \le \mu \le 1, E_0 \le E \le E_1} \left[\frac{1}{2} \int_{E_0}^{E_1} \int_{-1}^{1} \sigma(\mu, E; \mu', E') d\mu' dE' \right] = 1.$$

Here, for simplicity, we have taken the total cross section as one. (Thus, the length $2a$ may be considered as the optical length of the slab.)

The system (2.1), (2.2) is said to be multiplying if $c > 1$

 Copyright © 1977 American Institute of Physics

and nonmultiplying if $c < 1$. For physical reasons we assume that the scattering cross section σ and the input data q, ϕ_0, ϕ_1 are all nonnegative continuous (or piecewise continuous) functions of their respective arguments. For notational convenience, we set $D = [-a, a] \times [-1, 1] \times [E_0, E_1]$ and

$$(f(N))(x, \mu, E) \equiv \frac{c}{2} \int_{E_0}^{E_1} \int_{-1}^{1} \sigma(\mu, E; \mu', E') \times N(x, \mu', E') d\mu' dE' + q(x, \mu, E). \quad (2.3)$$

By the change of variable $x' = x - s\mu$, Eq. (2.1) may be written as

$$\frac{dN}{ds}(x' + s\mu, \mu, E) + N(x' + s\mu, \mu, E) = (f(N))(x' + s\mu, \mu, E).$$

Multiplication by e^s, integration from 0 to s, and replacement of x' by $x - s\mu$ lead to

$$N(x, \mu, E) = e^{-s} N(x - s\mu, \mu, E) + \int_0^s e^{-\tau} (f(N))(x - \tau\mu, \mu, E) d\tau. \quad (2.4)$$

Choose $s = (x + a)/\mu$ for $\mu > 0$ and $s = (a - x)/(-\mu)$ for $\mu < 0$. Then from the boundary condition (2.2), the above equation is reduced to

$$N(x, \mu, E) = (F(N))(x, \mu, E) \qquad [(x, \mu, E) \in D], \quad (2.5)$$

where

$$(F(N))(x, \mu, E) \equiv \begin{cases} e^{-(x+a)/\mu} \psi_0(\mu, E) + \int_0^{(x+a)/\mu} e^{-\tau} (f(N)) \times (x - \tau\mu, \mu, E) d\tau & (0 < \mu \le 1), \\ (f(N))(x, 0, E) & (\mu = 0), \\ e^{(a-x)/\mu} \psi_1(\mu, E) + \int_0^{(a-x)/(-\mu)} e^{-\tau} (f(N)) \times (x - \tau\mu, \mu, E) d\tau & (-1 \le \mu < 0) \\ (-a \le x \le a, \ E_0 \le E \le E_1). \end{cases} \quad (2.6)$$

The relation for $\mu = 0$ in (2.6) is obtained from (2.1) with $\mu = 0$. Notice that this relation can also be obtained from (2.6) either by letting $\mu \to 0^+$ from above in the expression for $\mu > 0$ or by letting $\mu \to 0^-$ from below in the expression for $\mu < 0$. Thus the boundary-value problem (2.1), (2.2) is reduced to the integral equation (2.5) with $F(N)$ given by (2.6). It is to be noted that $F(N)$ is positive if N is positive, and $F(N_1) \ge F(N_2)$ for any $N_1 \ge N_2 \ge 0$.

We say that the transport system is subcritical if for any nonnegative source q and incoming fluxes ϕ_0, ϕ_1 the integral equation (2.5) has a unique nonnegative solution, and it is said to be critical if there is a nontrivial nonnegative solution when $q = \phi_0 = \phi_1 = 0$. In order to investigate whether there exists a nonnegative solution, we use the method of successive approximations. Specifically, we construct, by a suitable choice of the initial iteration $N^{(0)}$, a sequence $\{N^{(k)}\}$ from the recursion formula

$$N^{(k)}(x, \mu, E) = (F(N^{(k-1)}))(x, \mu, E), \qquad k = 1, 2, \cdots. \quad (2.7)$$

To determine the convergence or divergence of this sequence, we require the notion of an upper solution which is defined as follows:

Definition 2.1: A continuous nonnegative function $\tilde{N}(x, \mu, E)$ is said to be an upper solution if it satisfies the relations

$$\tilde{N}(x, \mu, E) \ge (F(\tilde{N}))(x, \mu, E) \qquad [(x, \mu, E) \in D],$$

or, more specifically, $\tilde{N}$ satisfies the condition

$$\tilde{N}(x, \mu, E) \ge \begin{cases} e^{-(x+a)/\mu} \psi_0(\mu, E) + \int_0^{(x+a)/\mu} e^{-\tau} (f(\tilde{N})) \times (x - \tau\mu, \mu, E) d\tau & (0 < \mu \le 1), \\ (f(\tilde{N}))(x, 0, E) & (\mu = 0), \\ e^{(a-x)/\mu} \psi_1(\mu, E) + \int_0^{(a-x)/(-\mu)} e^{-\tau} (f(\tilde{N})) \times (x - \tau\mu, E) d\tau & (-1 \le \mu < 0) \\ (-a \le x \le a, \ E_0 \le E \le E_1). \end{cases} \quad (2.8)$$

Notice that if N is a nonnegative solution of (2.5), then it is also an upper solution. As will be seen in the following discussions, upper solutions play an essential role in the study of criticality problems.

Let $\tilde{N}$ be an upper solution, and let $N^{(0)} = \tilde{N}$ and $N^{(0)} = 0$ as two distinct initial iterations. Then we can obtain two sequences, denoted by $\{\bar{N}^{(k)}\}$ and $\{\underline{N}^{(k)}\}$, respectively, from (2.7). These two sequences possess the following monotone properties:

Lemma 2.1: The sequence $\{\bar{N}^{(k)}\}$ is monotone nonincreasing while the sqeuence $(\underline{N}^{(k)})$ is monotone nondecreasing. Moreover, $\underline{N}^{(k)} \le \bar{N}^{(k)}$ for every $k = 1, 2, \cdots$.

Proof: A proof of the above results is contained in Ref. 9 for the case $\phi_0 = \phi_1 = 0$. Here we give a brief sketch. In view of the recursion relation (2.7) and the definition of $\tilde{N}$ in (2.8) it is clear that $\bar{N}^{(0)} - \bar{N}^{(1)} = \tilde{N} - F(\tilde{N}) \ge 0$, where we have suppressed the independent variables (x, μ, E). Suppose that $\bar{N}^{(k)} \le \bar{N}^{(k-1)}$. Then from the monotone incressing property of $F(N)$, we have

$$\bar{N}^{(k)} - \bar{N}^{(k+1)} = F(\bar{N}^{(k-1)}) - F(\bar{N}^{(k)}) \ge 0.$$

It follows from the induction principle that $\bar{N}^{(k)} \ge \bar{N}^{(k+1)}$ for every k. This proves the montone nonincrasing property of $\bar{N}\{^{(k)}\}$. The proofs of $\underline{N}^{(k)} \le \underline{N}^{(k+1)}$ and $\underline{N}^{(k)} \le \bar{N}^{(k)}$ are similar and are omitted.

In view of Lemma 2.1 we immediately have the following conclusion.

Theorem 2.1: Let $\tilde{N}$ be an upper solution. Then the pointwise limits

$$\lim \bar{N}^{(k)}(x, \mu, E) = \bar{N}(x, \mu, E), \quad \lim \underline{N}^{(k)}(x, \mu, E) = \underline{N}(x, \mu, E), \qquad \text{as } k \to \infty \quad (2.9)$$

exist, are solutions of (2.5), and satisfy the relation

$$0 \le \underline{N}^{(1)} \le \underline{N}^{(2)} \le \cdots \le \underline{N} \le \bar{N} \le \cdots \le \bar{N}^{(2)} \le \bar{N}^{(1)} \le \bar{N}. \tag{2.10}$$

Moreover, upper solutions exist unless Eq. (2.5) has no nonnegative solution.

Proof: By Lemma 2.1 the sequence $\{\bar{N}^{(k)}\}$ is monotone nonincreasing and is bounded from below while $\{\underline{N}^{(k)}\}$ is monotone nondecreasing and is bounded from above. Thus the pointwise limits in (2.9) exist and satisfy the relation (2.10). By letting $k \to \infty$ in (2.7) and applying the dominated convergence theorem we conclude that $\bar{N}$ and $\underline{N}$ are both solutions of (2.5). Finally, since every nonnegative solution of (2.5) is also an upper solution there must exist such a function unless the problem has no nonnegative solution. This proves the theorem.

Another interesting consequence of Lemma 2.1 is the following:

Theorem 2.2: Let $\tilde{N}$ be an upper solution and let N be any solution of (2.5) satisfying $0 \le N \le \tilde{N}$. Then

$$\underline{N} \le N \le \bar{N}.$$

Proof: It is easily seen by an inductive argument that $N \le \bar{N}^{(k)}$ and $N \ge \underline{N}^{(k)}$ for every $k = 1, 2, \cdots$. It follows by letting $k \to \infty$ that $\underline{N} \le N \le \bar{N}$.

In view of Theorem 2.2 we say that $\bar{N}$ is a maximal solution (with respect to $\tilde{N}$) while $\underline{N}$ is the minimal solution. Notice that maximal solutions may depend on upper solutions but the minimal solution is independent of any particular upper solution. Nevertheless, the existence of the minimal solution (as well as maximal solutions) depends on the existence of an upper solution. In fact, we have the following:

Corollary: The sequence $\{\underline{N}^{(k)}\}$ converges to the minimal solution $\underline{N}$ if and only if there exists an upper solution.

Proof: The proof follows from Theorem 2.1 and the definition of an upper solution.

3. CONSTRUCTION OF UPPER SOLUTIONS AND CRITICALITY PROBLEM

In view of Theorem 2.1, the criticality of the system can be insured if for any nonnegative functions q, ϕ_0, ϕ_1 we can find an upper solution. Clearly such a function may not exist unless there is some restriction on the magnitude of c or on the slab length $2a$. The purpose of this section is to give a sufficient condition on c in terms of σ and the slab length so that an explicit upper soluton can be constructed. To give a motivation in our construction of an upper solution, we observe that if $\tilde{N}$ is a function such that $f(\tilde{N}) \le M$ on D, where M is a constant then the first integral in (2.8) is bounded by $M(1 - e^{-(x+a)/\mu})$ for $0 < \mu \le 1$ while the second integral is bounded by $M(1 - e^{(a-x)/\mu})$ for $-1 \le \mu < 0$. This observation leads to the consideration of a possible upper solution in the form

$$\tilde{N}(x, \mu, E) = \begin{cases} e^{-(x+a)/\mu}\phi_0(\mu, E) + M(1 - e^{-(x+a)/\mu}) \\ \qquad (0 < \mu \le 1), \\ M \qquad (\mu = 0), \\ e^{(a-x)/\mu}\phi_1(\mu, E) + M(1 - e^{(a-x)/\mu}) \\ \qquad (-1 \le \mu < 0) \qquad (-a \le x \le a, \\ \qquad E_0 \le E \le E_1). \end{cases} \tag{3.1}$$

We show that under a rather simple condition on c the above function is indeed an upper solution when the constant M is suitably chosen. The choice of M depends on the physical parameters c, a, σ and the input data q, ϕ_0, ϕ_1. For convenience, we use the following notations:

$$\bar{\sigma}(\mu', E') = \sup\{\sigma(\mu, E; \mu', E'); \quad -1 \le \mu \le 1, \quad E_0 \le E \le E_1\}$$

$$\bar{\sigma}_m(\mu') = \max\left\{\int_{E_0}^{E_1} \bar{\sigma}(\mu', E')\, dE', \int_{E_0}^{E_1} \bar{\sigma}(-\mu', E')\, dE'\right\} \quad (0 \le \mu' \le 1). \tag{3.2}$$

Our first result for the criticality problem is the following

Theorem 3.1: Assume that

$$c\int_0^1 \bar{\sigma}_m(\mu')(1 - e^{-a/\mu'})\, d\mu' < 1. \tag{3.3}$$

Then for any nonnegative functions q, ϕ_0, ϕ_1 the integral equation (2.5) has a maximal solution $\bar{N}$ and a minimal solution $\underline{N}$ which can be constructed from (2.7) by iterations. Thus the system is either critical or subcritical.

Proof: Let $\rho = c\int_0^1 \bar{\sigma}_m(\mu')(1 - e^{-a/\mu'})\, d\mu'$, and let δ be any positive constant satisfying $\delta \le 1 - \rho$. Denote by $\bar{q}$, $\bar{\phi}_0$, $\bar{\phi}_1$ the respective least upper bound of $q(x, \mu, E)$, $\phi_0(\mu, E)$, and $\phi_1(\mu, E)$. We show that for any constant M satisfying

$$M \ge \delta^{-1}(\bar{q} + c(\bar{\phi}_0 + \bar{\phi}_1)\int_0^1 \bar{\sigma}_m(\mu')\, d\mu') \tag{3.4}$$

the function $\tilde{N}$ given by (3.1) is an upper solution. For this purpose, we first derive an estimate for $f(\tilde{N})$. In view of (2.3) and (3.1),

$$\begin{aligned}(f(\tilde{N}))(x, \mu, E) \le\ & \tfrac{1}{2}c\int_{E_0}^{E_1}\Big\{\int_0^1 \sigma(\mu, E; \mu', E')\,[\bar{\phi}_0\, e^{-(x+a)/\mu'} \\ & + M(1 - e^{-(x+a)/\mu')}]\, d\mu' \\ & + \int_{-1}^0 \sigma(\mu, E; \mu', E')\,[\bar{\phi}_1 e^{(a-x)/\mu'} \\ & + M(1 - e^{(a-x)/\mu'})]\, d\mu'\Big\} dE' + \bar{q}\end{aligned} \tag{3.5}$$

for every (x, μ, E) in D. By replacing μ' by $-\mu'$ in the second integral and using the notation in (3.2) we obtain

$$
\begin{aligned}
&(f(\tilde N))(x,\mu,E)\\
&\le \tfrac{1}{2}c\int_{E_0}^{E_1}\Big\{\int_0^1\bar\sigma(\mu',E')\,[\bar\psi_0+M(1-e^{-(x+a)/\mu'})]\,d\mu'\\
&\quad+\int_0^1\bar\sigma(-\mu',E')[\bar\psi_1+M(1-e^{-(a-x)/\mu'}]d\mu'\Big\}dE'+\bar q\\
&\le\tfrac{1}{2}c\int_0^1\bar\sigma_m(\mu')\,[(\bar\psi_0+\bar\psi_1)\\
&\quad+M(2-e^{-(x+a)/\mu'}-e^{-(a-x)/\mu'})]d\mu'+\bar q. \qquad (3.6)
\end{aligned}
$$

It is easily seen by a simple calculation that for each fixed $\mu'>0$ the maximum of the function $(2-e^{-(x+a)/\mu'}-e^{-(a-x)/\mu'})$ occurs at $x=0$ and is equal to $2(1-e^{-a/\mu'})$. Hence by (3.6), (3.4) and the fact that $\rho\le 1-\delta$ we have

$$
\begin{aligned}
&(f(\tilde N))(x,\mu,E)\le c\int_0^1\bar\sigma_m(\mu')\,[(\bar\psi_0+\bar\psi_1)\\
&\quad+M(1-e^{-a/\mu'})]d\mu'+\bar q\\
&\quad=c(\bar\psi_0+\bar\psi_1)\int_0^1\bar\sigma_m(\mu')d\mu'+M\rho+\bar q\\
&\quad\le M\delta+M\rho\le M \qquad (3.7)
\end{aligned}
$$

for all (x,μ,E) in D. Since

$$
\begin{aligned}
&-a\le x-\tau\mu\le x \quad\text{for}\quad 0\le\tau\le(x+a)/\mu,\\
&\qquad\qquad 0<\mu\le 1,\\
&x\le x-\tau\mu\le a \quad\text{for}\quad 0\le\tau\le(a-x)/(-\mu),\\
&\qquad\qquad -1\le\mu<0,
\end{aligned}
\qquad (3.8)
$$

we see that $(f(N))(x-\tau\mu,\mu,E)\le M$ for the above values of τ and all (x,μ,E) in D. It follows from (2.6) and (3.1) that $(F(\tilde N))(x,\mu,E)\le\tilde N(x,\mu,E)$ which shows that $\tilde N$ is an upper solution. The result of the theorem follows from Theorem 3.1.

When $\bar\sigma_m(\mu')\le 1$, the requirement (3.3) holds if $c\int_0^1(1-e^{-a/\mu'})\,d\mu'<1$. In terms of the exponential integrals $E_n(z)$ which is defined by

$$
E_n(z)=\int_1^\infty e^{-zt}t^{-n}dt=\int_0^1 e^{-z/\mu}\mu^{n-2}d\mu \qquad (n=0,1,2\cdots), \qquad (3.9)
$$

we have the following simple relation between c and a.

Corollary: Assume that $\bar\sigma_m(\mu')\le 1$ for $0<\mu'\le 1$ and

$$
c<[1-E_2(a)]^{-1}. \qquad (3.10)
$$

Then the slab problem is either critical or subcritical.

Remark 3.1: Since the values of $E_n(z)$ have been tabulated and can be found in standard handbooks (e.g., see Refs. 11, 12) the result in the above corollary is particularly convenient to use. It is to be noted that (3.10) holds for any finite slab length when $c\le 1$. However, as $a\to\infty$ (i.e., for an infinite slab) condition (3.10) becomes $c<1$ which is to be expected physically.

When the scattering process is anisotropic such as the linear case $\sigma=(1+b\mu\mu')\,h(E,E')$, condition (3.3) can sometimes be improved by writing $\sigma=\sigma^{(1)}+\sigma^{(2)}$, where

$$
\begin{aligned}
\sigma^{(1)}(\sigma,E;\mu',E')&=\tfrac{1}{2}[\sigma(\mu,E;\mu',E')\\
&\quad+\sigma(\mu,E;-\mu',E')],\\
\sigma^{(2)}(\mu,E;\mu',E')&=\tfrac{1}{2}[\sigma(\mu,E;\mu',E')\\
&\quad-\sigma(\mu,E;-\mu',E')].
\end{aligned}
\qquad (3.11)
$$

To see this, we define $\bar\sigma^{(j)}(\mu',E')$ as in (3.2) for $j=1,2$ and set

$$
\begin{aligned}
\Sigma_m(\mu')=\max\Big\{&\int_{E_0}^{E_1}\bar\sigma^{(1)}(\mu',E')\,dE',\\
&\int_{E_0}^{E_1}\bar\sigma^{(2)}(\mu',E')dE'\Big\}\\
&(0\le\mu'\le 1).
\end{aligned}
\qquad (3.12)
$$

With this notation, the functions $\bar\sigma^{(1)}$, Σ_m are nonnegative while $\bar\sigma^{(2)}$ may assume negative values in $[-1,1]\times[E_0,E_1]$. Now by letting $\sigma=\sigma^{(1)}+\sigma^{(2)}$ in (3.5) and replacing μ' by $-\mu'$ in the second integral we obtain

$$
\begin{aligned}
&(f(\tilde N))(x,\mu,E)\\
&\le\tfrac{1}{2}c\int_{E_0}^{E_1}\Big\{\int_0^1[\bar\sigma^{(1)}(\mu',E')+\bar\sigma^{(2)}(\mu',E')]\\
&\quad\times[\bar\psi_0+M(1-e^{-(x+a)/\mu'})]\,d\mu'\\
&\quad+\int_0^1[\bar\sigma^{(1)}(-\mu',E')+\bar\sigma^{(2)}(-\mu',E')]\\
&\quad\times[\bar\psi_1+M(1-e^{-(a-x)/\mu'})]\,du'\Big\}\,dE'+\bar q,
\end{aligned}
\qquad (3.13)
$$

where $\tilde N$ is the function given by (3.1). In view of the symmetric and the antisymmetric property of $\bar\sigma^{(1)}$ and $\bar\sigma^{(2)}$ the above inequality implies that

$$
\begin{aligned}
&(f(\tilde N))(x,\mu,E)\\
&\le\tfrac{1}{2}Mc\int_0^1[\bar\sigma^{(1)}(\mu')\,(2-e^{-(x+a)/\mu'}-e^{-(a-x)/\mu'})\\
&\quad+\bar\sigma^{(2)}(\mu')\,(e^{-(a-x)/\mu'}-e^{-(x+a)/\mu'})]du'+K, \qquad (3.14)
\end{aligned}
$$

where

$$
\bar\sigma^{(j)}(\mu')=\int_{E_0}^{E_1}\bar\sigma^{(j)}(\mu',E')dE',\quad j=1,2, \qquad (3.15)
$$

and

$$
\begin{aligned}
K=\tfrac{1}{2}c\int_0^1&[\bar\sigma^{(1)}(\mu')\,(\bar\psi_0+\bar\psi_1)\\
&+\bar\sigma^{(2)}(\mu')\,(\bar\psi_0-\bar\psi_1)]d\mu'+\bar q.
\end{aligned}
$$

Since the second term of the integrand in (3.14) is bounded by $\pm\Sigma_m(\mu')\,(e^{-(a-x)/\mu'}-e^{-(x+a)/\mu'})$ according to $x\ge 0$ or $x\le 0$ we see that

$$
\begin{aligned}
&(f(\tilde N))(x,\mu,E)\le\tfrac{1}{2}Mc\int_0^1\Sigma_m(\mu')\,(2\\
&\qquad-2e^{-(|x|+a)/\mu'})\,du'+K\\
&\qquad\le Mc\int_0^1\Sigma_m(\mu')\,(1-e^{-2a/\mu'})d\mu'+K\\
&\qquad\equiv M\rho_1+K. \qquad (3.16)
\end{aligned}
$$

Hence, if $\rho_1<1$, then by choosing any positive constant $\delta\le 1-\rho_1$ and $M\ge K/\delta$ we obtain

$$(f(\tilde{N}))(x, \mu, E) \le M(1 - \delta) + K \le M \quad \text{for all } (x, \mu, E) \in D. \tag{3.17}$$

It follows from (2.6), (3.8), and (3.17) that $(F(\tilde{N}))(x, \mu, E) \le \tilde{N}(x, \mu, E)$, and thus $\tilde{N}$ is again an upper solution. In conclusion, we have

Theorem 3.2: Assume that

$$c\int_0^1 \Sigma_m(\mu')(1 - e^{-2a/\mu'})\, d\mu' < 1, \tag{3.18}$$

where $\Sigma_m(\mu')$ is defined by (3.12). Then for any nonnegative functions q, ϕ_0, ϕ_1 the integral equation (2.5) has at least one nonnegative solution and thus the system is either critical or subcritical.

It is easily seen that if σ is symmetric with respect to μ', then better estimates for c can be obtained from condition (3.3) than from condition (3.18) since in this case $\sigma^{(1)} = \sigma$, $\sigma^{(2)} = 0$, and thus $\Sigma_m(\mu') = \bar{\sigma}_m(\mu')$. However, if σ is not symmetric with respect to μ', condition (3.18) sometimes leads to better estimates for c. As a demonstration, we consider the special scattering cross section in the form of

$$\sigma(\mu, E; \mu', E') = (1 + b\mu\mu')\, h(E, E') \quad (0 \le b \le 1), \tag{3.19}$$

where $\max_E [\int_{E_0}^{E_1} h(E, E')dE'] = 1$. In this case, $\bar{\sigma}_m(\mu') = 1 + b\mu'$ and thus (3.3) is equivalent to

$$c < \{(1 + b/2) - [E_2(a) + bE_3(a)]\}^{-1}, \tag{3.20}$$

where $E_n(a)$ is the exponential integral in (3.9). On the other hand, since $\sigma^{(1)} = h(E, E')$, $\sigma^{(2)} = b\mu\mu' h(E, E')$, we see that $\sigma^{(1)} = 1$, $\bar{\sigma}^{(2)} = b\mu'$, and so $\Sigma_m(\mu') = 1$. By (3.18), the requirement on c becomes

$$c < [1 - E_2(2a)]^{-1}. \tag{3.21}$$

It is clear that for $b \neq 0$ we may obtain better estimates from (3.21) than from (3.20).

4. SUBCRITICALITY PROBLEM

The results in Theorems 3.1 and 3.2 insure the existence of a maximal solution $\bar{N}$ and a minimal solution $\underline{N}$ for any nonnegative functions q, ϕ_0, ϕ_1 but not the uniqueness of the solution. In order to obtain subcriticality of the system it is necessary to show that Eq. (2.5) has exactly one nonnegative solution. For this purpose, we need an additional physically realistic condition, namely that σ be positive and be symmetric with respect to (μ, E) and (μ', E'). It should be pointed out that a novelty of our uniqueness proof is that the function $(\bar{N} - \underline{N})$ is nonnegative, which is not necessarily true for the difference of two arbitrary nonnegative solutions. The main result for the uniqueness problem is contained in the following:

Theorem 4.1: Assume that $\bar{\sigma}(\mu', E') > 0$ and $\sigma(\mu, E; \mu', E') = \sigma(\mu', E'; \mu, E)$. If (3.3) holds, then for any nonnegative functions q, ϕ_0, ϕ_1 the integral equation (2.5) has exactly one nonnegative solution which can be determined from (2.7) with the initial iteration $N^{(0)} = 0$. Thus the system is subcritical.

Proof: We first show that every maximal solution $\bar{N}$ coincides with the minimal solution $\underline{N}$. Let $N = \bar{N} - \underline{N}$ so that $N \ge 0$ in D. Since $\bar{N}$ and $\underline{N}$ are both solutions of (2.5), N is a solution of (2.5) with $q = \phi_0 = \phi_1 = 0$. Specifically, N satisfies the equation

$$N(x, \mu, E) = \begin{cases} \displaystyle\int_0^{(x+a)/\mu} e^{-\tau}\Big[\tfrac{1}{2}c\int_{E_0}^{E_1}\int_{-1}^{1} \sigma(\mu, E; \mu', E') \\ \quad \times N(x - \tau\mu, \mu', E')\, d\mu' dE'\Big]d\tau \\ \quad (0 < \mu \le 1), \\ \displaystyle\int_0^{(a-x)/(-\mu)} e^{-\tau}\Big[\tfrac{1}{2}c\int_{E_0}^{E_1}\int_{-1}^{1} \sigma(\mu, E; \mu', E') \\ \quad \times N(x - \tau\mu, \mu', E')\, d\mu' dE'\Big]d\tau \\ \quad (-1 \le \mu < 0) \\ (-a \le x \le a, \quad E_0 \le E \le E_1). \end{cases} \tag{4.1}$$

The case for $\mu = 0$ may be included in either one of the above integrals. In view of the notations in (3.2) and the hypothesis on σ we see that $\bar{\sigma}(\mu', E') = \bar{\sigma}(\mu, E)$. Now by multiplying (4.1) by $\bar{\sigma}(\mu, E)$, integrating over (μ, E), and setting

$$g(x) \equiv \int_{E_0}^{E_1}\int_{-1}^{1} \bar{\sigma}(\mu, E) N(x, \mu, E) d\mu dE \tag{4.2}$$

we obtain

$$g(x) = \int_{E_0}^{E_1}\int_{0}^{1} \bar{\sigma}(\mu, E) I_1(x, \mu, E) d\mu dE + \int_{E_0}^{E_1}\int_{-1}^{0} \bar{\sigma}(\mu, E) I_2(x, \mu, E) d\mu dE, \tag{4.3}$$

where I_1, I_2, are the two integrals on the right side of (4.1). Notice that $g(x) \ge 0$ on $[-a, a]$. Let $x_0 \in [-a, a]$ such that $g(x_0) = \max g(x)$ on $[-a, a]$. Then in view of the relation (3.8) and the nonnegative property of N we have

$$\begin{aligned} I_1(x_0, \mu, E) &\le \int_0^{(x_0+a)/\mu} e^{-\tau}\Big[\tfrac{1}{2}c\int_{E_0}^{E_1}\int_{-1}^{1} \bar{\sigma}(\mu', E') \\ &\quad \times N(x_0 - \tau\mu, \mu', E') d\mu' dE'\Big]d\tau \\ &\le \frac{c}{2}\int_0^{(x_0+a)/\mu} e^{-\tau} g(x_0) d\tau \\ &= (\tfrac{1}{2}c) g(x_0)(1 - e^{-(x_0+a)/\mu}), \quad (0 < \mu \le 1) \end{aligned} \tag{4.4}$$

and, similarly,

$$I_2(x_0, \mu, E) \le (\tfrac{1}{2}c) g(x_0)(1 - e^{(a-x_0)/\mu}) \quad (-1 \le \mu < 0). \tag{4.5}$$

It follows from (4.3)–(4.5) that

$$\begin{aligned} g(x_0) &\le (\tfrac{1}{2}c) g(x_0) \int_{E_0}^{E_1}\int_0^1 [\bar{\sigma}(\mu, E)(1 - e^{-(x_0+a)/\mu}) \\ &\quad + \bar{\sigma}(-\mu, E)(1 - e^{-(a-x_0)/\mu})]\, d\mu dE \\ &\le (\tfrac{1}{2}c) g(x_0) \int_0^1 \bar{\sigma}_m(\mu)(2 - e^{-(x_0+a)/\mu} - e^{-(a-x_0)/\mu}) d\mu \end{aligned}$$

$$\leq cg(x_0)\int_0^1 \bar{\sigma}_m(\mu)(1 - e^{-a/\mu})d\mu. \tag{4.6}$$

But, by the condition (3.3), the above inequality is impossible unless $g(x_0) = 0$. Hence $g(x) = 0$ for all x in $[-a, a]$. In view of the hypothesis $\sigma > 0$ and the nonnegative property of N we must have $N = 0$ which proves that $\bar{N} = \underline{N}$. Now if N^* is any other nonnegative solution then it is also an upper solution. The above proof implies that $N^* = \underline{N}$. This shows that there exists only one nonnegative solution which completes the proof of the theorem.

An immediate consequence of the above theorem is the following.

Corollary: Assume that $\bar{\sigma}_m(\mu') \leq 1$, $\sigma(\mu, E; \mu', E') = \sigma(\mu', E'; \mu, E) > 0$, and $c < (1 - E_2(a))^{-1}$. Then Eq. (2.5) has a unique nonnegative solution for any nonnegative q, ϕ_0 and ϕ_1.

We next give a uniqueness result for the solution obtained in Theorem 3.2. Write $\sigma = \sigma^{(1)} + \sigma^{(2)}$ and assume that $\sigma(\mu, E; \mu', E') = \sigma(\mu', E; \mu', E) > 0$. Then $\bar{\sigma}(\mu', E') = \bar{\sigma}(\mu, E)$ and $\bar{\sigma}(\mu, E) \leq \bar{\sigma}^{(1)}(\mu, E) + \bar{\sigma}^{(2)}(\mu, E)$. From the first inequality in (4.6) and the relations $\bar{\sigma}^{(1)}(-\mu, E) = \bar{\sigma}^{(1)}(\mu, E)$, $\bar{\sigma}^{(2)}(-\mu, E) = -\bar{\sigma}^{(2)}(\mu, E)$ we have

$$g(x_0) \leq (\tfrac{1}{2}c)g(x_0)\Big[\int_{E_0}^{E_1}\int_0^1 \bar{\sigma}^{(1)}(\mu, E)\,(2 - e^{-(x_0+a)/\mu} - e^{-(a-x_0)/\mu})d\mu dE + \int_{E_0}^{E_1}\int_0^1 \bar{\sigma}^{(2)}(\mu, E)\,(e^{-(a-x_0)/\mu} - e^{-(x_0+a)/\mu})\,d\mu dE\Big]. \tag{4.7}$$

Since the integrand in the second integral in the above relation is bounded by $\pm\,\Sigma_m(\mu)\,(e^{-(a-x_0)/\mu} - e^{-(x_0+a)/\mu})$ according to $x_0 \geq 0$ or $x_0 \leq 0$, we obtain

$$g(x_0) \leq (\tfrac{1}{2}c)g(x_0)\int_0^1 \Sigma_m(\mu)\,(2 - 2e^{-(|x_0|+a)/\mu})\,d\mu \leq cg(x_0)\int_0^1 \Sigma_m(\mu)\,(1 - e^{-2a/\mu})\,d\mu. \tag{4.8}$$

Hence, if (3.18) holds, then the above inequality implies that $g(x_0) = 0$. It follows that $g(x) = 0$ on $[-a, a]$ which proves that $\bar{N} = \underline{N}$. In fact $\underline{N}$ is the unique solution of (2.5). This leads to the following conclusion:

Theorem 4.2: Assume that $\sigma(\mu, E; \mu', E') = \sigma(\mu', E'; \mu, E) > 0$ and that (3.18) holds. Then for any nonnegative functions q, ϕ_0, ϕ_1, Eq. (2.5) has a unique nonnegative solution which can be determined from (2.7) with the initial iteration $N^{(0)} = 0$.

5. THE ENERGY-INDEPENDENT PROBLEM

When the system is energy-independent the equation governing the density function $N(x, \mu)$ becomes

$$\mu\frac{\partial N}{\partial x} + N = (\tfrac{1}{2}c)\int_{-1}^{1}\sigma(\mu, \mu')N(x, \mu')d\mu' + q(x, \mu) \qquad (-a \leq x \leq a,\ -1 \leq \mu \leq 1), \tag{5.1}$$

$$N(-a, \mu) = \phi_0(\mu) \qquad (0 < \mu \leq 1),$$
$$N(a, \mu) = \phi_1(\mu) \qquad (-1 \leq \mu < 0). \tag{5.2}$$

The corresponding integral equation, the process of iterations and the definition of an upper solution $\tilde{N}(x, \mu)$ are the same as for the energy-dependent problem except without the energy variable E (see (2.5)–(2.8)). By using the same function $\tilde{N}(x, \mu)$ given by (3.1) (which is independent of E), it is possible to show that $\tilde{N}$ is an upper solution if

$$c\int_0^1 \bar{\sigma}_M(\mu')\,(1 - e^{-a/\mu'})\,d\mu' < 1, \tag{5.3}$$

where

$$\bar{\sigma}_M(\mu') = \max\{\bar{\sigma}(\mu'), \bar{\sigma}(-\mu')\}$$

with

$$\bar{\sigma}(\mu') = \sup\{\sigma(\mu, \mu');\ -1 \leq \mu \leq 1\}. \tag{5.4}$$

On the other hand, if we write $\sigma = \sigma^{(1)} + \sigma^{(2)}$ and set

$$\Sigma_1(\mu') = \max\{\bar{\sigma}^{(1)}(\mu'), \bar{\sigma}^{(2)}(\mu')\}, \tag{5.5}$$

where $\sigma^{(1)}$, $\sigma^{(2)}$ are given by (3.11) without the variable E then the function $\tilde{N}$ given by (3.1) is again an upper solution if

$$c\int_0^1 \Sigma_1(\mu')\,(1 - e^{-2a/\mu'})d\mu' < 1. \tag{5.6}$$

Furthermore, if $\sigma(\mu, \mu') = \sigma(\mu', \mu) > 0$, then the maximal solution coincides with the minimal solution and there exists exactly one nonnegative solution. Since the proofs of the above statements are the same as for the energy-dependent problem, we only state some results in the following:

Theorem 5.1: Let $\sigma(\mu, \mu') = \sigma(\mu', \mu) > 0$ and let either (5.3) or (5.6) hold. Then the corresponding integral equation for (5.1), (5.2) has a unique nonnegative solution, and thus the system is subcritical.

The following conclusions are immediate consequences of Theorem 5.1.

Corollary 1: Let $0 < \sigma(\mu, \mu') = \sigma(\mu', \mu) \leq 1$ and let $c < [1 - E_2(a)]^{-1}$. Then the conclusion in Theorem 5.1 holds.

Corollary 2: Let $\sigma(\mu, \mu') = \sigma(\mu', \mu) > 0$, $\Sigma_1(\mu') \leq 1$, and let $c < [1 - E_2(2a)]^{-1}$. Then the conclusion in Theorem 5.1 holds.

[1]K.M. Case and P.F. Zweifel, J. Math. Phys. **4**, 1376 (1963).
[2]T.W. Mullikin, J. Math. Anal. Appl. **5**, 184 (1962).
[3]H.B. Keller, SIAM J. Appl. Math. **8**, 560 (1960).
[4]P. Nelson, Jr., J. Math. Anal. Appl. **35**, 90 (1971).
[5]R.J. Bednarz and J. Mika, J. Math. Phys. **4**, 1285 (1963).
[6]J.H. Ferziger and A. Leonard, Ann. Phys. (N.Y.) **22**, 192 (1963).
[7]A. Leonard and J.H. Ferziger, Nucl. Sci. Eng. **II, III**, 26, 170, and 180 (1966).
[8]C.V. Pao, SIAM J. Appl. Math. **30**, 501 (1976).

[9]C.V. Pao, SIAM J. Appl. Math. (to appear).
[10]C.V. Pao, J. Math. Phys. **17**, 1767 (1976).
[11]M. Abramowitz and I.A. Stegun, Eds., *Handbook of Mathematical Functions* (Dover, New York, 1965).
[12]K.M. Case, F. de Hoffmann, and G. Plazek, "Introduction to the Theory of Neutron Diffusion," Los Alamos Scientific Laboratory, Los Alamos, New Mexico, 1953.

Partial-wave analysis for supersymmetric scattering amplitudes

P. D. Jarvis

Blackett Laboratory of Physics, Imperial College, London SW7 2BZ, England
(Received 18 May 1976)

The reduction of the direct product of two unitary irreducible representations of supersymmetry into a direct sum of UIR's is carried out for the massive case, and the Clebsch–Gordan coefficients are written down. The supersymmetric coupling of the different spin components of a multiplet arises from the use of the "superhelicity" basis (superhelicity $\kappa = -j_0, -j_0+1, \ldots, j_0-1, j_0$) in which the spin is not diagonal. Here $j_0 = 0, 1/2, 1, \ldots$ is the "superspin," and the ordinary helicity, λ, is given by $\lambda = \kappa$ or $\kappa \pm 1/2$. The physical results are retrieved by transforming back to the spin basis after the reduction. The results of the reduction are used to analyze the scattering processes $1 \to 2+3$ and $1+2 \to 3+4$ for particles belonging to supersymmetric multiplets. The ordinary partial wave helicity amplitudes are given in terms of a small number of reduced partial wave superhelicity amplitudes corresponding to given total superspin. By continuing the latter to complex superspin, it is shown how, in the high-energy limit in the crossed channel, a singularity in one superhelicity amplitude contributes to the high-energy behavior of several different spin channels.

1. INTRODUCTION

In an earlier paper[1] a program was initiated for developing a general formalism for investigating the idea of fermion–boson "super-symmetry," in a more general context than the supersymmetric Lagrangian theories, recently much studied in the literature, in which the idea was first formulated.[2a,b] Applications envisaged for this formalism included such things as generalized supersymmetry schemes, supersymmetry invariance and constraints on scattering amplitudes, partial-wave analysis and high energy behavior, and so on. The first stage of this work, which was carried out in Ref. 1, was a detailed analysis of those unitary irreducible representations of the supersymmetry algebra which are of most immediate importance in physical applications, namely in the massive and massless cases (the latter being a restricted case of the lightlike UIR's). Convenient bases, of the form $|p^2 > 0, j_0; p\sigma j\lambda\rangle$ and $|p^2 = 0, \lambda_0; p\lambda\rangle$, were found in each case, where j_0 and λ_0 are the superspin and superhelicity, and $j = j_0, j_0 \pm \frac{1}{2}, \lambda = \lambda_0, \lambda_0 - \frac{1}{2}$, are the constituent spins and helicities occurring (which are diagonal in this basis). In the massive case, the label $\sigma = \pm\frac{1}{2}, 0$ serves to distinguish the different spins occurring, since $j = j_0 \pm \frac{1}{2} \mp |\sigma|$. [Our notations follow Refs. 1 and 2b. The commutation and anticommutation relations of the supersymmetry algebra, together with the definitions of the basis state labels, are given in Appendix A. This basis, with diagonal constituent spins, especially in the spin (parity) classification of Ref. 2b, $j_0^{\pm\eta}$, $(j_0 \pm \frac{1}{2})^{i\eta}$, is here referred to as the "spin basis," to distinguish it from the "super-helicity basis" of Sec. 2.] This work provides insights into the structure of these UIR's, especially in the massless case; it also provides a natural framework for describing the structure of higher-spin superfields, in the finite-dimensional case.

It is the purpose of this paper to present some results in one development of this program. As a first step in investigating the structure of supersymmetric multiparticle states and scattering amplitudes, we solve the problem of reducing the direct product of two UIR's of supersymmetry (in the massive case), into a direct sum of UIR's. In doing so, we pass from the spin basis, described above, to the "superhelicity basis," in which states are labeled as $|p^2 > 0, j_0; p\sigma\kappa\lambda\rangle$, where now $\kappa = -j_0, -j_0+1, \ldots, j_0$ is the superhelicity, and the helicity $\lambda = \kappa \pm \frac{1}{2} \mp |\sigma|$. The spin basis remains the basis to which all our physical results are to be referred, but as we shall see, the structure of the supersymmetry algebra, and the reduction of the direct product, are most easily described in the superhelicity basis (in which the ordinary spin, j, is nondiagonal). The interplay between the two bases is to be expected, in fact, since the effect of supersymmetry is precisely to couple together the different spins in the multiplet.

The reduced states are found to be of the form $|p\ j_0\ \sigma\kappa\lambda;\ \kappa_1\kappa_2;\ \tau\rangle$, where the degeneracy labels (in addition to j_{01}, j_{02}, p_1^2, and p_2^2) are now κ_1 and κ_2, rather than λ_1 and λ_2, and some additional invariants τ. The angular dependences of amplitudes such as $\langle 23|T|1\rangle$ and $\langle 34|T|12\rangle$ are extracted by expanding them in terms of angular functions (representing rotations in the superhelicity basis) and reduced partial-wave amplitudes $\langle 23\|T^{j_0}(s)\|\rangle$ and $\langle 34\|T^{j_0}(s)\|12\rangle$ with total superspin j_0.

In the spin basis, the various reduced partial-wave amplitudes $\langle 23\|A^j(s)\|\rangle$, and $\langle 34\|A^j(s)\|12\rangle$, with total angular momentum j, may be similarly expressed in terms of the $\langle 23\|T^{j_0}(s)\|\rangle$ and $\langle 34\|T^{j_0}(s)\|12\rangle$. Since each supersymmetry multiplet contains four different spin

 Copyright © 1977 American Institute of Physics

components, there are, in principle, $4^3 = 64$, and $4^4 = 256$, different spin channels for the processes $1 \to 2 + 3$ and $1 + 2 \to 3 + 4$ respectively; however, half of these are forbidden by conservation of total superspin (or, equivalently, total angular momentum). The surviving 32 or 128 different types of reduced amplitudes $\langle 23 \| A^j(s) \| \rangle$ and $\langle 34 \| A^j(s) \| 12 \rangle$ are found to be given by just two or eight different types of supersymmetric reduced amplitudes, respectively. (Many of the qualitative features described here are given in Ref. 2c.) These relations play the role of supersymmetric constraints imposed upon the helicity amplitudes. After a Sommerfeld–Watson transformation, and continuation to complex superspin, j_0, the leading singularities of the $T^{j_0}(s)$ are found to make contributions to the high-energy behavior of the various (crossed) spin channels, the residues being given by the different supersymmetric partial-wave series.

In Sec. 2, the reduction of the direct product is carried out, and in Sec. 3, the supersymmetric partial-wave expansions, for the processes $1 \to 2 + 3$ and $1 + 2 \to 3 + 4$, are written down. In Sec. 4, further symmetry constraints imposed on the reduced amplitudes, for parity conservation and identical-particle scattering, are described. Finally, in Sec. 5, the connection between the reduced helicity amplitudes, $\langle 23 \| A^j \| \rangle$, $\langle 34 \| A^j(s) \| 12 \rangle$, and the reduced supersymmetric amplitudes, is established, and the continuation to complex j_0, and high energy behavior is described.

2. REDUCTION OF THE DIRECT PRODUCT

In this section we consider the problem of reducing the direct product of two UIR's of the supersymmetry algebra $(j_{01}, m_1)^+ \times (j_{02}, m_2)^+$, where $p_1^2 = m_1^2 > 0$, and $p_2^2 = m_2^2 > 0$, and $\mathrm{sgn}\,(p_{01}) = +1 = \mathrm{sgn}\,(p_{02})$, into a direct sum of UIR's. We choose initially to describe each factor of the direct product by means of the spin basis, so that the problem is to reduce the product states

$$|j_{01}p_1\sigma_1 j_1\lambda_1;\, j_{02}p_2\sigma_2 j_2\lambda_2\rangle = |j_{01}p_1\sigma_1 j_1\lambda_1\rangle |j_{02}p_2\sigma_2 j_2\lambda_2\rangle$$

into a direct sum of states of the form

$$|j_0 p\sigma j\lambda; \cdots\rangle,$$

where we have indicated that, since each massive UIR requires eight labels for its states (two representation labels, j_0 and m, and six state labels $\mathbf{p}, \sigma, j, \lambda$), the reduced states still lack at least eight degeneracy labels.

The total supertranslation, momentum, and angular momentum generators for the reduced states are defined by

$$\begin{aligned} &S_\alpha = S_\alpha^1 + S_\alpha^2, \quad P_\mu = P_\mu^1 + P_\mu^2, \\ &J_{\mu\nu} = J_{\mu\nu}^1 + J_{\mu\nu}^2, \end{aligned} \tag{2.1}$$

where we must assume

$$\{S_\alpha^1, S_\beta^2\} = 0, \tag{2.2}$$

in order to have

$$\{S_\alpha, S_\beta\} = -(\not{P}C)_{\alpha\beta}. \tag{2.3}$$

For consistency it is also necessary to adopt the definition[2c]

$$S_\alpha^2 |1; 2\rangle = (-1)^{2j_1} |1; 2'\rangle \langle 2' | S_2^g | 2\rangle$$

for the generators S_α^2 acting in the product space.

With these definitions, by considering the translation invariance of the product

$$\langle j_0 p \cdots | j_{01} p_1 \cdots ;\, j_{02} p_2 \cdots \rangle \propto \delta^4(p - p_1 - p_2), \tag{2.4}$$

we may assume $\mathrm{sgn}\,(p_0) = +1$ and

$$p = p^1 + p^2. \tag{2.5}$$

In general,

$$\begin{aligned} &p_\mu^1 = m^1 (\cosh\zeta_1,\, \sinh\zeta_1 \hat{\mathbf{n}}), \\ &\hat{\mathbf{n}} = (\sin\theta_1 \cos\phi_1,\ \sin\theta_1 \sin\phi_1,\ \cos\phi_1) \end{aligned} \tag{2.6}$$

with a similar parametrization for p_μ^2, subject to Eq. (2.5). However, by considering the Lorentz invariance of (2.4), it is sufficient to consider the particular choice of orientations given by

$$\begin{aligned} p^{1\mu} &= m^1 (\cosh\zeta_1,\, 0,\, 0,\, +\sinh\zeta_1), \\ p^{2\mu} &= m^2 (\cosh\zeta_2,\, 0,\, 0,\, -\sinh\zeta_2), \\ p &= M(1, 0, 0, 0), \end{aligned} \tag{2.7}$$

where $s = M^2 = (p_1 + p_2)^2 > 0$ is the total energy in the center-of-mass frame. With this choice, we see that $\mathbf{p}^1$ points in the $+z$ direction, and $\mathbf{p}^2$ in the $-z$ direction, in the rest frame of $p_1 + p_2$. The Clebsch–Gordan coefficients for arbitrary orientations of p, p^1, and p^2 may be obtained from those for this special orientation by applying the appropriate Lorentz transformation.

Before proceeding we must settle the question of the degeneracy labels for the reduced states. Adopting the labeling suggested above, it is readily verified using the commutation relations and the definitions of the labels (Appendix A) that

$$[j_0, j_1] \neq 0, \quad [j_0, \lambda_1] \neq 0, \quad [j_0, \sigma_1] \neq 0,$$

so that, since j_0 must be used as a representation label for the reduced states, we are forced to abandon the use of $j_1, j_2, \lambda_1, \lambda_2$, and σ_1, σ_2 as degeneracy labels, and look for other invariant combinations. Two such are $\lambda_1 - \lambda_2$ and $\sigma^1 + \sigma^2$, but counting the labels j_{01}, j_{02}, m_1 and m_2, which are good degeneracy labels, we still lack two labels.

Since we can in any case no longer retain the usual labels for the reduced states, we may also consider the possibility of choosing different bases for the product states, such that the degeneracy problem is simplified. As suggested in Sec. 1, the philosophy is that the results of the supersymmetric analysis may always be reinterpreted in terms of the physical spin basis by transforming back to the latter at the end.

This approach does indeed lead to a more appropriate basis for the supersymmetric UIR's. We replace j, given by the eigenvalue of $W_\mu W^\mu$, by the "superhelicity" κ, defined, in analogy with ordinary helicity, by (see Ap-

pendix B) the eigenvalue of

$$K_0^{\pm}/|\mathbf{P}| = P_0^{-1}\mathbf{P}\cdot\mathbf{K}^1/|\mathbf{P}| = \Lambda - \Sigma_0^{\pm}/|\mathbf{P}| = \Lambda - \Delta, \tag{2.8}$$

where $\Lambda = |\mathbf{P}|^{-1}W_0 = P_0^{-1}|\mathbf{P}|^{-1}\mathbf{P}\cdot\mathbf{J}$. In the rest frame, we take instead the definition (again in analogy with ordinary helicity)

$$-K_3^{\pm}/P_0 = J_3 + \Sigma_3^{\pm}/P_0 = \Lambda - \Delta. \tag{2.9}$$

It is readily found that the eigenvalues δ of Δ are $\delta = 0, \pm\frac{1}{2}$, and that κ, the superhelicity, takes the values

$$\kappa = -j_0, -j_0 + 1, \ldots, j_0,$$

where

$$\lambda = \kappa + \delta \quad \text{and} \quad \delta = \pm\tfrac{1}{2} \mp |\sigma|. \tag{2.10}$$

The superhelicity basis with the labels σ, δ, κ admits a very convenient graphical notation. We represent the pair (σ, δ) by a symbol $+$, for example [in view of (2.10)],

$$(+\tfrac{1}{2}, 0) \equiv +, \qquad (0, \tfrac{1}{2}) \equiv +,$$

and so forth. We may therefore represent basis states in the form $|j_0\ p + \kappa\rangle$, where $+ = +, +, +$, or $+$. For such a state, the values of σ, λ, and κ may be visualized as in Fig. 1, where $\kappa = -j_0, \ldots, j_0$ is plotted against $\sigma = 0, \pm\frac{1}{2}$. The center of the $+$ gives the value of κ, and the dot gives the values of λ and σ. Clearly there are $4(2j_0 + 1)$ helicity states for superspin j_0.[2b]

The utility of the superhelicity basis lies in the fact [Eq. (A9)] that κ is invariant under supertranslations, so that the latter operate only on the symbol $+$. In the rest frame, their matrix elements have a very simple form in this basis, and correspondingly, in an arbitrary frame, the S_α are best written in terms of auxiliary operators $R_\kappa^{\pm}$, with $\kappa = \pm\frac{1}{2}$, which also have a simple form in the superhelicity basis. For conciseness, a summary of the superhelicity basis, and the matrix elements of the $R_\kappa^{\pm}$, and their definitions, are given in Appendix B, as well as another method of constructing the states in the basis.

We note finally that the basis transformation between the superhelicity basis and the spin basis for given p and j_0 is

$$|+\kappa\rangle = |\tfrac{1}{2}j_0\kappa\rangle, \qquad |+\kappa\rangle = |-\tfrac{1}{2}j_0\kappa\rangle,$$

$$|+\kappa\rangle = -\left(\frac{j_0 - \kappa + 1}{2j_0 + 1}\right)^{1/2}|0\ \ j_0 + \tfrac{1}{2}\ \ \kappa - \tfrac{1}{2}\rangle + \left(\frac{j_0 + \kappa}{2j_0 + 1}\right)^{1/2}|0\ \ j_0 - \tfrac{1}{2}\ \ \kappa - \tfrac{1}{2}\rangle, \tag{2.12}$$

$$|+\kappa\rangle = \left(\frac{j_0 + \kappa + 1}{2j_0 + 1}\right)^{1/2}|0\ \ j_0 + \tfrac{1}{2}\ \ \kappa + \tfrac{1}{2}\rangle + \left(\frac{j_0 - \kappa}{2j_0 + 1}\right)^{1/2}|0\ \ j_0 - \tfrac{1}{2}\ \ \kappa + \tfrac{1}{2}\rangle.$$

The labeling problem for the reduced states in the direct product is now solved, since we find in the superhelicity basis that κ^1, κ^2 are good labels, so that the degeneracy labels are

$$j_{01}m_1 j_{02}m_2;\ \kappa^1\kappa^2;\ \sigma^1 + \sigma^2\ \ \lambda^1 - \lambda^2 = 8 \tag{2.13}$$

with the other degeneracy labels, suggested above, included. However, the last two labels, while Lorentz-invariant, are not supertranslation invariant, since

$$[\sigma^1 + \sigma^2, (S_\pm)_a] = \pm\tfrac{1}{2}(S_\pm)_a, \qquad [\lambda^1 - \lambda^2, (S_\pm)_a] = a(S_\pm)_a, \tag{2.14}$$

where $a = (1, 2) \equiv (-\frac{1}{2}, +\frac{1}{2})$. On the other hand, in the direct product algebra $\mathscr{S}_1 \times \mathscr{S}_2$, the number of labels commuting with, for example, the algebra $\mathscr{S}_1$, is ten, namely the two representation labels (Casimirs) of $\mathscr{S}_1$, and any of the eight independent labels of $\mathscr{S}_2$. We might therefore expect a similar accounting to be valid also if we choose to reduce $\mathscr{S}_1 \times \mathscr{S}_2$ with respect to the diagonal generators $\mathscr{S}_1 + \mathscr{S}_2$, namely the total supertranslation generators, momentum, and angular momentum; we therefore seek to replace $\sigma^1 + \sigma^2$ and $\lambda^1 - \lambda^2$ by invariant combinations.

Such combinations may, indeed, by constructed: for, if from the supertranslation generators S_α^1, and S_α^2, which anticommute, we construct some nontrivial linear combination, then there exists another linear combination which anticommutes with the first. We define, corre sponding to the total supertranslation

$$S_\alpha = S_\alpha^1 + S_\alpha^2$$

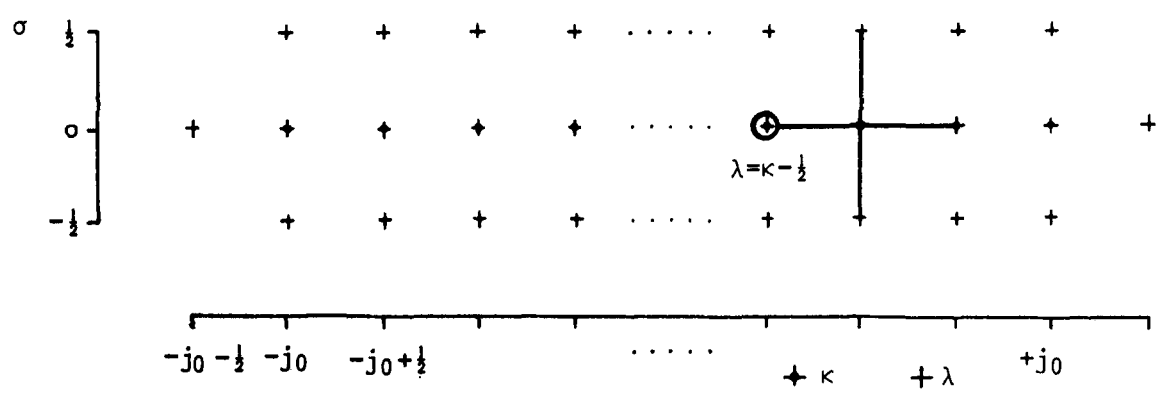


FIG. 1. Visualization of helicity states in the superhelicity basis.

the "relative supertranslation"

$$\tilde{S}_\alpha = (\not{P}^1)^{-1} S^1_\alpha - (\not{P}^2)^{-1} S^2_\alpha, \tag{2.15}$$

(which is anti-Majorana), and observe that

$$\{S_\alpha, \tilde{S}_\beta\} = 0, \quad \{\tilde{S}_\alpha, \tilde{S}_\beta\} = -(\not{\tilde{P}} C)_{\alpha\beta}, \tag{2.16}$$

where

$$\tilde{P}_\mu = \frac{1}{m_1^2} P^1_\mu + \frac{1}{m_2^2} P^2_\mu. \tag{2.17}$$

Moreover, $\tilde{S}_\alpha$ is a spinor under total angular momentum, and commutes with $\tilde{P}_\mu$. It follows that we may define labels $\tilde{\sigma}$, $\tilde{\kappa}$, $\tilde{\delta}$ for $\tilde{S}_\alpha$ in precisely the same way as for S_α replacing S_α by $\tilde{S}_\alpha$, and P_μ by $\tilde{P}_\mu$ with

$$\tilde{P}^2 = \tilde{M}^2 = \left(\frac{M}{m_1 m_2}\right)^2. \tag{2.18}$$

We find, for example, $\tilde{\lambda} = \lambda$, $\tilde{j}_0 = j_0$, $\tilde{\kappa} = \lambda - \tilde{\delta}$, and

$$\tilde{\sigma} = \sigma - \sigma^1 - \sigma^2. \tag{2.19a}$$

In general [Eq. (C9)], $\tilde{\delta}$ is a complicated combination of δ_1, δ_2, δ, and $\tilde{\sigma}$, but, as explained in Appendix C, for the reduced states in the center of mass, the formula

$$\tilde{\delta} = \delta^1 - \delta^2 - \delta \tag{2.19b}$$

(which is obviously not Lorentz-invariant) is true, and we shall use this in the following. The definition of the Lorentz- and supertranslation-invariant labels $(\tilde{\sigma}, \tilde{\delta})$ is discussed in detail in Appendix C.

From these considerations, we may adopt the following labels for the reduced states:

$$|\, j_0\, p + \kappa;\ j_{01} m_1 j_{02} m_2 \kappa_1 \kappa_2;\ \mp \,\rangle = 16, \tag{2.20}$$

where $+$ specifies (σ, δ) and $\mp$ specifies $(\tilde{\sigma}, \tilde{\delta})$, as in Eq. (2.11) above. The problem at hand is therefore to specify the Clebsch–Gordan coefficients

$$\langle j_0\, p + \kappa;\ j_{01} m_1 j_{02} m_2 \kappa_1 \kappa_2;\ \mp \,|\, j_{01} p_1 + {}_1\kappa_1;\ j_{02}\, p_2 + {}_2\kappa_2 \rangle, \tag{2.21}$$

which we shall often simplify, omitting inessential labels, to

$$\langle + \kappa;\ \kappa_1 \kappa_2;\ \mp \,|\, + {}_1\kappa_1;\ + {}_2\kappa_2 \rangle. \tag{2.22}$$

The evaluation of the Clebsch–Gordan coefficients is straightforward. Using the operators $(S_\pm)_a$ as shifting operators, in particular [1]

$$[\sigma, (S_\pm)_a] = \pm \tfrac{1}{2}(S_\pm)_a, \quad [\delta, (S_\pm)_a] = a(S_\pm)_a, \tag{2.23}$$

we apply the operators

$$\bar{S} S_\pm = \pm 4 (S_\pm)_1 (S_\pm)_2$$

to the product states, to construct states of total $\sigma = \pm \frac{1}{2}$, which may be further resolved into eigenstates of δ. Once $+$ is given, $\mp$ follows from Eqs. (2.19a) and (2.19b). As usual we find,

$$\Lambda\, |\, + {}_1\kappa_1;\ + {}_2\kappa_2 \rangle = (\lambda_1 - \lambda_2)\, |\, + {}_1\kappa_1;\ + {}_2\kappa_2 \rangle, \tag{2.24}$$

so that by definition

$$\kappa = \lambda - \delta = \kappa_1 - \kappa_2 + \delta_1 - \delta_2 - \delta \equiv \kappa_{12}. \tag{2.25}$$

Hence we construct total eigenstates

$$|\, + \kappa;\ \kappa_1 \kappa_2;\ \mp \,\rangle_t$$

for $\sigma = \pm \frac{1}{2}$; other total eigenstates may be constructed by applying the $(S_\pm)_a$. Notice that these states are not yet eigenstates of j_0.

The distribution of such states $|\, + \kappa;\ \mp \,\rangle$ is plotted against $\sigma^1 + \sigma^2$ and $\delta^1 - \delta^2$ in Fig. 2., together with the $|\, + {}_1 + {}_2 \rangle$ states with which they couple. For example, the state $|\, + + \,\rangle$ has $\sigma^1 + \sigma^2 = 0$ and $\delta^1 - \delta^2 = 0$ and so is coupled to $|\, + ;\ \mp \,\rangle$, $|\, + ;\ \mp \,\rangle$, $|\, + ;\ \mp \,\rangle$, and $|\, + , \mp \,\rangle$. Notice that, at this stage, the counting of states is consistent: The sixteen $|\, + {}_1 + {}_2 \rangle$ states, for fixed κ_1 and κ_2, have been resolved into sixteen $|\, + \ \mp \,\rangle$ states.

These considerations already allow us to deduce the Clebsch–Gordan series. For fixed κ_1 and κ_2, but allowing the values $+_1$ and $+_2$ to vary, we have

$$\mathscr{D}_{j_{01}\kappa_1} \times \mathscr{D}_{j_{02}\kappa_2} = \sum_{j_0 = |\kappa_1 - \kappa_2 - 1/2|}^{\infty} \mathscr{D}_{j_0} + \sum_{j_0 = |\kappa_1 - \kappa_2|}^{\infty} \mathscr{D}^2_{j_0} + \sum_{j_0 = |\kappa_1 - \kappa_2 + 1/2|}^{\infty} \mathscr{D}_{j_0}, \tag{2.26}$$

giving for example for $j_{01} = j_{02} = 0$

$$0 \times 0 = 0^2 + \tfrac{1}{2}^2 + 1^2 + (3/2)^2 + \cdots,$$

which is to be compared with the corresponding series for the Poincaré group, for $j_1 = j_2 = 0$

$$0 \times 0 = 0 + 1 + 2 + 3 + \cdots.$$

In general, for a given state, $|\, + {}_1 + {}_2 \rangle$, there will be one or more different UIR's with total superspin j_0 coupling to it, distinguished by their values of $\mp$ (Fig. 2). Overall there will be four such UIR's, and the projections of the following states, onto superspin j_0, are each contained in a different $\mp$ representation,

$$\begin{aligned}
&|\, + \kappa;\ \mp \,\rangle_t = |\, + + \,\rangle, \\
&|\, + \kappa - \tfrac{1}{2};\ \mp \,\rangle_t = \left(\frac{m_2}{M}\right)^{1/2} e^{\zeta_2/2} |\, + + \,\rangle \\
&\qquad\qquad + i \left(\frac{m_1}{M}\right)^{1/2} e^{-\zeta_1/2} |\, + + \,\rangle, \\
&|\, + \kappa + \tfrac{1}{2};\ \mp \,\rangle_t = \left(\frac{m_2}{M}\right)^{1/2} e^{\zeta_2/2} |\, - + \,\rangle \\
&\qquad\qquad + i \left(\frac{m_1}{M}\right)^{1/2} e^{-\zeta_1/2} |\, + - \,\rangle, \\
&|\, + \kappa;\ \mp \,\rangle_t = |\, + + \,\rangle, \quad \text{where } \kappa = \kappa_1 - \kappa_2.
\end{aligned} \tag{2.27}$$

The distribution of these special states is indicated in Fig. 2. Their projections onto total superspin j_0 are easily obtained: Since $|\sigma| = \frac{1}{2}$ and $j = j_0 \pm \frac{1}{2} \mp |\sigma| = j_0$, the superspin coincides with the spin, and the projections simply involve the Clebsch–Gordan coefficients for the Poincare group, on passing back to the spin basis on the right-hand sides of Eqs. (2.27). Consider for example

$$|\, + ;\ \mp \,\rangle_t = \alpha |\, + + \,\rangle + i\beta |\, - + \,\rangle.$$

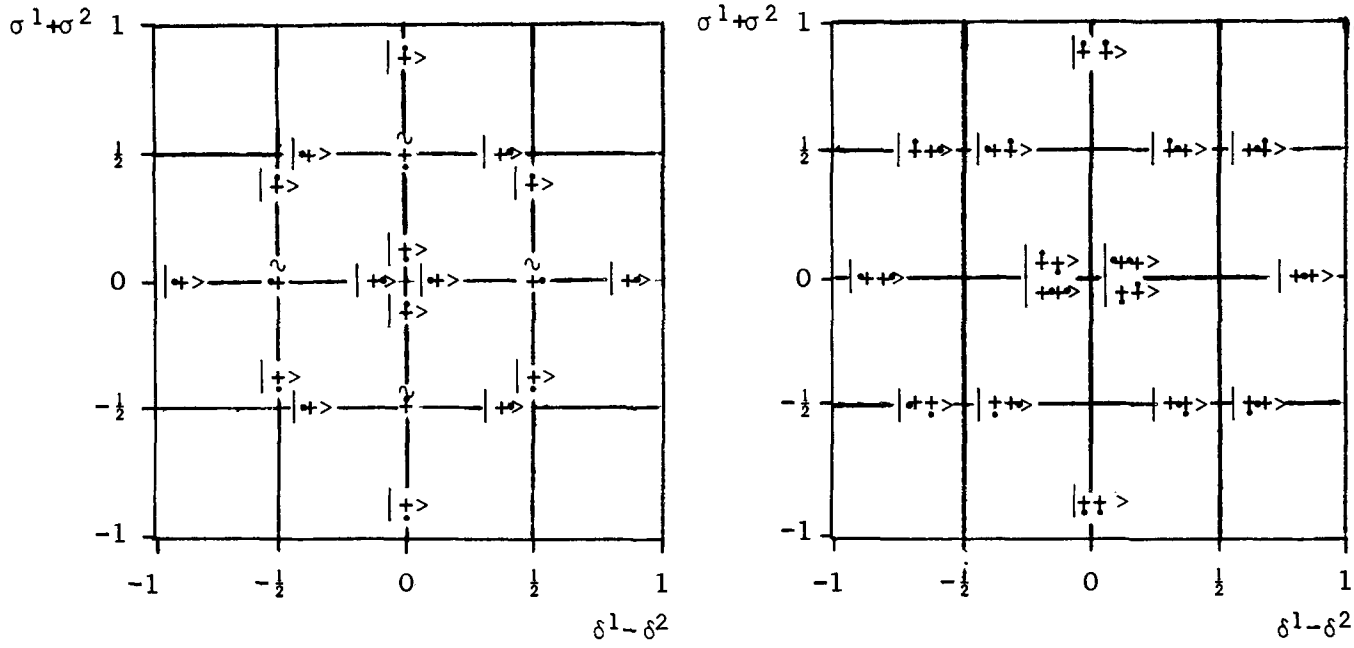


FIG. 2. Plot of the distribution of $|+\ \mp\rangle$ and $|+_1+_2\rangle$ against $\sigma^1+\sigma^2$ and $\delta^1-\delta^2$.

If the projection onto spin (= superspin) $-j_0$ is denoted by Γ, then

$$\Gamma|+;\ \mp\cdot\rangle_t \propto |+;\ \mp\cdot\rangle,$$

and, defining

$$D \equiv (\alpha^2\langle +\cdot + |\Gamma| +\cdot +\rangle + \beta^2\langle + \cdot+ |\Gamma| + \cdot+\rangle)^{1/2},$$

we find that

$$|+;\ \mp\cdot\rangle = \frac{\Gamma}{D}|+;\ \mp\cdot\rangle_t,$$

and

$$\begin{aligned}\langle +\cdot +|+\ \mp\cdot\rangle &= \frac{\alpha}{D}\langle +\cdot +|\Gamma|+\cdot +\rangle,\\ \langle + \cdot+|+\ \mp\cdot\rangle &= i\frac{\beta}{D}\langle + \cdot+|\Gamma|+\cdot+\rangle.\end{aligned} \tag{2.28}$$

The phase of $|+;\ \mp\cdot\rangle$ is also fixed by this definition. The second factors on the right-hand sides involve linear combinations of Poincaré group Clebsch–Gordan coefficients; we take these to be

$$\langle j\lambda;\ \lambda_1\lambda_2|j_1\lambda_1;\ j_2\lambda_2\rangle = \delta_{\lambda,\lambda_1-\lambda_2}(2j+1), \tag{2.29}$$

where $p_1 + p_2 = p = \hat{p}$, $j \geq |\lambda_1 - \lambda_2|$, and where p_1 and p_2 are as in Eq. (2.7).

The combinations needed in Eqs. (2.27) and (2.28), are given in Appendix D, where notations for the various kinematical factors, occurring in the Clebsch–Gordan coefficients, are introduced. In this notation, and using the method which led to Eqs. (2.28), we have, for example,

$$\begin{aligned}\langle +_1+_2|+;\ \mp\cdot\rangle &= \delta_{\kappa,\kappa_1-\kappa_2+1/2}(2j_0+1)^{1/2}\frac{N_+(\kappa_1)}{D_{++}(\kappa_1,\kappa_2)},\\ \langle +_1+_2|+;\ \mp\cdot\rangle &= \delta_{\kappa,\kappa_1-\kappa_2}(2j_0+1).\end{aligned} \tag{2.30}$$

Once the Clebsch–Gordan coefficients for these four "highest (or lowest) weight" vectors $|+\ \mp\cdot\rangle$, $|+\ \mp\cdot\rangle$, $|+\ \mp\cdot\rangle$, and $|+\ \mp\cdot\rangle$ have been found, the remainder are determined by supersymmetry, by acting on each side with the shift operators $(S_\pm)_a$. Thus, for example,

$$|+\rangle = i(2/M)^{1/2}(S_+)_1|+\rangle$$

and

$$(S_+)_1^\dagger = (S_-)_2 = (S_-^1)_2 + (S_-^2)_2$$

so

$$\langle +_1+_2|+\ \mp\cdot\rangle = i\left(\frac{2}{M}\right)^{1/2}\{\langle +\ \mp\cdot\ |(S_-^1)_2 + (S_-^2)_2|+_1+_2\rangle\}^*, \tag{2.31}$$

where we have used the matrix elements of $(S_\pm)_a$ in the rest frame [Eqs. (A12) and (B3)].

Using this method, the remaining Clebsch–Gordan coefficients (altogether 36 nonzero ones) may be written down. They are given in Table I, which is to be consulted in conjunction with Fig. 2.

As mentioned above, the Clebsch–Gordan coefficients for an arbitrary frame (and in particular, for p in the rest frame, and for arbitrary orientations of p_1 and p_2) are related to the above coefficients in the special frame, by Lorentz transformations. In the superhelicity basis, these have the form

$$U(\Lambda)|p\ +\rangle = \sum_{\kappa',+'}|\Lambda p\ +'\rangle\ \{+'\ +\}^{j_0}_{\kappa'\kappa}(\hat{\Lambda}), \tag{2.32}$$

where

$$\hat{\Lambda} = L^{-1}_{\Lambda p}\Lambda L_p,$$

and where the $\sum_{+'}$ must be such that $\sigma' = \sigma$. Hence there are only six nonzero $\{+'\ +\}^{j_0}$ matrices, namely $\{+\ +\}$,

TABLE 1. Clebsch-Gordan coefficients. A common factor of $x(2j_0+1)^{1-|\delta|}(-1)^{4|\delta|j_{01}}\delta_{\kappa,\kappa_1-\kappa_2+\delta}$ is understood.

$$\{++|+\}=1$$

$$\{+-|+\}=\frac{m_1}{M}\left(\frac{\rho+1}{\rho}\right)^{\frac{1}{2}} \qquad \{++|+\}=\left(\frac{\rho}{\rho+1}\right)^{\frac{1}{2}}$$

$$\{--|+\}=-i\frac{m_2}{M}(\rho+1)^{\frac{1}{2}} \qquad \{+-|+\}=-i(\rho+1)^{-\frac{1}{2}}$$

$$\{-+|-\}=\frac{m_2}{M}D_{-+} \qquad \{-+|-\}=\frac{N_+(1)}{D_{++}}$$

$$\{+-|-\}=-i\frac{m_1}{M}\rho^{-\frac{1}{2}}D_{-+} \qquad \{+-|-\}=i\rho^{\frac{1}{2}}\frac{N_-(2)}{D_{++}}$$

$$\{++|++\}=i(\rho+1)^{-\frac{1}{2}}\frac{(D_{--})^2}{D_{-+}}$$

$$\{+-|++\}=\frac{m_1}{M}\left(\frac{\rho+1}{\rho}\right)^{\frac{1}{2}}\frac{N_-(1)}{D_{-+}} \qquad \{++|++\}=\frac{m_2}{M}\left(\frac{\rho+1}{\rho}\right)^{\frac{1}{2}}\rho\frac{N_-(2)}{D_{++}}$$

$$\{--|++\}=\frac{m_2}{M}\left(\frac{\rho+1}{\rho}\right)^{\frac{1}{2}}\frac{N_+(2)}{D_{-+}} \qquad \{+-|++\}=\frac{m_1}{M}\left(\frac{\rho+1}{\rho}\right)^{\frac{1}{2}}\frac{N_+(1)}{D_{++}}$$

$$\{-+|++\}=i\frac{m_2}{M}(\rho+1)^{\frac{1}{2}}\frac{N_-(1)}{D_{-+}} \qquad \{-+|++\}=-i\frac{m_1}{M}(\rho+1)^{\frac{1}{2}}\frac{N_-(2)}{D_{++}}$$

$$\{+-|++\}=-i\frac{m_1}{M}(\rho+1)^{\frac{1}{2}}\frac{N_+(2)}{D_{-+}} \qquad \{+-|++\}=i\frac{m_2}{M}(\rho+1)^{\frac{1}{2}}\frac{N_+(1)}{D_{++}}$$

$$\{++|+-\}=-i(\rho+1)^{-\frac{1}{2}}\frac{(D_{+-})^2}{D_{++}}$$

$$\{--|+\}=\frac{m_1}{M}\left(\frac{\rho+1}{\rho}\right)^{\frac{1}{2}}\rho^{-\frac{1}{2}}$$

$$\{+-|+\}=i\frac{m_2}{M}\rho^{\frac{1}{2}}$$

$$\{--|-\}=\frac{m_2}{M}$$

$$\{+-|-\}=\frac{m_1}{M}$$

$$\{+-|+\}=-i\frac{m_1}{M}\rho^{-\frac{1}{2}}$$

$$\{--|+\}=i\frac{m_2}{M}\rho^{\frac{1}{2}}$$

$$\{--|-\}=\frac{m_1}{M}$$

$$\{+-|-\}=\frac{m_2}{M}$$

$$\{--|+\}=i\frac{m_2}{M}\rho^{\frac{1}{2}}$$

$$\{+-|+\}=-i\frac{m_1}{M}\rho^{-\frac{1}{2}}$$

$$\{+-|+\}=\frac{N_-(1)}{D_{-+}}$$

$$\{--|+\}=i\rho^{\frac{1}{2}}\frac{N_+(2)}{D_{-+}}$$

$$\{--|+\}=i(\rho+1)^{-\frac{1}{2}}$$

$$\{-+|+\}=\left(\frac{\rho}{\rho+1}\right)^{\frac{1}{2}}$$

$$\{--|-\}=\frac{m_2}{M}D_{++}$$

$$\{+-|-\}=-i\frac{m_1}{M}\rho^{-\frac{1}{2}}D_{++}$$

$$\{-+|+\}=\frac{m_1}{M}\left(\frac{\rho+1}{\rho}\right)^{\frac{1}{2}}$$

$$\{+-|+\}=-i\frac{m_2}{M}(\rho+1)^{\frac{1}{2}}$$

$$\{--|--\}=1$$

$\{+\,-+\}$, $\{+\,++\}$, $\{+\,-+\}$, $\{+\,++\}$, and $\{+\,+\}$. They are easily written down in terms of the ordinary Wigner matrices by using the basis transformations (2.12); for example,

$$\{+'\,+\}^{j_0}_{\kappa'\kappa} = \mathscr{D}^{j_0}_{\kappa'\kappa},$$

$$\{+'\,-+\}^{j_0}_{\kappa'\kappa} = \left(\frac{j_0-\kappa'+1}{2j_0+1}\right)^{1/2}\left(\frac{j_0-\kappa+1}{2j_0+1}\right)^{1/2}\mathscr{D}^{j_0+1/2}_{\kappa'-1/2\,\kappa-1/2} + \left(\frac{j_0+\kappa'}{2j_0+1}\right)^{1/2}\left(\frac{j_0+\kappa}{2j_0+1}\right)^{1/2}\mathscr{D}^{j_0-1/2}_{\kappa'-1/2\,\kappa-1/2}. \tag{2.33}$$

The remaining four matrices are written down in Appendix D, where their symmetry properties are also discussed. They are clearly the analogues for the superhelicity basis of the ordinary Wigner functions in the spin basis, and they will play a crucial role in the supersymmetric partial-wave analysis to be developed.

By using the $\{+'\,+\}^{j_0}$ functions, the Clebsch-Gordan coefficients for arbitrary orientations of the initial momenta p_1 and p_2 may be immediately written down. We consider only the case of azimuthal angle $\phi = \pi$ [Eq. (2.7)], and define

$$|+_1\,+_2\rangle^\theta = e^{-i\theta J_2}|+_1\,+_2\rangle^0, \tag{2.34}$$

where $|+_1\,+_2\rangle^0$ refers to the states with the z axis orientations of p_1 and p_2 which we have been considering. It then follows that

$$\begin{aligned}{}^\theta\langle +_1\,+_2|+\kappa;\mp\rangle &\equiv \{+_1\,+_2|+\mp\}^{j_0}(\theta)\\ &= \textstyle\sum_{+'}{}^0\langle +_1\,+_2|+'\kappa_{12};\mp\rangle\{+'\,+\}^{j_0}_{\kappa_{12}\kappa}(\theta),\end{aligned}$$

where

$$j_0 \geq \max(|\kappa|, |\kappa_{12}|), \qquad \kappa_{12} = \kappa_1 - \kappa_2 + \tilde{\delta}. \tag{2.35}$$

Notice that, in fact, $+'$ is fixed by $+_1$, $+_2$, $\mp$, and Eqs. (2.19), so in Eq. (2.35) there is only one summation label, corresponding to the occurrence of one relevant variable, the angle θ.

Finally we may use the completeness property of the reduced states to expand any state ${}^\theta\langle +_1\,+_2|$ as

$${}^\theta\langle +_1\,+_2| = \sum_{\mp}\sum_{j_0=|\kappa_{12}|}^{\infty}\{+_1\,+_2|+\mp\}^{j_0}(\theta)\langle +\kappa;\mp|. \tag{2.36}$$

3. PARTIAL-WAVE ANALYSIS

We now use the formalism of the last section to analyze the angular dependence of three- and four-particle scattering in the superhelicity basis, by making a partial-wave expansion in terms of the total superspin, j_0. We take

$$i\langle p';f|S-1|p;i\rangle = (2\pi)^4\delta^4(p'-p)\langle f|T(s)|i\rangle, \tag{3.1}$$

where $s = p^2$, and as before adopting the center-of-mass frame so that the total 4-momentum is $p_\mu = \hat{p}_\mu$. Assuming T to be invariant under supertranslations as well as Lorentz transformations, we have by a generalization of the Wigner–Eckart theorem:

$$\begin{aligned}&\langle +'\kappa';\kappa_2\kappa_3;\mp'|T|+_1\kappa_1\rangle\\ &\quad = \delta_{j_0'j_0}\delta_{+'+_1}\delta_{\kappa'\kappa_1}\langle\kappa_2\kappa_3;\mp'\|T^{j_0}(s)\|\rangle,\\ &\langle +'\kappa';\kappa_3\kappa_4;\mp'|T|+\kappa;\kappa_1\kappa_2;\mp\rangle\\ &\quad = \delta_{j_0'j_0}\delta_{+'+}\delta_{\kappa'\kappa}\langle\kappa_3\kappa_4;\mp'\|T^{j_0}(s)\|\kappa_1\kappa_2;\mp\rangle,\end{aligned} \tag{3.2}$$

for the decay process $1 \to 2 + 3$ and the scattering $1 + 2 \to 3 + 4$, respectively.

The partial-wave expansion is now obtained by writing the two-particle states $|+_1+_2\rangle^\theta$ in terms of reduced states $|+\kappa;\mp\rangle$,

$$\begin{aligned}|+_1+_2\rangle^\theta = \sum_{\mp}\sum_{j_0=|\kappa_{12}|}^{\infty}\sum_{\kappa=-j_0}^{j_0}&|+\kappa;\kappa_1\kappa_2;\mp\rangle\\ &\times\langle+\kappa;\kappa_1\kappa_2;\mp|+_1+_2\rangle^\theta,\end{aligned}$$

or

$$\begin{aligned}|+_1+_2\rangle^\theta = \sum_{\mp}\sum_{j_0=|\kappa_{12}|}^{\infty}\sum_{\kappa}&|+\kappa;\kappa_1\kappa_2;\mp\rangle\\ &\times\{+;\mp|+_1+_2\}^{j_0}(-\theta),\end{aligned} \tag{3.3}$$

where $\kappa_{12} = |\kappa_1 - \kappa_2 + \tilde{\delta}|$. Note that in this expansion $+$ is given once $\mp$ is specified, since

$$\tilde{\sigma} = \sigma - \sigma_1 - \sigma_2, \qquad \tilde{\delta} = \delta_1 - \delta_2 - \delta, \tag{2.19}$$

the latter being true only in the rest frame.

The angular dependence for a decay process $1 \to 2 + 3$ is therefore

$$\begin{aligned}{}^\theta\langle +_2+_3|T|+_1\kappa_1\rangle = \sum_{\mp}&\{+_2+_3|+\mp\}^{j_{01}}_{\kappa_{23}\kappa_1}(\theta)\\ &\times\langle\kappa_2\kappa_3;\mp\|T^{j_{01}}\|\rangle,\end{aligned} \tag{3.4}$$

where $j_{01} \geq |\kappa_{23}|$. For the process $1 + 2 \to 3 + 4$ we have similarly

$$\begin{aligned}{}^\theta\langle +_3+_4|T|+_1+_2\rangle^0 = \sum_{\mp}\sum_{j_0=J_0}^{\infty}&\{+_3+_4|\mp'\mp|\\ &+_1+_2\}^{j_0}(\theta)\langle\kappa_3\kappa_4;\mp'\|T^{j_0}\|\kappa_1\kappa_2;\mp\rangle,\end{aligned} \tag{3.5}$$

where $2|\kappa_{12}| \equiv 2|\kappa_{34}|$ modulo 2 (from conservation of total angular momentum), $J_0 = \max(|\kappa_{12}|, |\kappa_{34}|)$, and we define

$$\begin{aligned}&\{+_3+_4|\mp'\mp|+_1+_2\}^{j_0}(\theta)\\ &\quad = {}^0\langle +_3+_4|+'\kappa_{34};\mp'\rangle\{+'+\}^{j_0}_{\kappa_{34}\kappa_{12}}(\theta)\\ &\qquad\times\langle+\kappa_{12};\mp|+_1+_2\rangle,\end{aligned} \tag{3.6}$$

where $+'$ and $+$ are fixed by $\mp'$ and $\mp$ as in Eq. (2.19). The angular functions of Eqs. (3.4) and (3.5) are once again linear combinations of $d^{j_0}(\theta)$ and $d^{j_0\pm(1/2)}(\theta)$, with coefficients depending upon κ_1, κ_2, κ_3, and κ_4, and invariants such as m_1 and s. In these expansions $(s)^{1/2} = M$ and θ are the center-of-mass energy, and scattering angle, respectively.

We defer for the moment the questions of convergence, inversion, and interpretation of the expansions (3.4) and (3.5). These will be dealt with in Sec. 5, where we consider the relationship of the supersymmetric reduced amplitudes

$\langle \kappa_3\kappa_4; \mp' \| T^{j_0} \| \kappa_1\kappa_2; \mp \rangle$ to the ordinary partia-wave amplitudes $\langle \lambda_3\lambda_4 \| A^j \| \lambda_1\lambda_2 \rangle$, with total angular momentum j. In the meantime we investigate the further constraints imposed upon our reduced amplitudes by symmetries such as parity.

4. SYMMETRY PROPERTIES OF THE T MATRIX

In the previous section we exploited the invariance of T under Lorentz transformations and supertranslations, namely

$$[T,\ J_{\mu\nu}]=0=[T,\ S_\alpha],$$

to separate the angular dependence of the amplitudes and derive the partial-wave series in the reduced amplitudes. However, the latter, $8\prod_{i=1}^4(2j_{0_i}+1)\langle \kappa_3\kappa_4; \mp' \| T^{j_0}(s) \| \kappa_1\kappa_2; \mp \rangle$'s for the process $1+2\to 3+4$, and the $2\prod_{i=1}^3(2j_{0_i}+1)\langle \kappa_2\kappa_3; \mp \| T^{j_0}(s) \| \rangle$ for the process $1\to 2+3$, are further constrained if we assume that T has additional symmetry properties. In particular, we shall investigate here the consequences of parity conservation,

$$U_{\mathbb{P}}^{-1}\, T\, U_{\mathbb{P}} = T. \tag{4.1}$$

For the purposes of this paper we shall assume [2b] that

$$U_{\mathbb{P}}\, S_\alpha U_{\mathbb{P}}^{-1} = (i\gamma_0 S)_\alpha \quad \text{or} \quad U_{\mathbb{P}}\,(S_\mp)_\alpha U_{\mathbb{P}}^{-1} = i(S_\mp)_\alpha. \tag{4.2}$$

In Ref. 1 it was shown how to define parity acting in a UIR of supersymmetry, in the spin basis. Since σ is a pseudoscalar, the states of definite spin and parity are

$$|pj_0\lambda\rangle^\pm = (1/\sqrt{2})(|p\ \tfrac{1}{2}\ j_0\ \lambda\rangle \pm |p\ -\tfrac{1}{2}\ j_0\ \lambda\rangle),\ |p\,0\ j_0\ \pm\tfrac{1}{2}\ \lambda\rangle, \tag{4.3}$$

and we assume that

$$U_{\mathbb{P}}|pj_0\lambda\rangle^\pm = \pm\eta(-1)^{-j_0}|\mathbb{P}p\ j_0 - \lambda\rangle. \tag{4.4}$$

Furthermore, since for these $\sigma = \pm\frac{1}{2}$ states the superhelicity equals the helicity [Eq. (2.10)], this implies

$$\begin{aligned} U_{\mathbb{P}}|p\ +\ \kappa\rangle &= \eta(-1)^{-j_0}|\mathbb{P}p\ +\ -\kappa\rangle, \\ U_{\mathbb{P}}|p\ +\ \kappa\rangle &= \eta(-1)^{-j_0}|\mathbb{P}p\ +\ -\kappa\rangle. \end{aligned} \tag{4.5}$$

The behavior under parity of the $\sigma = 0$ states $|p\ +\ k\rangle$ and $|p\ +\ k\rangle$ is found by using, for example,

$$|\ +\ \rangle = i(2/M)^{1/2}(S_+)_2|+\ \rangle, \tag{4.6}$$

and applying $U_{\mathbb{P}}$, using Eqs. (4.2) and (4.5) above.

The result is

$$U_{\mathbb{P}}|p\ +\ \kappa\rangle = -\eta(-1)^{-j_0}|\mathbb{P}p\ +\ -\kappa\rangle, \tag{4.7}$$

and similarly for $|p\ +\ k\rangle$. We may summarize the results as

$$U_{\mathbb{P}}|p;\ \sigma\kappa\delta\rangle = \eta(-1)^{2|\delta|-j_0}|\mathbb{P}p;\ -\sigma-\kappa-\delta\rangle, \tag{4.8}$$

and this corresponds for $\sigma = 0$ in the spin basis to

$$\begin{aligned} &U_{\mathbb{P}}|p;\ 0\ j_0\pm\tfrac{1}{2}\ \lambda\rangle \\ &\quad = i\eta(-1)^{-j_0\mp(1/2)}|\mathbb{P}p;\ 0\ j_0\pm\tfrac{1}{2}\ -\lambda\rangle, \end{aligned} \tag{4.9}$$

as in Ref. 2b, so that the J^P classification of states is $j_0^{\pm\eta}$, $(j_0\pm\frac{1}{2})^{i\eta}$.

The action of $U_{\mathbb{P}}$ on the reduced states $|+;\ \kappa_1\kappa_2;\ \mp\rangle$ is derived similarly. We consider firstly the projections $|pj\lambda;\ +_1+_2\rangle$ of states $|\ +_1\kappa_1\ +_2\kappa_2\rangle$ onto total spin j (not superspin). These projections may be expanded in terms of the $|pj\lambda; j_1\lambda_1; j_2\lambda_2\rangle$ by means of the basis transformations, Eq. (2.12); for the latter states we may use[3]

$$\begin{aligned} U_{\mathbb{P}}|pj\lambda;\ j_1\lambda_1 j_2\lambda_2\rangle &= \eta_1\eta_2(-1)^{j_1+j_2-2j} \\ &\quad\times |p\ j\ -\lambda;\ j_1\ -\lambda_1\quad j_2\ -\lambda_2\rangle. \end{aligned} \tag{4.10}$$

This procedure results in

$$\begin{aligned} U_{\mathbb{P}}|pj\lambda;\ +_1\kappa_1\ +_2\kappa_2\rangle &= \eta_1\eta_2(-1)^{j_{01}+j_{02}-2j} \\ &\quad\times |\mathbb{P}p\ j\quad -\lambda;\ \mathbb{P}+_1\quad -\kappa_1\quad \mathbb{P}+_2\quad -\kappa_2\rangle, \end{aligned} \tag{4.11}$$

where, for example, $\mathbb{P}+=+, \mathbb{P}+-=-+$, and so on. Finally the states $|p+\kappa;\ \kappa_1\kappa_2;\ \mp\rangle$ may be expressed as a sum of spin $-j_0$ and $j_0\pm\frac{1}{2}$ projections on combinations of $|\ +_1\kappa_1\ +_2\kappa_2\rangle$ states [cf. Eq. (2.27)]. The result is

$$\begin{aligned} U_{\mathbb{P}}|p+\kappa;\ \kappa_1\kappa_2;\ \mp\rangle &= \eta_1\eta_2(-1)^{j_{01}+j_{02}-2j_0+2|\delta|} \\ &\quad\times \mathbb{P}p\mathbb{P}+\quad -\kappa;\quad -\kappa_1\quad -\kappa_2;\quad \mathbb{P}\mp\rangle. \end{aligned} \tag{4.12}$$

Equations (4.1), (4.8), and (4.12) give the necessary constraints on the reduced amplitudes (3.2) which follow from parity conservation. The results are simply

$$\begin{aligned} &\langle \| T^{j_{01}}(s) \| \kappa_2\kappa_3;\ \mp\rangle = \eta_1^*\eta_2\eta_3(-1)^{j_{02}+j_{03}-j_{01}} \\ &\quad\times \langle \| T^{j_{01}}(s) \| -\kappa_2\ -\kappa_3;\ \mathbb{P}\mp\rangle \end{aligned}$$

and

$$\begin{aligned} &\langle \kappa_3\kappa_4;\ \mp' \| T^{j_0}(s) \| \kappa_1\kappa_2;\ \mp\rangle \\ &\quad = \eta_3^*\eta_4^*\eta_1\eta_2(-1)^{j_{03}+j_{04}-j_{01}-j_{02}} \\ &\qquad\times \langle -\kappa_3\quad -\kappa_4;\ \mathbb{P}\mp' \| T^{j_0}\| -\kappa_1\quad -\kappa_2;\ \mathbb{P}\mp\rangle. \end{aligned} \tag{4.13}$$

Thus under parity the number of independent amplitudes is reduced to $\prod_{i=2}^3(2j_{0_i}+1)$ for $1\to 2+3$, and to $4\prod_{i=1}^4(2j_{0_i}+1)$ for the process $1+2\to 3+4$.

Clearly it is possible to study analogously other symmetries of the T matrix, such as time reversal invariance, and charge conjugation (crossing). In general the results are no more than the transferral to the superhelicity basis, of the corresponding well-known properties of the amplitudes in the spin basis.

The frequent use of superspin-zero models in the literature lends some interest to the general question of identical-particle states in supersymmetry. In the spin basis it is well known that the physical states with permutation properties appropriate to the connection between spin and statistics are of the form[3]

$$|j\lambda(12)\rangle = (1/\sqrt{2})(|j\lambda; 12\rangle + (-1)^j|j\ -\lambda; 21\rangle)$$

for total spin j. Inspection of these states, for all possible spins $j_0^{\pm\eta}$, $(j_0\pm\frac{1}{2})^{i\eta}$ in the multiplet, with the basis transformations (2.12), shows that they may be expressed in terms of the symmetrized states $|j_0\pm(12)\ \mp\rangle$, with the same basis transformations, if we define

$$|j_0\sigma\delta\kappa(\kappa_1\kappa_2)\bar{\sigma}\bar{\delta}\rangle = (1/\sqrt{2})(|j_0\sigma\delta\kappa;\ \kappa_1\kappa_2\bar{\sigma}\bar{\delta}\rangle + (-1)^{j_0-|\delta|}|j_0\sigma\ -\delta\ -\kappa;\ \kappa_2\kappa_1\ \bar{\sigma}\ -\bar{\delta}\rangle). \quad (4.14)$$

There is one further comment to be made in connection with the conservation of total superspin, with regard to the total number of fermions occurring in the multiparticle states in matrix elements of T, such as $\langle +_3 +_4|T|+_1 +_2\rangle$. Consider for example, $\langle + + |T| + + \rangle$, and suppose the superspins j_{0i} are all integral. The only possible term in the partial-wave expansion, Eq. (3.5), is that involving $\langle \mp \| T^{j_0} \| \mp \rangle$ (Fig. 2.). However, for this case $|\kappa_{12}| = |\kappa_1 - \kappa_2|$ is integral, but $|\kappa_{34}| = |\kappa_3 - \kappa_4 + \frac{1}{2}|$ is half-integral. This means that, in the amplitudes $\langle +'\kappa';\ \mp |T| + \kappa;\ \mp \rangle$ leading to the partial-wave series, we have j_0' half-integral, and j_0 integral, so that, in view of Eq. (2.36), we have

$$\langle +_3 +_4|T|+_1 +_2\rangle = 0.$$

Of course, referring to the spin basis, this is no more than the statement that

$$j_{01} \times j_{02} \nrightarrow (j_{03} \pm \tfrac{1}{2}) \times j_{04}, \quad j_{0i} \text{ integer},$$

and we expect conservation of total superspin, j_0, to be consistent with this. It does, at least, emphasize that supersymmetry is consistent with $\Delta F \equiv 0$ modulo 2, and it devolves upon the model in question to give the correct antifermion assignments to ensure $\Delta F = 0$. This question has received considerable attention in the literature, e.g., Ref. 4. We hope to return to a more complete discussion of it, in the present formalism, at a later time.

5. INTERPRETATION OF THE REDUCED AMPLITUDE

We take up now the questions of interpretation of the supersymmetric reduced amplitudes introduced in the partial-wave expansions, Eqs. (3.4) and (3.5). Our approach is to refer back to the conventional total angular momentum partial-wave expansions,[3] expressing the latter in the superhelicity basis by means of Eq. (2.12).

Consider, for example, a process $1 \to 2 + 3$ in which a spin component of a supersymmetric multiplet with superspin j_{01} decays into components of multiplets with j_{02} and j_{03}. We assume for simplicity $j$that $_{01}$, j_{02}, and j_{03} are integral. We have, for example,[3] in the spin basis,

$$\begin{aligned}&{}^{\theta}\langle 0\ j_{02}+\tfrac{1}{2}\ \lambda_2;\ \tfrac{1}{2}j_{03}\lambda_3|T|0\ j_{01}-\tfrac{1}{2}\ \lambda_1\rangle \\ &\quad = d^{j_{01}}_{\lambda\lambda_1}{}^{-1/2}(\theta)\,\langle \lambda_2\lambda_3 \| A^{j_{01}-1/2}(s)\|\rangle (2j_{01}),\end{aligned} \quad (5.1)$$

where

$$\lambda = \lambda_2 - \lambda_3.$$

On the other hand, using Eqs. (2.12), (2.35), and (D4),

$$\begin{aligned}{}^{\theta}\langle 0\, j_{02}+\tfrac{1}{2}\lambda_2;\ \tfrac{1}{2}j_{03}\lambda_3|T|0\ j_{01}-\tfrac{1}{2}\ \lambda_1\rangle &= \{++|+\mp\}^{j_0}\langle \lambda_2-\tfrac{1}{2}\lambda_3;\ \mp\|T^{j_{01}}(s)\|\rangle \\ &\times [C_1C_2\{+-\}^{j_{01}}_{\lambda+1/2\ \lambda_1+1/2}(\theta) \\ &+ S_1C_2\{+-\}^{j_{01}}_{\lambda-1/2\ \lambda_1-1/2}(\theta)] \\ &- \{++|+\mp\}^{j_0}\langle \lambda_2+\tfrac{1}{2}\lambda_3;\ \mp\|T^{j_{01}}(s)\|\rangle \\ &\times [C_1S_2\{+-\}^{j_0}_{\lambda+1/2\ \lambda_1+1/2}(\theta) \\ &+ S_1S_2\{+-\}^{j_0}_{\lambda+1/2\ \lambda_1-1/2}(\theta)],\end{aligned} \quad (5.2)$$

where $C_1 \equiv C(\lambda_1)$, etc. Using Eqs. (D10), and Table I, we have simply

$$\begin{aligned}&d^{j_{01}-(1/2)}_{\lambda\lambda_1}(\theta)\,\langle \lambda_2\lambda_3\|A^{j_{01}-(1/2)}(s)\|\rangle (2j_{01}) \\ &\quad = d^{j_{01}-1/2}_{\lambda\lambda_1}(\theta)\,(2j_{01}+1) \\ &\quad \times \Big[\frac{m_1}{M}\Big(\frac{\rho+1}{\rho}\Big)^{1/2} C_\lambda S_2\,\langle \lambda_2+\tfrac{1}{2}\lambda_3;\ \mp\|T^{j_{01}}(s)\|\rangle \\ &\quad - \Big(\frac{\rho}{\rho+1}\Big)^{1/2} S_\lambda C_2\,\langle \lambda_2-\tfrac{1}{2}\lambda_3;\ \mp\|T^{j_{01}}(s)\|\rangle\Big]\end{aligned} \quad (5.3)$$

so that $\langle \lambda_2\lambda_3\|A^j(s)\|\rangle$ is given as a linear combination of $\langle \lambda_2 \pm \tfrac{1}{2}\lambda_3\ \mp\|T^{j_{01}}\|\rangle$.

In general, for the decay $1 \to 2 + 3$ there will be $32\prod_{i=1}^{3}(2j_{0i}+1)$ reduced amplitudes $\langle \lambda_2\lambda_3\|A^j(s)\|\rangle$ for total angular momentum j; clearly, by repeating the analysis leading to Eq. (5.3), these may all be expressed in terms of the $2\prod_{i=1}^{3}(2j_{0i}+1)$ reduced amplitudes $\langle \kappa_2\kappa_3;\ \mp\|T^{j_0}(s)\|\rangle$ for total superspin j_0; these relations may alternatively be viewed as helicity constraints on the $\langle \lambda_2\lambda_3\|A^j(s)\|\rangle$ arising from supersymmetry.

The analysis of the scattering process $1 + 2 \to 3 + 4$, for components of supersymmetry multiplets, proceeds similarly. Assuming j_{01}, j_{02}, j_{03} and j_{04} to be integral as before, we have[3]

$$\begin{aligned}&{}^{\theta}\langle 0\, j_{03}+\tfrac{1}{2}\lambda_3;\ \tfrac{1}{2}j_{04}\lambda_4|T|0\, j_{01}+\tfrac{1}{2}\lambda_1;\ -\tfrac{1}{2}j_{02}\lambda_2\rangle \\ &\quad = \sum_{j=J}^{\infty}(2j+1)\,d^j_{\mu\lambda}(\theta)\,\langle \lambda_3\lambda_4\|A^j(s)\|\lambda_1\lambda_2\rangle,\end{aligned} \quad (5.4)$$

where $\mu = \lambda_3 - \lambda_4$, $\lambda = \lambda_1 - \lambda_2$, $J = \max(|\mu|, |\lambda|)$. On the other hand, using Eqs. (2.12),

$$\begin{aligned}&{}^{\theta}\langle 0\, j_{03}+\tfrac{1}{2}\lambda_3;\ \tfrac{1}{2}j_{04}\lambda_4|T|0\, j_{01}+\tfrac{1}{2}\ \lambda_1;\ -\tfrac{1}{2}j_{02}\lambda_2\rangle \\ &\quad = S_1S_3\,{}^{\theta}\langle +\ \lambda_3+\tfrac{1}{2};\ +\ \lambda_4|T|+\ \lambda_1+\tfrac{1}{2};\ +\ \lambda_2\rangle \\ &\quad - C_1S_3\,{}^{\theta}\langle +\ \lambda_3+\tfrac{1}{2};\ +\ \lambda_4|T|+\ \lambda_1-\tfrac{1}{2};\ +\ \lambda_2\rangle \\ &\quad - S_1C_3\,{}^{\theta}\langle +\ \lambda_3-\tfrac{1}{2};\ +\ \lambda_4|T|+\ \lambda_1+\tfrac{1}{2};\ +\ \lambda_2\rangle \\ &\quad + C_1C_3\,{}^{\theta}\langle +\ \lambda_3-\tfrac{1}{2};\ +\ \lambda_4|T|+\ \lambda_1-\tfrac{1}{2};\ +\ \lambda_2\rangle,\end{aligned} \quad (5.5)$$

where, for example, using Eqs. (3.5) and (3.6)

$$\begin{aligned}&{}^{\theta}\langle +\ \lambda_3+\tfrac{1}{2};\ +\ \lambda_4|T|+\ \lambda_1+\tfrac{1}{2};\ +\ \lambda_2\rangle \\ &\quad = \sum_{j_0=j_0}^{\infty}\frac{m_3}{M}\Big(\frac{1+\rho'}{\rho'}\Big)^{\frac{1}{2}}\Big(\frac{\rho}{1+\rho}\Big)^{\frac{1}{2}} \\ &\quad \times (-1)^{2j_{01}+2j_{03}}\{+-\}^{j_0}_{\mu+1/2\ \lambda+1/2}(\theta) \\ &\quad \times \langle \lambda_3+\tfrac{1}{2}\ \lambda_4;\ \mp\|T^{j_0}(s)\|_{,}\lambda_1+\tfrac{1}{2}\ \lambda_2;\ \mp\rangle,\end{aligned} \quad (5.6)$$

where $J_0 = \max(|\mu + \tfrac{1}{2}|, |\lambda + \tfrac{1}{2}|)$, and similar equations for the other components, with different lower limits for the summations. [The identification of $\langle \lambda_3\lambda_4\| A^j(s)\|\lambda_1\lambda_2\rangle$ in terms of $\langle \kappa_3\kappa_4\ \mp'\|T^{j_0}(s)\|\kappa_1\kappa_2\ \mp\rangle$ for

$j = J$, must be treated separately.] By substituting these expressions into Eq. (5.5), and expanding the $\{++\}^{j_0}(\theta)$, $\{+\,+\cdot\}^{j_0}(\theta)$, and so on, in terms of the $d^{j_0\pm1/2}(\theta)$, we obtain a form comparable with the right-hand side of Eq. (5.4). For ease of writing we specialize to the case $m_1 = m_2 = m_3 = m_4 = m$, and $j_{01} = j_{02} = j_{03} = j_{04} = 0$. We then have $\lambda_2 = \lambda_4 = 0$, and the remaining dependence upon λ_3 and λ_1 is given explicitly by

$$\sum_{j=1/2}^{\infty}(2j+1)\,d^j_{\mu\lambda}(\theta)\,A^j_{\lambda_3\lambda_1}(s) = \sum_{j=1/2}^{\infty} d^j_{\mu\lambda}(\theta)\Big(\frac{m}{M}(2j+2)^2 \langle \mp \| T^{j+1/2}(s) \| \tilde{\mp} \rangle F^{j+1/2}_{\lambda_3\lambda_1} + \frac{m}{M}(2j)^2 \langle \mp \| T^{j-1/2}(s) \| \tilde{\mp} \rangle G^{j-1/2}_{\lambda_3\lambda_1}\Big), \tag{5.7}$$

where $G^0_{\lambda_3\lambda_1} \equiv 0$, and

$$F^{j_0}_{\lambda_3\lambda_1} \equiv \Big[\frac{M}{m}\Big(\frac{\rho}{\rho+1}\Big)C_3S_\mu - S_3C_\mu\Big] \times \Big[\frac{m}{M}\Big(\frac{\rho+1}{\rho}\Big)C_1S_\lambda - S_1C_\lambda\Big],$$

$$G^{j_0}_{\lambda_3\lambda_1} = \Big[\frac{M}{m}\Big(\frac{\rho}{\rho+1}\Big)C_3C_\mu + S_3S_\mu\Big] \times \Big[\frac{m}{M}\Big(\frac{\rho+1}{\rho}\Big)C_1C_\lambda + S_1S_\lambda\Big].$$

For example,

$$F^{j_0}_{++} = \Big(\frac{j_0}{2j_0+1}\Big), \quad G^{j_0}_{+-} = \frac{M}{m}\Big(\frac{\rho}{\rho+1}\Big)\Big(\frac{j_0+1}{2j_0+1}\Big).$$

A similar analysis of $\langle \frac{1}{2}j_{03}\lambda_3; \frac{1}{2}j_{04}\lambda_4 \,|\, T \,|\, -\frac{1}{2}j_{01}\lambda_1; \frac{1}{2}j_{02}\lambda_2\rangle$ into its total angular momentum and total super-spin partial waves leads to

$$\sum_{j=J}^{\infty}(2j+1)\,d^j_{\mu\lambda}(\theta)\langle \lambda_3\lambda_4 \| A^j(s) \| \lambda_1\lambda_2\rangle = \sum_{j_0=J}^{\infty}(2j_0+1)^2\Big(\frac{m_2}{M}\Big)d^{j_0}_{\mu\lambda}(\theta) \times \langle \lambda_3\lambda_4; \tilde{\mp} \| T^{j_0}(s) \| \lambda_1\lambda_2; \tilde{\mp}\rangle. \tag{5.8}$$

Obviously a similar identification may be made for any of the amplitudes $\langle\lambda_3\lambda_4 \| A^j(s) \| \lambda_1\lambda_2\rangle$. Allowing for conservation of total angular momentum (or, equivalently, of total superspin), there will be 128 $\prod_{i=1}^4(2j_{0_i}+1)$ such amplitudes. These may all be expressed in terms of the 8 $\prod_{i=1}^4(2j_{0_i}+1)$ reduced amplitudes, $\langle \kappa_3\kappa_4; \tilde{\mp}' \| T^{j_0}(s) \| \kappa_1\kappa_2; \tilde{\mp}\rangle$. Alternatively, these relations may be viewed as constraints on the $\langle \lambda_3\lambda_4 \| A^j(s) \| \lambda_1\lambda_2\rangle$ arising from supersymmetry.

The above also demonstrates the implications of supersymmetry for high-energy behavior of the amplitudes. We extend the analyticity postulates in a natural way, and assume that the S matrix can be continued to complex superspin, j_0, with only isolated sigularities.[5] We shall also assume that the above identifications, such as (in a simplified notation)

$$(2j+1)A(j,s) = (2j+1)^2\,[m/(s)^{1/2}]\,T_{++}(j,s),$$

$$(2j+1)A_{\lambda_3\lambda_1}(j,s) = (2j+2)^2\,[m/(s)^{1/2}]T_{++}(j+\tfrac{1}{2},s)F_{\lambda_3\lambda_1}(j+\tfrac{1}{2},s) + (2j)^2\,[m/(s)^{1/2}]\,T_{++}(j-\tfrac{1}{2},s)G_{\lambda_3\lambda_1}(j-\tfrac{1}{2},s) \tag{5.9}$$

continue to be valid: thus the analyticity postulate applies in particular to continuation in complex angular momentum. This may be performed for each spin channel separately by established techniques (e.g., Ref. 6). Supersymmetry then requires that all of these continued amplitudes be given in terms of a small number of supersymmetric amplitudes.

Thus, in the example which we have been considering, the leading singularities of $T_{++}(j_0, s)$ dominate the high-energy behavior in the crossed channels of both $A(j, s)$ and $A_{\lambda_3\lambda_1}(j, s)$. For example, a pole at $j_0 = \alpha_0(s)$ will generate, in $A(j, s)$, contributions $\propto t^{\alpha_0(s)}$, and will occur in $A_{\lambda_3\lambda_1}(j, s)$ when $j \pm \frac{1}{2} = \alpha_0(s)$, giving contributions $\propto t^{\alpha_0(s)\mp1/2}$: a "super-Regge-trajectory" $\alpha_0(s)$ would thus correspond to ordinary Regge trajectories, $\alpha_0(s)$ and $\alpha_0(s) \pm \frac{1}{2}$, and the residues of the various contributions would be given in terms of the residue of $T(\alpha_0(s), s)$. A more detailed investigation of this effect is in progress.

It should be pointed out that recent work on the high-energy behavior of the Wess–Zumino model[7] indicates that the leading singularity is not a simple pole. Nevertheless, the effect of supersymmetry is still to tie together the asymptotic behavior of the various spin channels. However, it may be that in a gauge theory model, e.g., Ref. 8, Regge pole behavior will be reinstated also in the supersymmetric case.

ACKNOWLEDGMENTS

I should like to thank Dr. R. Delbourgo for suggesting this work, and for his guidance and encouragement during its progress. I am indebted to the Association of Commonwealth Universities for the award of a Scholarship.

APPENDIX A: SUPERSYMMETRY ALGEBRA AND SPIN BASIS

We use $\eta_{\mu\nu} = \text{diag}\,(+\,-\,-\,-)$, $\mu, \nu = 0, 1, 2, 3$, and $\varepsilon_{0123} = +1$. The γ matrices are generated by $\{\gamma_\mu, \gamma_\nu\} = 2\eta_{\mu\nu}$, and we define $\gamma_5 = \gamma_0\gamma_1\gamma_2\gamma_3$. Adjoint and conjugate spinors are defined by $\bar{S}^\alpha = (A^{-1})^{\alpha\beta}S_\beta^\dagger$, $\bar{S}^{c\alpha} = (C^{-1})^{\alpha\beta}S_\beta$, respectively. We define also $(S_\pm)_\alpha = \frac{1}{2}(1 \pm i\gamma_5)_\alpha{}^\beta S_\beta$.

The supersymmetry algebra is a graded Lie algebra generated by the ordinary Poincaré group generators P_μ, $J_{\mu\nu}$, together with the 4-spinor charge S_α, with the following commutation and anticommutation relations[2b]:

$$\{S_\alpha, S_\beta\} = -(\gamma_\mu C)_{\alpha\beta}P^\mu,$$
$$[S_\alpha, P_\mu] = 0, \tag{A1}$$
$$[S_\alpha, J_{\mu\nu}] = \tfrac{1}{2}(\sigma_{\mu\nu})_\alpha{}^\beta S_\beta,$$

and in a unitary representation, S_α satisfies the reality condition of a Majorana spinor:

$$S_\alpha = C_{\alpha\beta}\bar{S}^\beta. \tag{A2}$$

For the massive case, the Casimirs are given by

$$P^2 = M^2 > 0, \quad K_\mu^\perp K^{\mu\perp} = -P^2 j_0(j_0+1) \tag{A3}$$

in the UIR (M, j_0), where we define[2b]

$$\begin{aligned} K_\mu^\perp &= W_\mu - \Sigma_\mu^\perp = W_\mu - \Sigma_\mu + P_\mu P^{-2} P\cdot\Sigma, \\ \Sigma_\mu &= -\tfrac{1}{4}\bar{S}i\gamma_\mu\gamma_5 S, \quad W_\mu = \tfrac{1}{2}\varepsilon_{\mu\nu\rho\sigma}P^\nu J^{\rho\sigma}. \end{aligned} \tag{A4}$$

[For the massless case, we have instead,[1] $W_\mu = \Lambda P_\mu$, $\Sigma_\mu = NP_\mu$, and $K = \Lambda - \frac{1}{2}(N - \frac{1}{2}) = \lambda_0$.]
In the spin basis, we also have

$$\begin{aligned} &P\cdot\Sigma\,|\,P^2 > 0,\ j_0;\ p\sigma j\lambda\rangle = P^2\sigma\,|\,P^2 > 0,\ j_0;\ p\sigma j\lambda\rangle, \\ &W^2\,|\,P^2 > 0,\ j_0;\ p\sigma j\lambda\rangle \\ &\quad = -P^2 j(j+1)\,|\,P^2 > 0,\ j_0;\ p\sigma j\lambda\rangle, \\ &W_0\,|\,P^2 > 0,\ j_0;\ p\sigma j\lambda\rangle \\ &\quad = \lambda\,|\,\mathbf{P}\,|\,P^2 > 0,\ j_0;\ p\sigma j\lambda\rangle, \end{aligned} \tag{A5}$$

and

$$\begin{aligned} &\langle j_0';\ p'\sigma' j'\lambda'\,|\,j_0;\ p\sigma j\lambda\rangle \\ &\quad = \delta_{j'_0 j_0}\delta_{j' j}\delta_{\lambda'\lambda}\langle p'\,|\,p\rangle, \end{aligned} \tag{A6}$$

where

$$\langle p'\,|\,p\rangle = (2\pi)^3 2p_0\delta^3(\mathbf{p}' - \mathbf{p}). \tag{A7}$$

The spin content is given by

$$\sigma = 0, \pm\tfrac{1}{2}, \qquad j = j_0 \pm \tfrac{1}{2} \mp |\sigma|. \tag{A8}$$

The following properties are important:

$$\begin{aligned} &[S_\alpha, K_\mu^\perp] = 0, \\ &[P\cdot\Sigma, S_\pm] = \pm\tfrac{1}{2}P^2 S_\pm, \end{aligned} \tag{A9}$$

$$(S_\pm)_a^\dagger = \pm\varepsilon_{ab}(S_\mp)_b, \tag{A10}$$

where $a, b = 1, 2$, and ε_{ab} is the alternating tensor.

The effect of S_α on helicity states is most simply discussed in terms of the auxiliary operators

$$R_\kappa^\pm(p) = (-1)^\kappa \bar{u}_\kappa \tfrac{1}{2}(1 \pm i\gamma_5)S, \tag{A11}$$

where $\kappa = \pm\frac{1}{2}$, and where $u_\kappa(p)$ are a pair of normalized, positive-frequency Dirac spinors.[1] In the rest frame, we have

$$\begin{aligned} (S_\pm)_1 &= [i/(2m)^{1/2}]\,R_{\mp 1/2}^\pm, \\ (S_\pm)_2 &= [-i/(2m)^{1/2}]\,R_{\mp 1/2}^\pm. \end{aligned} \tag{A12}$$

The matrix elements of $(S_\pm)_a$ in the spin basis are given in Ref. 1.

APPENDIX B: SUPERHELICITY BASIS

The "superhelicity," κ, is by definition the eigenvalue of

$$\frac{K_0^\perp}{|\mathbf{P}|} = \frac{1}{P_0}\frac{\mathbf{P}\cdot\mathbf{K}^\perp}{|\mathbf{P}|} = -\frac{1}{P_0}\frac{\mathbf{P}\cdot\mathbf{J}}{|\mathbf{P}|} - \frac{\mathbf{P}\cdot\boldsymbol{\Sigma}^\perp}{P_0\,|\mathbf{P}|} = \Lambda - \Delta \tag{B1}$$

in an arbitrary frame, and of

$$-\frac{K_3^\perp}{P_0} = J_3 + \frac{\Sigma_3^\perp}{P_0} = \Lambda - \Delta \tag{B2}$$

in the rest frame. The relation of λ, κ, and δ is given by Eq. (2.10).

The fact [Eq. (A9)] that supertranslations leave κ invariant means that they only act on the + symbols introduced in Eq. (2.11). The nonzero matrix elements of $R_\kappa^\pm$ in this basis are given by

$$\begin{aligned} \langle +|R_+^+|+\rangle &= M = -\langle -+|R_-^-|+\rangle, \\ \langle +-|R_+^+|+-\rangle &= M = -\langle +-|R_-^-|+-\rangle, \\ \langle +-|R_-^+|-+\rangle &= M = -\langle -+|R_+^-|+-\rangle, \\ \langle -+|R_-^+|+-\rangle &= -M = \langle +-|R_+^-|-+\rangle, \end{aligned} \tag{B3}$$

from which, using Eq. (2.12), the matrix elements in the spin basis may be written down.[1]

The normalization of states in the κ basis is

$$\langle p'\ +'\kappa'\,|\,p\ +\ \kappa\rangle = \delta_{j'_0 j_0}\delta_{+'+}\delta_{\kappa'\kappa}\langle p'\,|\,p\rangle. \tag{B4}$$

There is a very natural interpretation of the superhelicity which derives from the induced representation method of Salam and Strathdee.[2b] There one considers UIR's of the "little algebra" of transformations leaving invariant the momentum P_μ, which, for the rest frame in the massive case, consists of the supertranslations adjoined to the ordinary rotation group; the UIR's are constructed by applying supertranslations to a chosen "vacuum" state, and extending to the whole algebra by boosting. Here we choose the vacuum states to be the states $|\,j_0\kappa\rangle$ of a UIR of the rotation group, with spin j_0, and helicity κ, with the condition

$$(S_-)_a\,|\,j_0\kappa\rangle = 0.$$

Then the superhelicity basis in the rest frame is defined by

$$\begin{aligned} |\,+\kappa\rangle &= |\,j_0\kappa\rangle, \\ |\,-+\kappa\rangle &= i\left(\frac{2}{M}\right)^{1/2}(S_+)_1\,|\,j_0\kappa\rangle, \\ |\,+-\kappa\rangle &= i\left(\frac{2}{M}\right)^{1/2}(S_+)_2\,|\,j_0\kappa\rangle, \\ |\,--\kappa\rangle &= \left(\frac{2}{M}\right)(S_+)_1(S_+)_2\,|\,j_0\kappa\rangle, \end{aligned} \tag{B5}$$

and is extended to an arbitrary frame by boosting.

APPENDIX C: THE DURECED STATES $|\,+\ \mp\rangle_t$

In the center of mass frame we have $(S_\pm)_\alpha$ given as in Eq. (A12) in terms of $R_\kappa^\pm$; however using $S_\alpha = (S^1)_\alpha + (S^2)_\alpha$, we have

$$(S_+)_1 = \frac{i}{(2m_1)^{1/2}}e^{-\zeta_1/2}R_-^{1+} - \frac{1}{(2m_2)^{1/2}}e^{\zeta_2/2}R_+^{2+},$$

$$(S_+)_2 = -\frac{i}{(2m_1)^{1/2}} e^{\zeta_1/2} R_+^{1+} + \frac{1}{(2m_2)^{1/2}} e^{-\zeta_2/2} R_-^{2+},$$
$$(S_-)_1 = \frac{i}{(2m_1)^{1/2}} e^{\zeta_1/2} R_-^{1-} - \frac{1}{(2m_2)^{1/2}} e^{-\zeta_2/2} R_+^{2-},$$
$$(S_-)_2 = -\frac{i}{(2m_1)^{1/2}} e^{-\zeta_1/2} R_+^{1-} + \frac{1}{(2m_2)^{1/2}} e^{\zeta_2/2} R_-^{2-}, \tag{C1}$$

where $R_\pm \equiv R_{\pm\frac{1}{2}}$, for p^1 and p^2 in the orientation of Eq. (2.7).

Expressing J_1 and J_2 in terms of invariants m_1^2, m_2^2, and M^2, by means of Eq. (2.7), we can write the labels κ^1 (and similarly κ^2) in a manifestly Lorentz-invariant form, as the eigenvalue of

$$\frac{2}{Q} P\cdot K^{1\perp} = \Lambda^1 - \frac{2}{Q} P\cdot \Sigma^{1\perp}, \tag{C2}$$

where

$$Q(m_1^2, m_2^2, M^2) = [(M^2 - (m_1 + m_2)^2) \times (M^2 - (m_1 - m_2)^2)]^{1/2}. \tag{C3}$$

Moreover, Eq. (A9) ensures that κ^1 and κ^2 are supertranslation-invariant.

The structure of the algebra generated by $J_{\mu\nu}$, $\tilde{P}_\mu$, $\tilde{S}_\alpha$ of Eq. (2.16) is identical to Eqs. (A1), and so we may define analogously $\tilde{W}_\mu$ and $\tilde{\Sigma}_\mu$. Noting that $P\cdot\tilde{W} = -\tilde{P}\cdot W = -P_0P_3\Lambda$, where

$$\tilde{P}_3 = -\tilde{M}\sinh(\zeta_1 - \zeta_2) = -\frac{1}{2M}\frac{m_2^2 - m_1^2}{m_1^2 m_2^2} Q(m_1^2, m_2^2, M^2), \tag{C4}$$

we have

$$\tilde{\Lambda} = \frac{2}{Q}\frac{m_1^2 m_2^2}{m_2^2 - m_1^2} P\cdot\tilde{W} = \Lambda, \tag{C5}$$

so that

$$\tilde{\kappa} = \tilde{\lambda} - \tilde{\delta} = \lambda - \tilde{\delta}, \tag{C6}$$

where

$$\tilde{\Delta} = \frac{2}{Q}\frac{m_1^2 m_2^2}{m_2^2 - m_1^2} P\cdot\tilde{\Sigma}^{\perp}. \tag{C7}$$

Similarly we may define $\tilde{\sigma}$ as the eigenvalue of $\tilde{P}^{-2}\tilde{P}\cdot\tilde{\Sigma}$.

Once again, $\tilde{\sigma}$ and $\tilde{\delta}$ are manifestly Lorentz-invariant; but because of Eq. (2.16a), they are also supertranslation-invariant, with j_{01}, m_1, j_{02}, m_2, κ_1, and κ_2 they complete the set of invariant degeneracy labels needed for the reduction of the direct product. Since they replace the Lorentz- but not supertranslation-invariant combinations $\sigma^1 + \sigma^2$ and $\delta^1 - \delta^2$, they may be written in terms of the latter. We find in general,

$$\tilde{\sigma} = \sigma - \sigma^1 - \sigma^2 \tag{C8}$$

and

$$\tilde{\delta} = \frac{1}{2\cosh(\zeta_1 - \zeta_2)}\left[\frac{2}{\tilde{M}}\left(\frac{\cosh\zeta_1}{m_1}\delta^1 - \frac{\cosh\zeta_2}{m_2}\delta^2\right) - \sinh\zeta_1\sinh\zeta_2(\delta^1 - \delta^2 - 2\delta) - [\sinh\zeta_1\cosh\zeta_2\,\sigma^2 - \sinh\zeta_2\cosh\zeta_1\,\sigma^1 + \sinh(\zeta_1 - \zeta_2)\tilde{\sigma}]\right]. \tag{C9}$$

However, for the special orientation of p^1 and p^2 given by Eq. (2.7), we know from Fig. 2 the structure of the eigenstate of κ, δ, $\sigma^1 + \sigma^2$, and $\delta^1 - \delta^2$. In particular, we have for the extremal states:

$$|+_1 +_2\rangle = |+ \mp\rangle, \quad |+_1 +_2\rangle = |+ \mp\rangle, \quad |+_1 +_2\rangle = |+ \mp\rangle, \quad |+_1 +_2\rangle = |+ \mp\rangle, \tag{C10}$$

for which evaluation of $\tilde{\delta}$ as given by Eq. (C9) on the left-hand sides, shows that the simpler relationship

$$\tilde{\delta} = \delta^1 - \delta^2 - \delta \tag{C11}$$

holds for this case (as indicated on the right-hand sides). Since the other $|+ \mp\rangle$ may be obtained by application of the $(S_\pm)_a$ to these extremal states, and since $\tilde{\delta}$ is supertranslation invariant, we conclude that in this frame Eq. (C11) is generally true, and we use it rather than Eq. (C9) in our discussions of reduced states in the center-of-mass frame.

APPENDIX D: CLEBSCH-GORDAN COEFFICIENTS $\{+_1 +_2| + \mp\}^{j_0}(\theta)$

ζ_1 and ζ_2 are given in terms of the invariants m_1^2, m_2^2 and M^2 by Eqs. (2.5) and (2.7). Explicitly we have [cf. Eq. (C3)]

$$\cosh\zeta_1 = \frac{M^2 + m_1^2 - m_2^2}{2m_1 M}, \qquad m_1\sinh\zeta_1 = \frac{Q(m_1^2, m_2^2, M^2)}{2M}. \tag{D1}$$

We also define

$$\rho = \frac{m_1}{m_2} e^{-(\zeta_1+\zeta_2)} = \frac{1}{2m_2^2}(M^2 - m_1^2 - m_2^2 - \Delta) \tag{D2}$$

in terms of which

$$e^{-\zeta_1} = \frac{M}{m_1}\left(\frac{\rho}{\rho+1}\right), \qquad e^{-\zeta_2} = \frac{m_2}{M}(\rho+1). \tag{D3}$$

If we define

$$C(\kappa) = \left(\frac{j_0 + \kappa + \frac{1}{2}}{2j_0 + 1}\right)^{1/2} = S(-\kappa), \qquad N_\pm(\kappa) = C(\kappa \pm \tfrac{1}{2}) \pm S(\kappa \pm \tfrac{1}{2}), \tag{D4}$$

$$D_{\pm\pm}(1, 2) = [N_\pm(1) \pm \rho\, N_\mp(2)]^{1/2}, \qquad D_{\pm\mp}(1, 2) = [N_\pm(1) \mp \rho N_\mp(2)]^{1/2}, \tag{D5}$$

then the combinations of Poincaré-group Clebsch–Gorden coefficients required for Eqs. (2.27) are

$$\langle + + |\Gamma_{j_0}| + + \rangle = \delta_{\lambda,\kappa_1-\kappa_2}(2j_0+1) = \langle + + |\Gamma_{j_0}| + + \rangle,$$
$$\langle + - |\Gamma_{j_0}| + - \rangle$$

$$= \delta_{\lambda,\kappa_1-\kappa_2+1/2}(2j_0+1)N_-(\kappa_2-\tfrac{1}{2}),$$
$$\langle + + | \Gamma_{j_0} | + + \rangle$$
$$= \delta_{\lambda,\kappa_1-\kappa_2+1/2}(2j_0+1)N_+(\kappa_1+\tfrac{1}{2}), \quad \text{(D6)}$$
$$\langle + + | \Gamma_{j_0} | + + \rangle$$
$$= \delta_{\lambda,\kappa_1-\kappa_2-1/2}(2j_0+1)N_-(\kappa_1-\tfrac{1}{2}),$$
$$\langle + + | \Gamma_{j_0} | + + \rangle$$
$$= \delta_{\lambda,\kappa_1-\kappa_2-1/2}(2j_0+1)\,N_+(\kappa_2+\tfrac{1}{2}).$$

The $\theta = 0$ Clebsch–Gordan coefficients are given, in this notation, in Table I. For arbitrary θ, we use the definition

$$\langle +' | e^{i\theta J_2} | + \rangle = \{ +' + \}^{j_0}_{\kappa'\kappa}(\theta) \quad \text{(D7)}$$

together with Eq. (2.35). Thus, for example, we have

$$^{\theta}\langle + - | + \mp \rangle = i\frac{m_2}{M}\rho^{1/2}\{ + + \}^{j_0}_{\kappa_1-\kappa_2,\kappa}(\theta) \quad \text{(D8)}$$

and the remaining ones are readily written down. In this notation, in the case $\theta = 0$, the $\{+'+\}^{j_0}_{\kappa_1\kappa}$ function is replaced by $\delta_{\kappa_1\kappa}$.

The $\{+'+\}^{j_0}_{\kappa'\kappa}(\theta)$ functions themselves are combinations of the ordinary $d^{j_0}_{\kappa'\kappa}$ functions[9]:

$$\{+ +\}^{j_0}_{\kappa'\kappa}(\theta) = d^{j_0}_{\kappa'\kappa}(\theta) = \{+ +\}^{j_0}_{\kappa'\kappa}(\theta), \quad \text{(D9)}$$
$$\{+ -\}^{j_0}_{\kappa'\kappa}(\theta) = S(\kappa'-\tfrac{1}{2})S(\kappa-\tfrac{1}{2})d^{j_0+1/2}_{\kappa'-1/2\ \kappa-1/2}(\theta)$$
$$+ C(\kappa'-\tfrac{1}{2})\,C(\kappa-\tfrac{1}{2})d^{j_0-1/2}_{\kappa'-1/2\ \kappa-1/2}(\theta),$$
$$\{+ +\}^{j_0}_{\kappa'\kappa}(\theta) = -S(\kappa'-\tfrac{1}{2})\,C(\kappa+\tfrac{1}{2})d^{j_0+1/2}_{\kappa'-1/2\ \kappa+1/2}(\theta)$$
$$+ C(\kappa'-\tfrac{1}{2})\,S(\kappa+\tfrac{1}{2})d^{j_0-1/2}_{\kappa'-1/2\ \kappa+1/2}(\theta), \quad \text{(D10)}$$
$$\{+ -\}^{j_0}_{\kappa'\kappa}(\theta) = -C(\kappa'+\tfrac{1}{2})S(\kappa-\tfrac{1}{2})d^{j_0+1/2}_{\kappa'+1/2\ \kappa-1/2}(\theta)$$
$$+ S(\kappa'+\tfrac{1}{2})\,C(\kappa-\tfrac{1}{2})d^{j_0-1/2}_{\kappa'+1/2\ \kappa-1/2}(\theta),$$
$$\{+ +\}^{j_0}_{\kappa'\kappa}(\theta) = C(\kappa'+\tfrac{1}{2})\,C(\kappa+\tfrac{1}{2})d^{j_0+1/2}_{\kappa'+1/2\ \kappa+1/2}(\theta)$$
$$+ S(\kappa'+\tfrac{1}{2})\,S(\kappa+\tfrac{1}{2})d^{j_0-1/2}_{\kappa'+1/2\ \kappa+1/2}(\theta).$$

They are related through the symmetry properties

$$\{+ -\}^{j_0}_{\kappa'\kappa}(\theta) = (-1)^{\kappa'-\kappa}\{+ -\}^{j_0}_{-\kappa'-\kappa}(\theta),$$
$$\{+ +\}^{j_0}_{\kappa'\kappa}(\theta) = -(-1)^{j_0-\kappa'}\{+ +\}^{j_0}_{-\kappa'\kappa}(\theta-\pi), \quad \text{(D11)}$$
$$\{+ -\}^{j_0}_{\kappa'\kappa}(\theta) = (-1)^{j_0+\kappa}\{+ +\}^{j_0}_{\kappa'-\kappa}(\theta-\pi).$$

*Present address: Department of Physics, University of Southampton, Southampton S09 5NH, United Kingdom.

[1]P.D. Jarvis, J. Math. Phys. **17,** (1976).
[2](a) J. Wess and B. Zumino, Nucl. Phys. B **70,** 39 (1974); (b) Abdus Salam and J. Strathdee, Phys. Rev. D **11,** 1509 (1975); (c) Nucl. Phys. B **80,** 499 (1974).
[3]M. Jacob and G.C. Wick, Ann. Phys. **7,** 404 (1959).
[4]Abdus Salam and J. Strathdee, Nucl. Phys. B **87,** 85 (1975).
[5]P.D.B. Collins and E. J. Squires, *Regge Poles in Particle Physics* (Springer-Verlag, Berlin, 1968).
[6]F. Calogero, J.M. Charap, and E.J. Squires, Ann. Phys. **25,** 325 (1963).
[7]R. Kee (unpublished).
[8]L. Tyburski, Phys. Lett. B **59,** 49 (1975).
[9]A.R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U.P., Princeton, N.J., 1960).

A Lagrangian formulation for noninteracting high-spin fields*

Francisco Antonio Doria

Theoretical Physics Department, Instituto de Física, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
(Received 28 October 1976)

We develop here some properties of Clifford algebras and of "algebraic" or Sauter spinors that lead to a very simple classification scheme for the Lorentz group representations. We apply our results to a Dirac-like Lagrangian and get a general formulation for noninteracting high-spin fields; the equations for spin 1/2, 0 and 1, 3/2 and 2 are explicitly calculated; the first three are seen to reproduce some previous results by Teitler.

I. INTRODUCTION

A classical difficulty in the theory of higher-spin equations concerns its Lagrangian formulation. The literature review[1,2] furnishes but one example of an easy Lagrangian treatment for high spin.[2] The structural simplicity of the Bargmann–Wigner equations[3] and the elegant group-theoretic reasoning that led to them appears in strong contrast to the above somewhat ad hoc formulations, as already noted by Corson.[4] This suggests that a "natural" Lagrangian for higher-spin equations may be obtained if we are able to parallel, in an adequate way, the Bargmann–Wigner theory. This is the chief idea behind the present paper. We present here a very simple Lagrangian formulation for noninteracting high spin field equations based on a not very much explored algebraic treatment for the classification of Lorentz group representations.

We start from the theory of Clifford algebras and its automorphisms; this is presented in Sec. II, the connection between Clifford algebras and the Lorentz group being established in Sec. III. Section IV reviews some properties of "algebraic" or Sauter spinors,[5] that is, spinors as objects that are embedded in a given Clifford algebra. Section V introduces our Lagrangian formulation through a classification scheme for Lorentz group representations. The first (for spin $\frac{1}{2}$, 0, 1, $\frac{3}{2}$, and 2) field equations are explicitly presented and explored in the massive and massless cases; they are seen to reproduce the well-known Fierz[1] and Rarita–Schwinger[6] schemes. Our formulation differs from Durand's[2] in the algebraic structures considered (the objects we use are explicitly representation-free and are usually defined on lower-dimensional Clifford algebras than those employed by Durand). We also get Teitler's very nice equations for spin-0 and 1 as a consequence of our analysis.[7]

II. THE (+) AND (-) ADJUNCTIONS WITHIN SIMPLE CLIFFORD ALGEBRAS

Our basic results will be derived for simple Clifford algebras; they are however easily extended to the semisimple case. Let $C(V, g)$ be the Clifford algebra $\otimes V/I$ associated to a complex vector space V, $\otimes V = C \oplus V \otimes V \oplus \cdots$ being the weak direct sum of tensor products (that is, the tensor algebra) of V, and I being the two-sided ideal generated by $x \otimes x - g(x, x)$, $x \in V$ and g being a symmetric and nondegenerate bilinear form with real values for V's real subspace. Now let's define three inner products within $C(V, g)$:

(a) The standard inner product:

$$g(x, y) = \tfrac{1}{2}[x, y]_+ = \tfrac{1}{2}(xy + yx) \tag{2.1}$$

(juxtaposition will always denote the Clifford product).

(b) The positive-definite inner product:

$$\langle x, y\rangle = g(x^*, y), \tag{2.2}$$

$$\langle x, x\rangle \in \mathbf{R}, \quad \langle x, x\rangle \geq 0 \text{ and equality just when } x = 0. \tag{2.3}$$

(c) The real-indefinite inner product:

$$\langle x, y\rangle_L = g(x^-, y), \tag{2.4}$$

$$\langle x, x\rangle_L \in \mathbf{R}, \tag{2.5}$$

$\langle\ ,\ \rangle_L$ has the same signature as g acting on V's real subspace. (2.6)

The $(^*)$ and $(^-)$ operations are defined through conditions (2.2)–(2.6); they are also supposed to act as a homothety on any $x \in V$, that is, given any $\lambda \in \mathbb{C}$,

$$(\lambda x)^* = f(\lambda)x \quad \text{and} \quad (\lambda x)^- = h(\lambda)x, \tag{2.7}$$

f and h being complex-valued functions to be specified. It is then easily concluded that $\langle \alpha x, y\rangle = \alpha^* \langle x, y\rangle$, and $\langle \alpha x, y\rangle_L = \alpha^* \langle x, y\rangle_L$, $\alpha \in C$, the * denoting here a complex conjugation (take g in a diagonalized basis and discuss the $\langle\ ,\ \rangle$ and $\langle\ ,\ \rangle_L$ functions with the help of (2.2)–(2.7). Again with the help of an orthonormal set of generators for $C(V, g)$, denoted by $\{e_\mu, e_{\dot\mu}\}$, the undotted index running over the + valued coordinates in the diagonalized g, and the dotted one over the − valued coordinates, we can see that

$$e_\mu^* = e_\mu \quad \text{and} \quad e_{\dot\mu}^* = -e_{\dot\mu}, \tag{2.8}$$

$$e_\mu^- = e_\mu \quad \text{and} \quad e_{\dot\mu}^- = -e_{\dot\mu}. \tag{2.9}$$

The transformed sets of orthonormal generators obey the same anticommutation relations as the untransformed ones, that is,

$$[e_a^*, e_b^*]_+ = [e_a^-, e_b^-]_+ = [e_a, e_b] \tag{2.10}$$

(a, b running over the whole index set of the basis),

but this equality doesn't hold for a generic $x \in V$. There is thus a great similarity between both our adjunctions.

Copyright © 1977 American Institute of Physics

In order to explore it and to connect $(^*)$ and $(^-)$ we need a lemma first derived by Brauer and Weyl[8]:

Lemma: Given two orthonormal sets of generators $\{e_a\}$, $\{f_a\}$, in a simple Clifford algebra $C(V, g)$, there exists an invertible and unimodular $u \in C(V, g)$ such that $\{f_a\} = \pm u\{e_a\}u^{-1}$, the sign ambiguity being an essential feature.

Brauer and Weyl's proof is based on representations for simple Clifford algebras; a purely algebraic one will be given in the next section. The lemma is easily extended to the semisimple case; it will be used here to obtain the well-known[9] connection between the hermitized and Dirac-adjoint elements in any Clifford algebra. For the Lorentzian case, that is, the four-dimensional complex Dirac algebra with a + 2 metric, one gets

$$p^- = + e_0 p^* e_0, \tag{2.11}$$

e_0 denoting the timelike unit vector, for the $\{e_a^*\}$ and $\{e_a^-\}$ sets of generators obey the same anticommutation rules. The + sign is chosen due to (2.6), and this result is easily extended to any g in any simple Clifford algebra.

The $(^*)$ and $(^-)$ conjugations are thus merely the standard Hermitian conjugation and the (general) Dirac adjunction with respect to a given real and nonsingular metric g; our definition is however basis and representation-free.

One further property will be needed here. Using the exterior product

$$x\Lambda y = \tfrac{1}{2}[xy - yx] = \tfrac{1}{2}[x, y]_-, \quad x, y \in V, \tag{2.12}$$

we define

$$g(x\Lambda y, z\Lambda w) = g(y, z)g(x, w), \quad x, y, z, w \in V. \tag{2.13}$$

We now ask that the extension of $\langle\ ,\ \rangle$ be positive definite. That is,

$$\langle x\Lambda y, x\Lambda y\rangle = \langle x, x\rangle\langle y, y\rangle. \tag{2.14}$$

One must then have $g((x\Lambda y)^*, x\Lambda y) = g(x^*, x)g(y^*, y)$, wherefrom we get

$$(x\Lambda y)^* = y^*\Lambda x^*. \tag{2.15}$$

Given (2.15) and (2.11) we can extend the $(^-)$ conjugation to products within $C(V, g)$.

III. PSEUDO-ORTHOGONAL GROUPS AND THE (-) CONJUGATION

We will need some further properties of the $(^-)$ conjugation in order to get to our classification scheme for the Lorentz group representations. The first step is given by the lemma stated in the previous section, which we now prove:

Lemma 3.1: Given two orthonormal sets of generators $\{e_a\}$, $\{f_a\}$ in a simple Clifford algebra $C(V, g)$, there exists an invertible and unimodular $u \in C(V, g)$ such that $\{f_a\} = \pm u\{e_a\}u^{-1}$, the sign ambiguity being an essential feature.

Proof: (A) Define $f_a = p_{(a)}e_a q_{(a)}$ (no sum), $p_{(a)}, q_{(a)} \in C(V, g)$, invertibles. This means that $(f_a)^2 = p_{(a)}e_a q_{(a)}p_{(a)}e_a q_{(a)} = g_{aa}$ (no sum). Since we are on an orthonormal frame, $g_{aa}p_{(a)}e_a q_{(a)} = (p_{(a)}e_a q_{(a)})^{-1} = q_{(a)}^{-1}e_a^{-1}p_{(a)}^{-1} = g_{aa}q_{(a)}^{-1}e_a p_{(a)}^{-1}$. Thus $(q_{(a)}p_{(a)})e_a(q_{(a)}p_{(a)}) = e_a$. This is valid for any basis within the algebra, and so we must have:

(i) $q_{(a)}p_{(a)}$ commutes with e_a,

(ii) $(q_{(a)}p_{(a)})^2 = 1$.

There now appears a certain arbitrariness in the definition of $p_{(a)}$ and $q_{(a)}$. For suppose that $q_{(a)}p_{(a)} = u$. Then $q_{(a)}p_{(a)}u = 1$. We may choose a new $p'_{(a)} = p_{(a)}u$, and it is then sufficient to impose $q_{(a)}p_{(a)} = \lambda_{(a)}1$, and conclude from (i), (ii), and the preceding remarks that $\lambda_{(a)} = \pm 1$.

(B) The anticommutation relations $[f_a, f_b]_+ = 0$, $a \neq b$, and the already established $f_a = \pm q_{(a)}e_a q_{(a)}^{-1}$ (no sum) leads to $p_{(a)}e_a p_{(a)}^{-1}p_{(b)}e_b p_{(b)}^{-1} = -p_{(b)}e_b p_{(b)}^{-1}p_{(a)}e_a p_{(a)}^{-1}$. Thus $e_a p_{(a)}^{-1}p_{(b)}e_b = -p_{(a)}^{-1}p_{(b)}e_b p_{(b)}^{-1}p_{(a)}e_a p_{(a)}^{-1}p_{(b)}$. Since this is valid for all frames and all indices, one must forcibly have $p_{(a)}^{-1}p_{(b)} = \lambda_{(a,b)}1$. So, $e_a e_b = -\lambda_{(a,b)}\lambda_{(b,a)}e_b e_a$. Thus $\lambda_{(a,b)}\lambda_{(b,a)} = 1$. If one takes p to be unimodular, the coefficients will be just ± 1. Thus the lemma.

Lemma 3.1 describes *all possible* sets of generators for a simple $C(V, g)$, and not only those formed by linear combinations of another set of generators. The simplicity assumption is used when establishing the values of $\lambda_{(a)}$ and $\lambda_{(a,b)}$. To extend Lemma 3.1 to a semisimple Clifford algebra one may decompose this algebra into two two-sided ideals and apply the lemma to each ideal.

We will now write $O(g)$ for the real group of matrices that preserves g; given any $L \in O(g)$, one can state:

Proposition 3.2: $L \in O(g)$ if and only if it preserves the anticommutation relations for orthonormalized generators.

Proof: $2\lambda_a^b g_{bc}\lambda_d^c = 2g_{ad} = [f_a, f_d]_+ = \lambda_a^b[e_b, e_c]_+\lambda_d^c$, where g_{ab} denotes the diagonal form of g and $L = (\lambda_b^a)$.

This entails our chief consequence:

Proposition 3.3: There exists a $(^-)$ unitary or $(^-)$ antiunitary $u (u u^- = \pm 1)$ such that $u e_a u^- = f_a = L(e_a)$.

Proof: First see that $(u^-)^{-1} = (u^{-1})^-$. Then, note that $u e_a u^{-1} = \lambda_a^b e_b$, where all coefficients are real. Then its adjoint $(u^-)^{-1}e_a u^- = \lambda_a^b e_b$. This implies that $e_a = u^{-1}(u^-)^{-1} \times e_a \bar{u}u = (\bar{u}u)^{-1}e_a(\bar{u}u)$. For simple algebras, Schur's lemma allows us to conclude that $u\bar{u} = \lambda 1$, $\lambda \in \mathbb{C}$, and one can choose $\lambda = \pm 1$, sign indefinition being an essential feature. For semisimple algebras, one can use the trick described after Lemma 3.1.

Proposition 3.3 means that every element in the pseudo-orthogonal group $O(g)$ may be represented by a $(^-)$ unitary $u \in C(V, g)$, up to a sign. That is, the $(^-)$ conjugation stands in respect to g as the Hermitian conjugation $(^*)$ with respect to positive definite bilinear forms and may be similarly used to classify $O(g)$'s linear representations in a straightforward manner.

IV. SAUTER SPINORS

The classification scheme for Lorentz group representations that we are going to develop in the present

section and in the next is based on a very simple idea: The tensor product of two copies of a representation space for the orthogonal group decomposes into three irreducible pieces, diagonal tensors, traceless symmetric tensors, and skewsymmetric tensors. We repeat this for the Lorentz group, using Sauter (or "algebraic") spinors as the basis space for linear representations; the product of a Sauter spinor space and its Hermitian adjoint, when linearly extended, is easily seen to be the simple Clifford algebra where those spaces are embedded. As for the orthogonal group, this product decomposes into three irreducible pieces (for the Lorentz group, these pieces have dimensions one, five, and ten), the scalar, the $(\bar{})$ adjoint and the skew $(\bar{})$ adjoint objects of the Clifford algebra.

Spinors are usually defined as objects belonging to the representation space of a simple Clifford algebra. However a well-known idea first advanced by Sauter in the late twenties[5] provides a natural embedding for spinor space within the algebra that acts on it; Sauter spinors (S-spinors) thus have an extra structure besides the original linear one, the product inherited from the embedding algebra. S-spinors are defined as follows: Let e be a minimal projection within a simple $C(V,g)$. Then ψ is a left S-spinor only if $\psi=\psi e$. The space of left S-spinors is a minimal left ideal $I\subset C(V,g)$; its adjoint ideal $I^\dagger\subset C(V,g)$ is the space of right S-spinors $\psi^\dagger=e\psi^\dagger$.

Now when working with S-spinors one must face two questions: (i) given a spectral decomposition $\sum e_i=1$ for the algebra's identity, which projection are we to choose as a generating unit for the S-spinors' ideals? And again, since the identity's decomposition is given modulo a similarity transformation (see Lemma 4.1 below), then (ii) how are we to choose a particular projection to generate S-spinors? This last indeterminacy reproduces the fact that given a Clifford algebra representation, it will be given only up to a similarity mapping. Again if we put on an equal footing every ideal belonging to a given spectral resolution of the algebra, we are saying that no particular ordering (or labeling) of the generator set is to be preferred; actually left (and right) ideals are isomorphic, their only distinction lying in their generating projections. And since reordering of the generators is a linear transformation it will be also given by a similarity mapping.

As we know that the physics is representation independent and that calculations only involve components (and not algebraic bases) it seems natural to include the undeterminacies asserted in (i) and (ii) when defining S-spinors. Actually we do the opposite; fixing a certain ideal I and its adjoint $I^\dagger$ we are going to construct the higher order representations up to a similarity mapping.

We start from a frequently used property:

Lemma 4.1: Let $\sum e_i=1$ be a spectral resolution for the identity in a simple algebra A. Then $q(\sum e_i)q^{-1}$ describes all possible decompositions for the identity in A, q being invertible and unimodular.

Proof: Take two different spectral resolutions $\sum e_i=\sum e_i'=1$. Generate the isomorphic matrix algebra M with the e_i. Get the expression for the e_i' in M; they determine a second faithful representation M' for A. Then there must be a q satisfying the hypothesis' conditions and such that $M'=qMq^{-1}$, due to the equivalence of both representations. Note that this q isn't restricted to the ones given by Proposition 3.3, since we must include here every inner automorphism of the algebra.

Given this lemma one may act as follows to get the first higher order representation space for the Lorentz group. Let $I\subset C(V,g)$ be an arbitrary minimal left ideal within a simple Clifford algebra (an arbitrary left S-spinor space). Let $I^\dagger$ be its associated right ideal. Then we have:

Proposition 4.2: The set $I^\dagger I=I^\dagger\cap I$ is the set $\{\lambda e\}$, $\lambda\in\mathbb{C}$ and e being the generating projection. The set $II^\dagger$ when linearly extended is isomorphic to $C(V,g)$ and the isomorphism is given up to a similarity mapping.

Proof: These are the properties currently used when forming the quantum mechanical operators $\sum|\rangle\langle|$ and $\sum|\rangle\langle|$. To show the first part of the proposition one must apply the well-known property of minimal projections, $epe=\lambda e$, $\lambda\in\mathbb{C}$, and p being a generic element of the algebra. To show the second part, construct the matrix algebra isomorphic to $II^\dagger$'s linear extension. It is clearly $C(V,g)$ up to an equivalence mapping.

The undeterminacy in Proposition 4.2 means that the algebra's structure is given globally in the product $II^\dagger$; its objects, however, are fixed only up to an isomorphism. This is very important, for in the next section we are going to find products between Clifford objects and S-spinors, and we treat both as elements of different spaces, due to the above undeterminacy.

The next property gives the irreducible pieces of $C(V,g)$ when acted upon by the Lorentz group:

Proposition 4.3: Define $\psi_{(\pm)}=\frac{1}{2}(\psi\pm\psi^-)$. Denote by L_u the $O(g)$ transformation associated to a $(\bar{})$ unitary object u. Then $L_u(\psi_{(\pm)})=[L_u(\psi)]_{(\pm)}$.

Proof: For the Lorentzian case, use (2.11) and calculate directly. For a generic $O(g)$, extend (2.11) and again apply it.

Corollary 4.4: $C(V,g)=C_0\oplus C_{(+)}\oplus C_{(-)}$, where C_0 denotes the set of scalars and $C_{(\pm)}$ denotes the sets of $(\bar{})$ adjoint and $(\bar{})$ skew-adjoint objects, scalars excluded.

Note that these pieces are irreducible, despite the fact that given a real $\psi\in C_4$, the Clifford—Dirac algebra, the expansion

$$\psi=\psi_0+\psi_\mu\gamma^\mu+\tfrac{1}{2}\psi_{\mu\nu}\gamma^{\mu\nu}+\psi_{\mu 5}\gamma^\mu\gamma^5+\psi_5\gamma^5$$

seems to be such that a Lorentz transformation keeps the algebra's gradation, that is, maps vectors over vectors, tensors over tensors. But this is not the case. For consider a "nonstandard" set of generators for C_4,

$$(\gamma^0)'=-\gamma^1\gamma^3,\quad(\gamma^1)'=i\gamma^0\gamma^1\gamma^2,$$
$$(\gamma^2)'=\gamma^3,\quad(\gamma^3)'=-i\gamma^2\gamma^3.$$

This set is readily seen to satisfy $[(\gamma^\mu)',(\gamma^\nu)']_+=[\gamma^\mu,\gamma^\nu]_+=2g^{\mu\nu}$; if we now write ψ over the primed set, it is easily seen that no gradation is preserved when we make the mapping $\psi\to L(\psi)$. A product of ideals doesn't impose an unique choice of generators for the algebra,

and the above preserved gradation will only happen for a specific choice of the generator set.

Extension of the above results to semisimple Clifford algebras is straightforward once we form the semisimple algebra as a direct sum of two equal two-sided ideals having a simple Clifford algebra structure.

V. A LAGRANGIAN FORMULATION FOR HIGHER SPIN

Throughout this section we will only consider C_4, the Clifford–Dirac algebra with respect to a +2 metric and over the ring of complex-valued functions defined on Minkowski space. We are going to get the explicit forms of the Lorentz-irreducible wavefunctions that correspond to spins $\frac{1}{2}$, 0, 1, $\frac{3}{2}$, and 2, and to show the generic form for any spin.

Since we have been trying to parallel the Bargmann–Wigner (B–W) theory[3] and since the B–W equations are essentially a system of Dirac-like equations it seems natural to consider representations of the Dirac Lagrangian,

$$\mathcal{L} = 2m\psi^*\psi - (\nabla\psi)^*\psi + \psi^*(\nabla\psi), \tag{5.1}$$

where ∇ is a linear operator acting on the ψ-algebra whose square equals the +2 d'Alembertian, that is,

$$\nabla^2 = \Box = -\partial_0^2 + \partial_1^2 + \partial_2^2 + \partial_3^2$$

if one chooses an orthonormal frame. ∇ can then be represented by $\nabla = \gamma^\mu\partial_\mu$, $\mu = 0, \ldots, 3$, with $[\gamma^\mu, \gamma^\nu]_+ = 2g^{\mu\nu}$ and $\partial_\mu \equiv \partial/\partial X^\mu$. Equation (5.1) entails the Dirac equations

$$(\nabla + m)\psi = 0 \quad \text{or} \quad (\gamma^\mu\partial_\mu + m)\psi = 0, \tag{5.2a}$$

$$(\nabla\psi)^* + m\psi^* = 0 \quad \text{or} \quad (\partial_\mu\psi)^*(\gamma^\mu)^* + m\psi^* = 0 \tag{5.2b}$$

(concerning our choice for the γ's algebra, see the Appendix). The lowest dimensional representation for (5.2) and the γ-algebra is given by the usual Dirac matrices, that is, by the Clifford–Dirac algebra C_4. ψ must then be an element of C_4, a spinor or an irreducible object belonging to that algebra. We will consider both cases in detail:

A. The Spinor Representation

Let $\psi \in I$, a minimal left ideal within C_4. Proposition 3.3 shows that the $(^-)$ unitary u's are a four-spinor representation for the Lorentz group. That is, given a Lorentz transformation L, it induces a mapping

$$\psi' = u(L)\psi, \quad \psi', \psi \in I, \tag{5.3}$$

that maps the ideal over itself. Now in order to define an adequate ψ^-, we get the Hermitian adjoint of the above equation, $(\psi')^* = \psi^*(u(L))^*$. ψ^- will be the object acted upon by u^-, that is,

$$(\psi^*)(u(L))^*\gamma^0 = -(\psi^*\gamma^0)(\gamma^0(u(L))^*\gamma^0) = -\psi^-u^-. \tag{5.4}$$

The minus sign being irrelevant, we may define

$$\psi^- = \psi^*\gamma^0 \tag{5.5}$$

as usual. Right multiplication of (5.2b) by γ^0 leads to

$$(\partial_\mu\psi)^-\gamma^\mu - m\psi^- = 0. \tag{5.6}$$

Nothing more can be said about the spinor representation unless we choose a particular ideal I within C_4 and calculate to get the standard differential equations for the spin-$\frac{1}{2}$ particle. The massless case has also no particular features with respect to our goals.

B. The Vector Representations

For a general $\psi \in C_4$, the irreducible pieces with respect to the Lorentz group are easily obtained from the expansion

$$\psi = \psi_0 + \psi_\mu\gamma^\mu + \tfrac{1}{2}\psi_{\mu\nu}\gamma^{\mu\nu} + \psi_{\mu 5}\gamma^\mu\gamma^5 + \psi_5\gamma^5,$$

$$\psi_{(0)} = \mathrm{Re}(\psi_0) + i\,\mathrm{Im}(\psi_0), \tag{5.7a}$$

$$\psi_{(+)} = \mathrm{Re}(\psi_\mu\gamma^\mu + \tfrac{1}{2}\psi_{\mu\nu}\gamma^{\mu\nu}) + i\,\mathrm{Im}(\psi_{\mu 5}\gamma^\mu\gamma^5 + \psi_5\gamma^5), \tag{5.7b}$$

$$\psi_{(-)} = \mathrm{Re}(\psi_{\mu 5}\gamma^\mu\gamma^5 + \psi_5\gamma^5) + i\,\mathrm{Im}(\psi_\mu\gamma^\mu + \tfrac{1}{2}\psi_{\mu\nu}\gamma^{\mu\nu}). \tag{5.7c}$$

Since the Lorentz group is a real group when acting on space–time, real (imaginary) objects are mapped on real (imaginary) objects. The symmetry in the set (5.7) suggests that we consider only the corresponding real parts (anyway real and imaginary won't mix due to the reality of the operator $\gamma^\mu\partial_\mu + m$).

a. The scalar representation: The equation $(\nabla + m)\psi_{(0)} = 0$ implies $\psi_{(0)} = 0$.

b. The spin-0 field: We now take $\psi = \mathrm{Re}\psi_{(-)} = (\phi_\mu\gamma^\mu - m\phi)\gamma^5$, ϕ, ϕ_μ being real functions (the mass coefficient hasn't been introduced in an ad hoc fashion[10]). Taking this ψ into (5.2a),

$$(\gamma^\mu\partial_\mu + m)(-m\phi + \phi_\alpha\gamma^\alpha)\gamma^5 = 0,$$

one gets the system

$$\partial^\alpha\phi_\alpha = m^2\phi, \tag{5.8a}$$

$$(m)\partial_\alpha\phi = (m)\phi_\alpha, \tag{5.8b}$$

$$\partial_\alpha\phi_\mu = \partial_\mu\phi_\alpha, \tag{5.8c}$$

which describes a spin-0 field, since it trivially implies the Klein–Gordon equation $(\Box - m^2)\phi = 0$. The (5.8) system has been already derived by Teitler from a Dirac-like equation.[7]

Now for the massless case consider

$$(\gamma^\mu\partial_\mu)(-\phi + \phi_\alpha\gamma^\alpha) = 0.$$

The system is uncoupled, for this equation implies

$$\partial_\mu\phi = 0, \tag{5.9a}$$

$$\partial^\alpha\phi_\alpha = 0, \tag{5.9b}$$

$$\partial_\mu\phi_\alpha = \partial_\alpha\phi_\mu. \tag{5.9c}$$

ψ has been decomposed into a scalar and a massless spin-0 field. [Equation (5.9c) implies that there is, locally at least, a function χ such that $\phi_\alpha = \partial_\alpha\chi$, and the existence of such a function allows us to recover the massless Klein–Gordon equation $\Box\chi = 0$.] The irreducible wavefunction in the massless case is thus given by $\psi = \phi_\alpha\gamma^\alpha\gamma^5$.

c. The Proca field: We now take $\psi = \mathrm{Re}(\psi_{(+)}) = (-m\phi_\mu\gamma^\mu + \frac{1}{2}\phi_{[\mu\nu]}\gamma^{\mu\nu})$. The wave equation (5.2a) becomes

$$(\gamma^\mu\partial_\mu + m)(-m\phi_\alpha\gamma^\alpha + \tfrac{1}{2}\phi_{[\alpha\beta]}\gamma^{\alpha\beta}) = 0$$

in the massive case. This implies the Proca set

$$(m)\partial^\alpha \phi_\alpha = 0, \tag{5.10a}$$

$$(m)\phi_{[\alpha\beta]} = (m)(\partial_\alpha \phi_\beta - \partial_\beta \phi_\alpha), \tag{5.10b}$$

$$\partial^\alpha \phi_{[\alpha\beta]} = m^2 \phi_\beta, \tag{5.10c}$$

$$\partial_\rho \phi_{[\sigma\tau]} + \partial_\tau \phi_{[\rho\sigma]} + \partial_\sigma \phi_{[\tau\rho]} = 0. \tag{5.10d}$$

The square brackets [ ] indicate skew-symmetric indices; again the mass coefficient hasn't been introduced by hand,[10] and it has been explicitly presented in order to compare the massive case with the massless (below). For in the massless limit the wavefunction ψ splits into a spin-0 and a spin-1 zero-mass fields, to which one must add a gauge condition. The massless field equation is

$$(\gamma^\mu \partial_\mu)(-\phi_\alpha \gamma^\alpha + \tfrac{1}{2}\phi_{[\alpha\beta]}\gamma^{\alpha\beta}) = 0.$$

It implies

$$\partial^\alpha \phi_\alpha = 0, \tag{5.11a}$$

$$\partial_\alpha \phi_\beta = \partial_\beta \phi_\alpha, \tag{5.11b}$$

$$\partial^\alpha \phi_{[\alpha\beta]} = 0, \tag{5.11c}$$

$$\partial_\alpha \phi_{[\beta\delta]} + \partial_\delta \phi_{[\alpha\beta]} + \partial_\beta \phi_{[\delta\alpha]} = 0. \tag{5.11d}$$

The first two equations describe a massless Klein–Gordon field [as (5.9)]; the remaining two are the familiar Maxwellian microscopic equations, since the Lorentz condition (5.10a) now becomes part of a first-order system for the spin-0 field; it must be added to (5.11c) and (5.11d) in order to ensure the correct value for spin. We must note that when taking the limit $m \to 0$ in the Proca–Dirac field equation the correct gauge condition will remain in the derived equations [just cancel the unnecessary mass in (5.10) and go to the limit].

The equations considered in sections V. A and V. B of the present paper have been introduced by Riesz[11] and studied by Teitler,[7] Hestenes,[12] and Cercignani[13]; they are simpler than the Duffin–Kemmer set[14] due to their algebraic structure. The spin-1 (massive and massless) field ψ and the corresponding field equations have been derived by Lurié[15] directly from the Bargmann–Wigner set.

C. The first tensor representations for the Dirac Lagrangian

We are now going to obtain the spin-$\frac{3}{2}$ and spin-2 field equations. They arise from irreducible representations of the tensor product $C_4 \otimes C_4$. The Clifford–Dirac anticommutation condition $[\gamma^\mu, \gamma^\nu]_+ = 2g^{\mu\nu}$ is here represented by any one of the sets $e_1^\mu = \gamma^\mu \otimes 1$ and $e_2^\mu = 1 \otimes \gamma^\mu$, which are unique up to an equivalence. For the spin-2 field one takes the product $\mathrm{Re}C_{4(+)} \otimes \mathrm{Re}C_{4(+)}$ and decomposes it into three irreducible pieces, zero-trace elements of $\mathrm{Re}C_{4(+)} \odot \mathrm{Re}C_{4(+)}$, diagonal objects, and $\mathrm{Re}C_{4(+)} \Lambda\, \mathrm{Re}C_{4(+)}$ ($\odot$ denotes the symmetric tensor product, Λ the skew-symmetric; the null trace condition is due to the Lorentzian relation $L_\mu^\rho L_\nu^\sigma g_{\rho\sigma} = g_{\mu\nu}$). The first piece has the highest spin, as we will see. The spin-$\frac{3}{2}$ case is a bit more complicated: First note that given $I \subset C_4$, $I \otimes C_4 \oplus C_4 \otimes I$ is the representation space for the product transformation $L \otimes u = (L_\nu^\mu u_B^A)$, L acting on C_4 and u on I. Now the representation space again decomposes into a symmetric $I \odot C_4$ and a skew-symmetric $I \Lambda C_4$. Taking irreducible factors, $I \odot \mathrm{Re}C_{4(+)}$ is the space of spin-$\frac{3}{2}$ fields ($L \otimes u$ doesn't imply a further decomposition).

a. The Rarita–Schwinger[6] spin-$\frac{3}{2}$ field: Given $\psi \in I \odot \mathrm{Re}C_{4(+)}$ it can be expanded as

$$\psi = -m\psi_{\mu A}(\gamma^\mu \odot \sigma^A) + \tfrac{1}{2}\psi_{[\mu\nu]A}(\gamma^{\mu\nu} \odot \sigma^A),$$

where the σ^A, $A = 0, 1, 2, 3$, form a basis for S-spinors. The equation we start from is

$$(e_1^\mu \partial_\mu + m)\psi = 0. \tag{5.12}$$

It implies:

$$(m)\partial^\alpha \psi_{\alpha A} = 0, \tag{5.13a}$$

$$(m)(e_1^\alpha \partial_\alpha + m)\psi_{\mu A}\sigma_1^{\mu A} = 0 \tag{5.13b}$$

[where we have written $\sigma_1^{\mu A} = (\sigma^A \otimes \gamma^\mu)$];

$$(m)\psi_{[\mu\nu]A} = (m)(\partial_\mu \psi_{\nu A} - \partial_\nu \psi_{\mu A}), \tag{5.14a}$$

$$\partial^\alpha \psi_{[\alpha\nu]A} = m^2 \psi_{\nu A}, \tag{5.14b}$$

$$\partial_\alpha \psi_{[\mu\nu]A} + \partial_\nu \psi_{[\alpha\mu]A} + \partial_\mu \psi_{[\nu\alpha]A} = 0, \tag{5.14c}$$

$$(e_1^\alpha \partial_\alpha + m)\psi_{[\mu\nu]A}\sigma_1^{\mu\nu A} = 0 \tag{5.14d}$$

[where we have written $\sigma_1^{\mu\nu A} = (\sigma^A \otimes \gamma^{\mu\nu})$]. System (5.13) is equivalent to the Rarita–Schwinger spin-$\frac{3}{2}$ set; (5.14) is a derived system that endows this field with a Maxwell-like structure. The mass m hasn't been added by hand, but naturally appears when we calculate (5.12); the wavefunction ψ has been adjusted in order that (5.14) gets its familiar aspect.

The massless case shows a system of uncoupled fields to which one must add extra conditions in order to get spin-$\frac{1}{2}$ and spin-$\frac{3}{2}$ equations. We start from

$$(e_1^\mu \partial_\mu)\psi = 0, \tag{5.15a}$$

with

$$\psi = -m\psi_{\mu A}(\gamma^\mu \odot \sigma^A) + \tfrac{1}{2}\psi_{[\mu\nu]A}(\gamma^{\mu\nu} \odot \sigma^A). \tag{5.15b}$$

The first field is given by

$$(e_1^\mu \partial_\mu)\psi_{\alpha A} e_1^{\alpha A} = 0, \tag{5.16a}$$

$$\partial^\alpha \psi_{\alpha A} = 0, \tag{5.16b}$$

$$\partial_\alpha \psi_{\mu A} - \partial_\mu \psi_{\alpha A} = 0. \tag{5.16c}$$

This last equation implies the existence (local, at least) of a spinorial field ψ_A such that $\psi_{\mu A} = \partial_\mu \psi_A$. Thus the above system becomes

$$(\gamma^\mu \partial)_\mu \psi = \phi, \tag{5.17a}$$

$$\Box\psi = 0, \tag{5.17b}$$

where ϕ is a constant spinor. This set describes a noninteracting massless spin-$\frac{1}{2}$ particle if one imposes the extra condition $\phi = 0$.

The second field is described by three equations:

$$\partial^\alpha \psi_{[\alpha\nu]A} = 0, \tag{5.18a}$$

$$\partial_\alpha \psi_{[\beta\mu]A} + \partial_\mu \psi_{[\alpha\beta]A} + \partial_\beta \psi_{[\mu\alpha]A} = 0, \tag{5.18b}$$

$$(e_1^\alpha \partial_\alpha)\psi_{[\mu\nu]A}\sigma_1^{\mu\nu A} = 0. \tag{5.18c}$$

This system describes a massless spin-$\frac{3}{2}$ field, and the additional spin values are excluded by adequate gauge

conditions. The Bianchi-like relation (5.18c) implies the (local, at least) existence of a potential $\phi_{\alpha A}$ such that $\psi_{[\beta\mu]A} = \partial_\beta \phi_{\mu A} - \partial_\mu \phi_{\beta A}$. Taking this into (5.18c) one gets $\partial_\beta(\nabla_1 \phi_{\mu A}) - \partial_\mu(\nabla_1 \phi_{\beta A}) = 0$, where $\nabla_1 = e_1^\alpha \partial_\alpha$. This zero curl implies that there is, locally at least, a ϕ_A such that

$$(e_1^\alpha \partial_\alpha)\phi_{\mu A} = \partial_\mu \phi_A. \tag{5.19a}$$

On the other side, (5.18a) implies

$$\Box\phi_{\mu A} - \partial_\mu(\partial^\alpha \phi_{\alpha A}) = 0. \tag{5.19b}$$

The extra conditions are thus

$$\phi_A = \text{constant}, \tag{5.20a}$$

$$\partial^\alpha \phi_{\alpha A} = 0, \tag{5.20b}$$

which together with (5.19a) form a description for a massless spin-$\frac{3}{2}$ field. Note that (5.20) is implied by (5.13) and (5.14), when we pass there to the massless limit.

A system similar to (5.13) and (5.14) for the massive spin-$\frac{1}{2}$ field may be obtained from (5.12) with $\psi = -m\phi_A(1 \odot \sigma^A) + \phi_{\mu A}(\gamma^\mu \odot \sigma^A)$.

b. The spin-2 field and the weak gravitational field: A massive spin-2 field is given by

$$(e_1^\alpha \partial_\alpha + m)\psi = 0, \tag{5.21a}$$

$$\psi = -m\psi_{\mu\nu}(\gamma^\mu \odot \gamma^\nu) + \psi_{[\mu\nu]\rho}(\gamma^{\mu\nu} \odot \gamma^\rho) - (1/4m)\psi_{[\mu\nu][\rho\sigma]}(\gamma^{\mu\nu} \odot \gamma^{\rho\sigma}), \tag{5.21b}$$

$$m^2\psi_\rho + \tfrac{1}{4}\psi^{\rho\sigma}_{\rho\sigma} = 0. \tag{5.21c}$$

$\psi \in \mathrm{ReC}_{4(+)} \odot \mathrm{ReC}_{4(+)}$, with zero trace (5.21c). The associated tensor equations are:

$$(m)\psi^\rho_\rho = 0, \tag{5.22a}$$

$$(m)\partial^\alpha \psi_{\alpha\rho} = 0, \tag{5.22b}$$

$$(m)\psi_{[\mu\nu]\rho} = (m)(\partial_\mu \psi_{\nu\rho} - \partial_\nu \psi_{\mu\rho}), \tag{5.22c}$$

$$\partial^\alpha \psi_{[\alpha\nu]\rho} = m^2 \psi_{\nu\rho}, \tag{5.22d}$$

$$\partial_\alpha \psi_{[\mu\nu]\rho} + \partial_\nu \psi_{[\alpha\mu]\rho} + \partial_\mu \psi_{[\nu\alpha]\rho} = 0, \tag{5.22e}$$

$$(m)\psi_{[\mu\nu][\rho\sigma]} = (m)(\partial_\mu \psi_{\nu[\rho\sigma]} - \partial_\nu \psi_{[\rho\sigma]}), \tag{5.22f}$$

$$\partial^\alpha \psi_{[\alpha\nu][\rho\sigma]} = m^2 \psi_{\nu[\rho\sigma]}, \tag{5.22g}$$

$$\partial_\alpha \psi_{[\mu\nu][\rho\sigma]} + \partial_\nu \psi_{[\alpha\mu][\rho\sigma]} + \partial_\mu \psi_{[\nu\alpha][\rho\sigma]} = 0. \tag{5.22h}$$

This is the complete Fierz[1] system for massive spin-2 fields [Eq. (5.22a) is obtained through contraction from (5.21c) and (5.22f)].

The massless system is implied by the set

$$(e_1^\alpha \partial_\alpha)\psi = 0, \tag{5.23a}$$

$$\psi = -\psi_{\mu\nu}(\gamma^\mu \odot \gamma^\nu) + \psi_{[\mu\nu]\rho}(\gamma^{\mu\nu} \odot \gamma^\rho) - \tfrac{1}{4}\psi_{[\mu\nu][\rho\sigma]}(\gamma^{\mu\nu} \odot \gamma^{\rho\sigma}), \tag{5.23b}$$

$$\psi^\rho_\rho + \tfrac{1}{4}\psi^{\rho\sigma}_{\rho\sigma} = 0. \tag{5.23c}$$

This will imply an uncoupling of (5.22) that reveals the various tensor components as fields satisfying spin-0, spin-1, and spin-2 equations. One begins with

$$\partial^\alpha \psi_{\alpha\nu} = 0, \tag{5.24a}$$

$$\partial_\alpha \psi_{\mu\nu} - \partial_\mu \psi_{\alpha\nu} = 0. \tag{5.24b}$$

Here ν appears as an "idle" index, since it isn't operated or contracted in the above set—it only serves to enumerate equations in the system. We then write $\psi_{\mu\nu} = \partial_\mu \phi_{(\nu)}$ due to (5.24b), and (5.24a) becomes

$$\Box\phi_{(\nu)} = 0, \tag{5.24c}$$

that is, four noninteracting spin-0 fields. Equation (5.23) also entails

$$\partial^\alpha \psi_{[\alpha\nu]\rho} = 0, \tag{5.25a}$$

$$\partial_\alpha \psi_{[\mu\nu]\rho} + \partial_\nu \psi_{[\alpha\mu]\rho} + \partial_\mu \psi_{[\nu\alpha]\rho} = 0, \tag{5.25b}$$

$$\partial_\alpha \psi_{[\mu\nu]\rho} - \partial_\rho \psi_{[\mu\nu]\alpha} = 0. \tag{5.25c}$$

This system is Maxwell's when given adequate additional conditions. For (5.25c) implies that there is a $\phi_{[\mu\nu]}$ so that $\psi_{[\mu\nu]\rho} = \partial_\rho \phi_{[\mu\nu]}$. And getting this into (5.25a) and (5.25b), one has

$$\partial_\rho(\partial^\alpha \phi_{[\alpha\nu]}) = 0, \tag{5.26a}$$

$$\partial_\rho(\partial_\alpha \phi_{[\mu\nu]} + \partial_\nu \phi_{[\alpha\mu]} + \partial_\mu \phi_{[\nu\alpha]}) = 0. \tag{5.26b}$$

Choosing the integration constants to be zero, (5.26) becomes the Maxwellian system. The last set of equations implied by (5.23) is

$$\partial^\alpha \psi_{[\alpha\nu][\rho\sigma]} = 0, \tag{5.27a}$$

$$\partial_\alpha \psi_{[\mu\nu][\rho\sigma]} + \partial_\nu \psi_{[\alpha\mu][\rho\sigma]} + \partial_\mu \psi_{[\nu\alpha][\rho\sigma]} = 0. \tag{5.27b}$$

It is a linearized version of Lichnerowicz's set of equations for the gravitational field.[16] Note that the traceless condition (5.23c) now has the following meaning: If we choose $\psi^{[\rho\sigma]}_{[\rho\sigma]} = 0$, then $\psi^\rho_{(\rho)} = 0$ marks the linear dependence among the massless spin-0 fields associated to this weak gravitational field.

D. The general case

We will just indicate the irreducible expressions for ψ. In the integral spin case it is given by

$$\psi = \sum_{k=0}^{k=s} \binom{s}{k} (\tfrac{1}{2})^k (-m)^{s-k} \psi_{\mu\nu\ldots\lambda[\alpha\beta][\delta\xi]\ldots[\rho\sigma]} \cdot (\gamma^\mu \odot \gamma^\nu \odot \cdots \odot \gamma^\lambda \odot \gamma^{\alpha\beta} \odot \gamma^{\delta\xi} \odot \cdots \odot \gamma^{\rho\sigma}), \tag{5.28}$$

where the $\binom{s}{k}$ are combinatorial numbers; (5.28) has $s-k$ symmetric indices and k pairs of skew-symmetric indices, s being the particle's spin. In the half integral case [spin-$(s-\frac{1}{2})$] one symmetric index is substituted by a spinorial index, and correspondingly the tensor product

$$\psi = \sum_{k=0}^{k=s} \binom{s}{k} (\tfrac{1}{2})^k (-m)^{s-k} \psi_{\mu\nu\ldots\lambda A[\alpha\beta][\delta\xi]\ldots[\rho\sigma]} \cdot \gamma^\mu \odot \gamma^\nu \odot \cdots \odot \gamma^\lambda \odot \sigma^A \odot \gamma^{\alpha\beta} \odot \gamma^{\delta\xi} \odot \cdots \odot \gamma^{\rho\sigma}. \tag{5.29}$$

The irreducible nature of the wavefunction ψ implies that there is also a zero-trace condition like (5.21c) in the general case. Equations (5.28) and (5.29) belong to the product $C_4 \otimes C_4 \otimes \cdots \otimes C_4$, where the Clifford—Dirac anticommutation conditions are represented by the e_i^μ, $[e_i^\mu, e_i^\nu]_+ = 2g^{\mu\nu}$, $e_i^\mu = 1 \otimes 1 \otimes \cdots \otimes \gamma^\mu \otimes \cdots \otimes 1$, s being the number of factors in the direct product and γ^μ being at the ith position.

Some remarks must be made before we conclude.

(i) The above formulation has been developed only for *real* fields; the extension to the complex case is straight-

forward, since the differential Dirac operator $(e_i^\mu \partial_\mu + m)$ is a *real* operator (see the Appendix), and so real and imaginary components won't mix; given two irreducible and real fields ψ, ψ', belonging to the same representation class, the associated complex field is thus $\Psi = \psi + i\psi'$.

(ii) We must also emphasize that when passing to the massless limit from the massive equations one gets the whole set of field equations together with adequate gauge conditions that exclude unwanted values of the spin.

(iii) Also, analysis of the massless field equations as done above shows that the potentials for high spin describe lower spin fields; this parallels a similar result by Penrose[17] in a two-spinor formalism.

(iv) We have obtained higher-spin fields from direct products of lower spin fields. Our technique is thus closely related to de Broglie's *méthode de fusion*[18]; they differ however in that getting direct products has been for us just a way of obtaining irreducible representations for the Dirac-like Lagrangian (5.1); also the synthesis of spin-1 from spin-$\frac{1}{2}$ is obtained through a Clifford product of two minimal ideals in the algebra, and not through a direct product of the same ideals.

(v) The Dirac-like operator $(e_i^\mu \partial_\mu) = \nabla_i = d_i - \delta_i$, where d_i is the partial exterior derivative and δ_i its Hodge dual.[19] The above formalism may be translated into the language of exterior forms and it would be interesting to look for a connection between field equations (and the restrictions of a topological nature they impose on the fields) and the global structure of the manifolds where the fields are defined.

Lastly (vi) we have given a direct if not so elegant proof of Gelfand—Yaglom's theorem[20]: The only finite-dimensional Dirac-like equation without redundant components is the Dirac equation for spin-$\frac{1}{2}$ fields (the observation at the end of Sec. V C *b* shows that even that field can have a redundant-component formulation here).

VI. CONCLUSION

We have explored some features of Clifford algebras that lead to a very simple classification technique for irreducible representations of the Lorentz group; when applied to the search of representations of a Dirac-like Lagrangian, they entail well-known equations for high spin fields. Our proof technique has been adequately rigorous in Secs. II—IV of the present paper; it is frankly heuristic in Sec. V for the sake of clarity. However the results we obtain there can be given a solid foundation through the use of some well-known techniques[21]; we have avoided rigor so as not to make the exposition unnecessarily long.

ACKNOWLEDGMENTS

I must thank my colleagues from the Research Group on Symmetries, S. M. Abrahão, A. F. F. de Amaral, and Professor A. da Silveira for their many suggestions and comments about the present paper, as well as J. C. do Nascimento.

APPENDIX

We use a nonstandard Dirac operator $(\gamma^\mu \partial_\mu + m)$ in order to avoid imaginaries. It is thus obtained: From the Klein—Gordon equation

$$\left(-\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2 - \frac{m_0^2 c^2}{\hbar^2}\right)\psi(\mathbf{x}, t) = 0, \tag{A1}$$

defining the d'Alembertian with $a+2$ signature,

$$\Box^2 - m^2 = -\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \nabla^2 - m^2 = -\partial_0^2 + \partial_1^2 + \partial_2^2 + \partial_3^2 - m^2$$

$$(m = m_0 c/\hbar^1), \tag{A2}$$

we get a linear differential operator through

$$(\gamma^\alpha \partial_\alpha - m)(\gamma^\beta \partial_\beta + m) = \Box^2 - m^2 = -\partial_0^2 + \partial_1^2 + \partial_2^2 + \partial_3^2 - m^2 \tag{A3}$$

which implies the Clifford—Dirac condition

$$[\gamma^\alpha, \gamma^\beta]_+ = 2\dot{g}^{\alpha\beta}, \tag{A4}$$

where $\dot{g}^{\alpha\beta}$, the diagonalized metric $\dot{g} = \mathrm{diag}(-1, +1, +1, +1)$.

*Partially supported by FINEP.

[1]M. Fierz, Helv. Phys. Acta **12**, 3 (1939); W. Pauli and M. Fierz, Proc. Roy. Soc. (London) **A173**, 211 (1939); A. da Silveira, Phys. Rev. **97**, 1144 (1955); Nuovo Cimento **3**, 513 (1956); G.S. Guralnik and T.W. Kibble, Phys. Rev. **139**, 712B (1965); S. Kamefuchi and Y. Takahashi, Nuovo Cimento A **44**, 1 (1966); C. Fisk and W. Tait, J. Phys. A **6**, 3 (1973); M. Kawasaki, M. Kobayashi, and Y. Mori, Lett. Nuovo Cimento **14**, 611 (1975).

[2]E. Durand, Phys. Rev. D **11**, 3405 (1975).

[3]V. Bargmann and E.P. Wigner, Proc. Nat. Acad. Sci. **34**, 211 (1948).

[4]E.M. Corson, *Introduction to Tensors, Spinors and Relativistic Wave Equations* (Blackie, London, 1955).

[5]F. Sauter, Z. Phys. **63**, 803 (1930); **64**, 295 (1930); S. Teitler, Supp. Nuovo Cimento **3**, 1 (1965); F.A. Doria, "Spinors as Objects Defined Within Clifford Algebra," M.Sc. thesis, Centro Bras. de Pesq. Fis. Rio, Brazil (1973), in Portuguese.

[6]W. Rarita and J. Schwinger, Phys. Rev. **60**, 61 (1941).

[7]See Ref. 5 above.

[8]R. Brauer and H. Weyl, Am. J. Math. **57**, 425 (1935); H. Boerner, *Representations of Groups* (North-Holland, Amsterdam, 1970).

[9]See S. Teitler, Ref. 5, and T. Kahan, *Theory of Groups in Classical and Quantum Physics. I.* (Oliver and Boyd, Edinburgh, 1965).

[10]Unpublished calculations by A. da Silveira, A.F.F. do Amaral, and S.M. Abrahão.

[11]M. Riesz, Comptes Rendus du Xe. Congrès des Math. Scand. (Copenhagen, 1946) and Comptes Rendus du XIIe. Congrès des Math. Scand. (Lund, 1953).

[12]D. Hestenes, *Space—Time Algebra* (Gordon and Breach, New York, 1966).

[13]C. Cercignani, J. Math. Phys. **8**, 3 (1967).

[14]R.J. Duffin, Phys. Rev. **54**, 114 (1938); N. Kemmer, Proc. Roy. Soc. (London) **A173**, 91 (1939).

[15]D. Lurié, *Particles and Fields* (Interscience, New York, 1968).
[16]A. Lichnerowicz, Propagateurs et Commutateurs en Relativité Générale, Publ. Math. no. 10, Paris (1961); F. A. Doria, Lett. Nuovo Cimento **14**, 480 (1975); **7**, 153 (1973) introduces an equation similar to (5.23) which however does not give the correct spin value due to a condition that entails a zero curvature tensor. See also C. A. Coulter, Nuovo Cimento B **7**, 284 (1972), for a fairly complete discussion of Eqs. (5.24)—(5.27).
[17]R. Penrose, Proc. Roy. Soc. (London), **A284**, 159 (1965).
[18]L. de Broglie, *Theorie Générale des Particules à Spin* (Gauthier-Villars, Paris, 1954).
[19]This connection between the Clifford—Dirac differential operator and exterior algebra is a consequence of the deep interdependence between Clifford and exterior algebras. It has been thoroughly explored by M. F. Atiyah, R. Bott, and A. Shapiro, Topology Suppl. 3, 1 (1964). For the partial operators d_i, δ_i, see W. Greub, S. Halperin, and R. Vanstone, *Connections, Curvature and Cohomology* (Academic, New York, 1971).
[20]I. M. Gel'fand and A. M. Yaglom, Zh. Eksp. Teor. Fiz. **18**, 703 (1948).
[21]See Corson, Ref. 4 and Boerner Ref. 8.

Generalized invariants for the time-dependent harmonic oscillator

N. J. Günther and P. G. L. Leach

Department of Applied Mathematics, LaTrobe University, Bundoora/Victoria, Australia 3083
(Received 27 April 1976)

A generalized class of invariants, $I(t)$, for the three-dimensional, time-dependent harmonic oscillator is presented in both classical and quantum mechanics. For convenience a simple notation for types of harmonic oscillator is introduced. Two interpretations, one in terms of angular momentum and the other employing a canonical transformation, are offered for $I(t)$. An invariant symmetric tensor, $I_{mn}(t)$, is constructed and shown to reduce to Fradkin's invariant tensor for time-independent systems. The usual SU(3) (compact) or SU(2,1) (noncompact) is shown to be a noninvariance group for the time-dependent oscillator with $S\{U(2)\otimes U(1)\}$ as the invariance subgroup. Extensions to anisotropic systems and the singular quadratic perturbation problem are discussed.

1. INTRODUCTION

The study of the time-dependent harmonic oscillator has been greatly advanced in the one-dimensional case by the work of Lewis[1,2] and Riesenfeld.[3] Lewis showed that a previously known adiabatic invariant[4] was in fact an exact invariant for the time-dependent harmonic oscillator. Jointly, Lewis and Riesenfeld extended the application of this invariant to quantum mechanics. For the Hamiltonian

$$H=\tfrac{1}{2}p^2+\tfrac{1}{2}\omega^2(t)\,x^2, \tag{1}$$

where p and x are conjugate canonical coordinates, Lewis showed that

$$I=\tfrac{1}{2}\{\rho^{-2}x^2+(\rho p-\dot{\rho}x)^2\} \tag{2}$$

is an exact invariant when the auxiliary function $\rho(t)$ satisfies

$$\ddot{\rho}+\omega^2(t)\rho=\rho^{-3}. \tag{3}$$

Lewis attempted to provide an interpretation of I by pointing out that it is, to within an arbitrary multiplicative constant, the most general homogeneous quadratic invariant possible for the Hamiltonian given by Eq. (1). If a similar invariant was constructed from two linearly independent solutions of

$$\ddot{x}+\omega^2(t)x=0, \tag{4}$$

then I would be identical to it provided ρ was such that it satisfied (3). The connection between the solutions of (3) and (4) was first reported by Pinney.[5]

However, even with the above identification, Lewis claimed only limited progress[1] in understanding the nature of I. A more natural and physical interpretation has been suggested by Eliezer and Gray[6] in terms of a two-dimensional auxiliary motion. If (4) is regarded as one of the Cartesian equations for the motion of a two-dimensional isotropic time-dependent harmonic oscillator, (3) is the radial equation scaled to make the angular momentum of unit magnitude. The invariant I is proportional to the conserved angular momentum of the auxiliary motion.

In an independent development, Fradkin,[7,8] following the work of Jauch and Hill,[9] constructed a conserved symmetric tensor operator

$$H_{mn}=\tfrac{1}{2}(p_m p_n+\omega^2 q_m q_n) \tag{5}$$

for the three-dimensional time-independent harmonic oscillator. Fradkin found that this operator plays a role for the oscillator identical to that of the Runge-Lenz vector for the Kepler problem[10,11] and, together with the angular momentum, provides a suitable representation for the Lie algebra su(3).

In this paper we generalize the work of Lewis and Riesenfeld to the corresponding problem in three dimensions. A more precise interpretation than that given by Eliezer and Gray is provided for I in terms of angular momentum. Another interpretation using canonical transformations is made. The form of I suggests the construction of a symmetric tensor which reduced to that of Fradkin's under appropriate constraints on $\omega(t)$ and $\rho(t)$. The invariance group associated with the time-dependent oscillator is shown to be SU(2) in contrast to SU(3) for the time-independent problem. We conclude by discussing briefly the extension of our results to the singular perturbation problem in quantum mechanics[12,13] and to the anisotropic oscillator.

2. NOTATION

Cumbrous expressions such as "three-dimensional time-dependent anisotropic harmonic oscillator" are avoided by employing the symbolism below:

n: n-dimensional, $n=1,2,3,\ldots$,

(t): time-dependent, time independence has no (t),

H: isotropic harmonic oscillator,

A: anisotropic harmonic oscillator,

p (subscript): inclusion of a singular perturbation potential proportional to r^{-2}.

In this notation, the expression above becomes $3A(t)$. Units are chosen such that $m=1=\hbar$. The usual convention for summation over an explicitly repeated index applies unless otherwise stated. Lower case Roman indices take the values 1,2,3 and Greek indices take the values 1,2,3,4. Groups are identified by upper case letters and algebras by lower case letters, e.g., SU(3) has the algebra su(3).

 Copyright © 1977 American Institute of Physics

3. THE INVARIANT IN CLASSICAL MECHANICS

A. The invariant and associated tensor

For a $3H(t)$ with Hamiltonian

$$H=\tfrac{1}{2}p_m p_m+\tfrac{1}{2}\omega^2(t)\,q_m q_m, \tag{6}$$

the invariant corresponding to that of Lewis for the $1H(t)$ is

$$I=I_{mm}$$
$$=\tfrac{1}{2}\{\rho^{-2}q_m q_m+(\rho p_m-\dot\rho q_m)(\rho p_m-\dot\rho q_m)\}, \tag{7}$$

and ρ satisfies (3). The form of the invariant suggests a definition for a *tensor* invariant analogous to the Fradkin tensor for $3H$. We write

$$I_{mn}=\tfrac{1}{2}\{\rho^{-2}q_m q_n+(\rho p_m-\dot\rho q_m)(\rho p_n-\dot\rho q_n)\}, \tag{8}$$

whereby the following properties:

$$I_{mn}=I_{nm}, \tag{9}$$

$$\frac{d}{dt}(I_{mn})=0, \tag{10}$$

$$\mathrm{Tr}(I_{mn})=I, \tag{11}$$

$$I_{mm}I_{nn}-I_{mn}^2=\tfrac{1}{4}(\epsilon_{lmn}L_l)\ \text{(summation on } l \text{ only)}, \tag{12}$$

$$\det(I_{mn})=0, \tag{13}$$

$$I_{mn}L_n=0=L_m I_{mn}, \tag{14}$$

$$I_{mk}I_{kn}=II_{mn}+\tfrac{1}{4}(L_m L_n-\delta_{mn}L^2), \tag{15}$$

$$\mathrm{adj}(I_{mn})=\tfrac{1}{4}(L_m L_n), \tag{16}$$

$$q_m(I\delta_{mn}-I_{mn})q_n=\tfrac{1}{2}\rho^2 L^2, \tag{17}$$

$$p_m(I\delta_{mn}-I_{mn})q_n=\tfrac{1}{2}\rho\dot\rho L^2, \tag{18}$$

$$p_m(I\delta_{mn}-I_{mn})p_n=\tfrac{1}{2}(\rho^{-2}+\dot\rho^2)L^2, \tag{19}$$

confirm I_{mn} as a suitable tensor generalization of (7). In the above identities, the L_m are the components of angular momentum for the $3H(t)$ motion . With the exception of (18), the properties of I_{mn} are seen to be of the same form as those of the Fradkin tensor. We note that (18) and (19) follow from repeated time differentiation of (17), a remark which also applies to the corresponding Fradkin tensor identities. In the particular case when $\omega(t)$ is constant and $\dot\rho(t)$ is zero, (3) has the solution

$$\rho^2=\omega^{-1}. \tag{20}$$

Substitution of (20) in ωI_{mn} yields the H_{mn} of (5) so that the connection between the invariant tensor of $3H(t)$ and that of $3H$ is manifest.

B. Description of the motion

The eigenvalues of (I_{mn}) are

$$\lambda^{\pm}=\tfrac{1}{2}\{I\pm[(I^2-L^2)]^{1/2}\},\quad \lambda=0, \tag{21}$$

with associated eigenvectors (unnormalized)

$$\begin{pmatrix}\lambda^+I_{31}+\frac{1}{4}L_3L_1\\ \lambda^+I_{32}+\frac{1}{4}L_3L_2\\ \lambda^+I_{33}-\frac{1}{4}(L^2-L_3^2)\end{pmatrix},\quad \begin{pmatrix}\lambda^-I_{31}+\frac{1}{4}L_3L_1\\ \lambda^-I_{32}+\frac{1}{4}L_3L_2\\ \lambda^-I_{33}-\frac{1}{4}(L^2-L_3^2)\end{pmatrix},\quad \mathbf{L}. \tag{22}$$

Discussion of the relationship between the matrix (I_{mn}) and the motion is simplified by a suitable orthogonal transformation of the coordinates q_i to q_i' such that the axes are aligned with the eigenvectors and the transformed matrix is

$$(I'_{mn})=\begin{pmatrix}\frac{1}{2}\{I+[(I^2-L^2)]^{1/2}, & 0 & , 0\\ 0 & , \frac{1}{2}\{I-[(I^2-L^2)]^{1/2}\}, & 0\\ 0 & , 0 & , 0\end{pmatrix}. \tag{23}$$

In this coordinate system, (17) is

$$\frac{q_1'^2}{\rho^2\{I+[(I^2-L^2)]^{1/2}\}}+\frac{q_2'^2}{\rho^2\{I-[(I^2-L^2)]^{1/2}\}}=1. \tag{24}$$

In the particular case when $\omega(t)$ is constant and ρ is given by (20), (24) is the orbit equation of a $3H$, viz., an ellipse with semi-axes of length given by

$$a^2=\rho^2\{I+[(I^2-L^2)]^{1/2}\},$$
$$b^2=\rho^2\{I-[(I^2-L^2)]^{1/2}\}. \tag{25}$$

In the general $3H(t)$ case, (24) does not provide a complete description of the path of a particle since, although at any instant (24) describes an ellipse, for successive instants it describes a family of ellipses which are expanding or contracting in accordance with the variation in ρ. At all times, because the eigenvalues and eigenvectors of (I_{mn}) are invariant, the orientation of these ellipses is unaltered. The path which the particle takes depends upon the initial conditions and the time evolution of ρ. However, at any instant the particle must be on the ellipse prescribed by the value of ρ at that instant and be subject to the additional constraints (18) and (19).

C. Angular momentum interpretation

The auxiliary motion used by Eliezer and Gray to interpret the Lewis invariant for a $1H(t)$ was two-dimensional. Here we use a $4H(t)$ for the auxiliary motion, the $3H(t)$ being the three-dimensional projection of this motion. The angular momentum tensor is

$$L_{\mu\nu}=x_\mu p_\nu-x_\nu p_\mu, \tag{26}$$

with which we define

$$L^2=\tfrac{1}{2}L_{\mu\nu}L_{\mu\nu}. \tag{27}$$

The $4H(t)$ Hamiltonian,

$$H=\tfrac{1}{2}p_\mu p_\mu+\tfrac{1}{2}\omega^2(t)\,x_\mu x_\mu, \tag{28}$$

becomes, in hyperspherical polars,

$$H=\tfrac{1}{2}p_\rho^2+\tfrac{1}{2}L^2\rho^{-2}+\tfrac{1}{2}\omega^2\rho^2, \tag{29}$$

where

$$x_1=\rho\cos\theta,\quad x_2=\rho\sin\theta\cos\phi,$$
$$x_3=\rho\sin\theta\sin\phi\cos\psi,\quad x_4=\rho\sin\theta\sin\phi\sin\psi. \tag{30}$$

The space coordinates of the $3H(t)$ are x_2, x_3, and x_4 while x_1 is the space coordinate introduced with the auxiliary motion. Not only is the square of the angular momentum (27) conserved, but also

$$L_{34}^2 = p_\psi^2, \quad L_{23}^2 + L_{24}^2 = p_\phi^2 + p_\psi^2 \cot^2\phi,$$
$$L_{12}^2 + L_{13}^2 + L_{14}^2 = p_\theta^2 + p_\phi^2 \cot^2\theta + p_\psi^2 \cot^2\theta \operatorname{cosec}^2\phi. \quad (31)$$

In terms of the $4H(t)$ auxiliary motion, the invariant I given by (7) is

$$I = \tfrac{1}{2}\{L^2 + (L_{23}^2 + L_{24}^2 + L_{34}^2)\}$$
$$= L^2 - \tfrac{1}{2}(L_{12}^2 + L_{13}^2 + L_{14}^2) \quad (32)$$

and the Eq. (3) governing ρ is the radial equation (scaled to make L^2 unity). The invariant may be interpreted either as half the sum of the square of the angular momentum of the auxiliary motion and of the original motion or as the square of the angular momentum of the auxiliary motion less half the sum of the squares of the angular momentum components introduced with the auxiliary motion. This result, given by (32) in the case of a $3H(t)$, holds generally and reduces to the result of Eliezer and Gray for a $1H(t)$, since in that case the angular momentum is zero.

D. Canonical transformation interpretation

For a $3H(t)$ with Hamiltonian (6) and invariant (7), let the canonical coordinates (p_m, q_m) be transformed to (P_m, Q_m) such that

$$p_m = \rho^{-1}P_m + \dot\rho Q_m \qquad P_m = \rho p_m - \dot\rho q_m$$
$$\Longleftrightarrow \quad (33)$$
$$q_m = \rho Q_m \qquad Q_m = \rho^{-1} q_m,$$

where ρ is finite and nonzero. Then the coordinates (P_m, Q_m) are canonical since

$$[Q_m, Q_n]_{PB} = [\rho^{-1}q_m, \rho^{-2}q_m]_{PB} = 0, \quad (34)$$

$$[P_m, P_n]_{PB} = [\rho P_m - \dot\rho q_m, \rho P_n - \dot\rho q_n]_{PB} = 0, \quad (35)$$

$$[Q_m, P_n]_{PB} = [\rho^{-1}q_m, \rho p_n - \dot\rho q_n]_{PB} = \delta_{mn}. \quad (36)$$

For convenience we write the transformation as one of the third type[14] with generating function $F_3(Q_m, p_m, t)$ such that

$$q_m = -\frac{\partial F_3}{\partial p_m}, \quad P_m = -\frac{\partial F_3}{\partial Q_m}. \quad (37)$$

Combining (33) with (37), we see that

$$F_3(Q_m, p_m, t) = -\rho p_m Q_m + \tfrac{1}{2}\rho\dot\rho\, Q_m Q_m, \quad (38)$$

where an arbitrary function of time has been taken as zero. The transformed Hamiltonian is

$$H' = H + \frac{\partial F_3}{\partial t} = \tfrac{1}{2}\rho^{-2}\{P_m P_m + Q_m Q_m\}. \quad (39)$$

In the new coordinates,

$$I = \tfrac{1}{2}\{P_m P_m + Q_m Q_m\}. \quad (40)$$

We see that I has the form of the Hamiltonian of a $3H$ in the new coordinates. Further,

$$I_{mn} = \tfrac{1}{2}\{P_m P_n + Q_m Q_n\},$$

which is the Fradkin tensor for (40). The advantage of this coordinate system is that the explicit time dependence in the Hamiltonian is confined to a multiplicative factor. Furthermore, the components of angular momentum remain unchanged under the transformation since

$$Q_m P_n - Q_n P_m = q_m p_n - q_n p_m. \quad (41)$$

The two interpretations offered here have separate attractions. The one based on the work of Eliezer and Gray gives a natural origin to the equation defining (3). The canonical transformation approach which uses this equation leads to a ready explanation of the symmetry groups appropriate to this problem (see Sec. 4, Part B).

4. QUANTUM INVARIANT OPERATOR

A. Definition and properties

The $3H(t)$ Hamiltonian has the form of (6) in which p_m and q_m are now operators with the usual commutation relations

$$[q_m, q_n] = 0, \quad [p_m, p_n] = 0, \quad [q_m, p_n] = i\delta_{mn}. \quad (42)$$

We construct a tensor operator I_{mn} which has the same form as the classical tensor (8) with ρ being the classical function. The operator I_{mn} does not commute with the Hamiltonian, but it is an invariant since

$$\frac{d}{dt}\{I_{mn}(p, q, \rho, \dot\rho)\}$$
$$= [I_{mn}, H] + \dot\rho\frac{\partial}{\partial\rho}(I_{mn}) + \ddot\rho\frac{\partial}{\partial\dot\rho}(I_{mn}) = 0. \quad (43)$$

The algebraic properties of the operator I_{mn} are similar to those of the classical invariant. For completeness they are recorded[15]:

$$I_{mn} = I_{nm}, \quad (44)$$

$$I_{mn}L_n = 0 = L_m I_{mn}, \quad (45)$$

$$I_{mk}I_{kn} = I I_{mn} - \tfrac{1}{4}\{L_m L_n - \delta_{mn}L^2 + 2i[L_m, L_n] - 2\delta_{mn}\}, \quad (46)$$

$$\mathrm{Tr}(I_{mn}) = I,$$
$$I_{mm}I_{nn} - I_{mn}^2 = \tfrac{1}{4}\{(\epsilon_{lmn}L_l) + 1 \quad (47)$$

no summation on m, n),

$$q_m(I\delta_{mn} - I_{mn})q_n = \tfrac{1}{2}\rho^2 L^2, \quad (48)$$

$$p_m(I\delta_{mn} - I_{mn})q_n = q_m(I\delta_{mn} - I_{mn})p_n = \tfrac{1}{2}\rho\dot\rho L^2, \quad (49)$$

$$p_m(I\delta_{mn} - I_{mn})p_n = \tfrac{1}{2}(\rho^{-2} + \dot\rho^2)L^2. \quad (50)$$

B. Lie algebra for $3H(t)$

The conventional motion of a symmetry group relates to an exact invariance group of the Hamiltonian which incorporates conservation laws and selection rules. More recently, consideration has been given to the relevance of larger symmetry groups that are not themselves exact symmetries of the Hamiltonian, but which contain the usual invariance group as a proper subgroup.[16] As an illustration, for nH it has been

shown[17,18] that the noninvariance group SU($n+1$) (compact) or SU($n,1$) (noncompact) has the maximal compact subgroup S{U(n)⊗ U(1)}, the usual invariant symmetry group of the Hamiltonian, $H=(a_k^\dagger a_k+\frac{1}{2})$.

The $3H(t)$ exhibits similar behavior but with lower order symmetries involved. The relationship discussed in Sec. 3, Part D, of I to a $3H$ Hamiltonian and the preservation of angular momentum under the canonical transformation suggest that the symmetry group under I is SU(3). We form linear combinations of the six components of I_{mn} and the three components of $\mathbf{L}$ which generate a closed Lie algebra under I corresponding to the compact SU(3). The commutation relations are:

$$\begin{aligned}
&[I_{mn}, I]=0,\\
&[L_m, I]=0,\\
&[L_m, L_n]=i\epsilon_{lmn}L_l,\\
&[L_k, I_{mn}]=i(\epsilon_{kml}I_{ln}+\epsilon_{knl}I_{ml}),\\
&[I_{kl}, I_{mn}]=i(\epsilon_{lnr}\delta_{kl}+\epsilon_{lmr}\delta_{kn}\\
&\qquad +\epsilon_{knr}\delta_{lm}+\epsilon_{kmr}\delta_{ln})L_r.
\end{aligned}\tag{51}$$

The eight generators may be represented as

$$\begin{aligned}
&2\sqrt{3}H_1=L_3,\quad 12N_2=I_{11}+I_{22}-2I_{33},\\
&4\sqrt{3}E_1^\pm=L_1\pm iL_2-I_{13}\pm iI_{23},\\
&4\sqrt{3}E_2^\pm=L_1\mp iL_2-I_{13}\mp iI_{23},\\
&4\sqrt{6}E_3^\pm=I_{11}-I_{22}\pm 2iI_{12}.
\end{aligned}\tag{52}$$

These generators furnish the standard form of the SU(3) commutation relations, viz.,

$$\begin{aligned}
&[H_1,H_2]=0,\quad [H_1,E_1^\pm]=\pm(i/2\sqrt{3})E_1^\pm\\
&[H_1,E_2^\pm]=\pm(i/2\sqrt{3})E_2^\pm,\quad [H_2,E_1^\pm]=\pm(i/2)E_1^\pm\\
&[H_2,E_2^\pm]=\pm(i/2)E_2^\pm,\quad [H_1,E_3^\pm]=\pm(i/\sqrt{3})E_3^\pm\\
&[H_2,E_3^\pm]=0,\\
&[E_1^-,E_2^-]=[E_1^-,E_3^+]=[E_2^+,E_3^-]=0,\\
&[E_1^+,E_1^-]=(i/2\sqrt{3})H_1+(i/2)H_2,\\
&[E_2^+,E_2^-]=-(i/2\sqrt{3})H_1+(i/2)H_2,\\
&[E_3^+,E_3^-]=(i/\sqrt{3})H_1,\quad [E_1^+,E_2^+]=(i/\sqrt{6})E_3,\\
&[E_1^+,E_3^-]=(i/\sqrt{6})E_2^-,\quad [E_1^-,E_3^+]=-(i/\sqrt{6})E_2.
\end{aligned}\tag{53}$$

Evidently SU(3) is not an invariance symmetry group for the $3H(t)$. That SU(2) is the maximal compact symmetry subgroup, is confirmed by the commutivity of the angular momentum with the $3H(t)$ Hamiltonian. We emphasize that, since the metrics $g_{ij}=\mathrm{diag}(+++)$, preserved by SU(3) and $g_{ij}=\mathrm{diag}(++-)$, preserved by SU(2,1) are related by the Weyl unitary trick (WUT),[19] the algebras su(3) and su(2,1) are also related by a WUT and hence their complex extensions are identical to sl(3). For this reason, both SU(3) and SU(2,1) are noninvariance groups for the $3H(t)$ having finite dimensional and infinite dimensional unitary irreducible representations respectively furnished by the eigenvalues of the invariant, I, for $3H(t)$.

C. Raising and lowering operators

The invariant tensor operator I_{mn} permits the construction of the raising and lowering operators

$$\begin{aligned}
a_m^\dagger&=(1/\sqrt{2})\{\rho^{-1}q_m-i(\rho p_m-\rho q_m)\}\\
a_m&=(1/\sqrt{2})\rho^{-1}q_m+i(\rho p_m-\rho q_m)\},
\end{aligned}\tag{54}$$

such that

$$I_{mn}=\tfrac{1}{2}(a_m^\dagger a_n+a_m a_n^\dagger).\tag{55}$$

Closure under commutation with the (number) generators

$$N_m=a_m^\dagger a_m\tag{56}$$

and the identity $\mathcal{J}$ realizes a nonsemisimple Lie algebra which is a generalization of $\mathcal{A}(4)$, the usual algebra of $1H$.[19] In particular, it shows that the eigenvalues of I are $(N_m+\frac{3}{2})\hbar$. The tensor

$$J_{mn}=a_m^\dagger a_n-a_m a_n^\dagger\tag{57}$$

has the relations

$$\begin{aligned}
&I_{mn}=\tfrac{1}{2}(J_{mn}+J_{nm}),\\
&[I,J_{mn}]=0,\\
&L_k=\tfrac{1}{2}i\epsilon_{kmn}J_{mn}.
\end{aligned}\tag{58}$$

5. DISCUSSION

A. Singular perturbation

Two problems which are logical extensions of this work are the $3A(t)$ (see Sec. 5, Part B) and the $3H(t)$ with a singular perturbation of the form gr^{-2} [$g > -\frac{1}{2}(l+\frac{1}{2})^2$ to prevent collapse] in the Hamiltonian

$$H_p=\tfrac{1}{2}[p_m p_m+\omega^2(t)q_m q_m+g/r^2].\tag{59}$$

The one-dimensional form of this problem has been solved using two different approaches,[12,13] both of which rely on the ρ equation. That the eigenfunctions of I are significant factors in the solutions has also been noted.[12] For the three-dimensional problem, the appropriate invariant is

$$I=\sum_{m=1}^{3}\tfrac{1}{2}\{\rho^{-2}q_m^2+(\rho p_m-\rho q_m)^2+g\rho^2 q_m^2/r^4\},\tag{60}$$

although each component is *not* an invariant. Since the angular momentum is conserved and commutes with H, the appropriate invariance symmetry group for the $3H_p(t)$ is SU(2).

Under the canonical transformation (33), we see that

$$H_P'=\tfrac{1}{2}\rho^{-2}(P_m P_m+Q_m Q_m+g/R^2),\tag{61}$$

$$I=\tfrac{1}{2}(P_m P_m+Q_m Q_m+g/R^2),\tag{62}$$

i.e., in the new coordinate system, I corresponds to the Hamiltonian of a $3H_P$. As $\mathbf{L}$ is invariant under the

transformation and commutes with I, SU(2) is a symmetry group for I. As there are no other generators, there is no higher order symmetry group.

B. Anisotropic oscillator

The $3A(t)$ has Hamiltonian

$$H=\tfrac{1}{2}(p_m p_m+\omega_m^2(t)q_m^2). \tag{63}$$

Such a time-dependent system does not appear to have attracted much attention although the related time-independent problems, $2A$[20] and nA[21] have seen considerable controversy. The invariant for (62) is

$$I=\tfrac{1}{2}\{\rho_m^{-2}q_m^2+(\rho_m p_m-\dot{\rho}_m q_m)^2\} \tag{64}$$

where each ρ_m satisfies

$$\ddot{\rho}_m+\omega_m^2(t)\rho_m=\rho_m^{-3}. \tag{65}$$

The four-dimensional auxiliary motion interpretation does not apply to this invariant. However, canonical transformations of the type (33) with separate ρ's for each pair (p_m, q_m) yield

$$H'=\tfrac{1}{2}\rho_m^{-2}(P_m^2+Q_m^2) \tag{66}$$

and

$$I=\tfrac{1}{2}(P_m P_m+Q_m Q_m). \tag{67}$$

Each element I_{mm} (not summed) is an invariant and commutes with I. It is not possible to construct invariants of the form I_{mn} $(m\neq n)$. Raising and lowering operators

$$a_m^\dagger=(1/\sqrt{2})[\rho_m^{-1}q_m-i(\rho_m p_m-\dot{\rho}_m q_m)] \tag{68}$$
$$a_m=(1/\sqrt{2})[\rho_m^{-1}q_m+i(\rho_m p_m-\rho_m q_m)] \quad \text{(no summation on m)},$$

can be constructed. As for the $3H(t)$, $3A(t)$ possesses an $\mathcal{K}(4)$ algebra for each I_{mm} and the eigenvalues of I, as is evident from (67), are $(N_m+\frac{3}{2})\hbar$.

For both the $3H_p(t)$ and $3A(t)$, our investigations indicate an overall lowering of symmetry, a result which confirms intuitive expectations. For example, in the $3A(t)$, one would anticipate the removal of geometrical symmetry by anisotropy to be reflected in a lower order algebra. The noncommutation of certain linear combinations of the components of I_{mn} with the time-dependent Hamiltonian leads to the determination of the noninvariance symmetry group for the Hamiltonian in the case of $3H(t)$. What happens for the $3H_p(t)$ and $3A(t)$ is not yet clear to us and will be the subject of further research.

6. CONCLUSION

In this paper we have presented a generalized class of invariants $I(t)$ and constructed an invariant symmetric tensor $I_{mn}(t)$ for the system $3H(t)$ in both classical and quantum mechanics. The utility of these invariants in determining analytic solutions for problems of the type $nH(t)$ is manifest. Moreover, the work of Sec. 5 suggests that such invariants may provide a way of dealing with more complex Hamiltonians.

ACKNOWLEDGMENTS

We are grateful to Professor C.J. Eliezer for his stimulating comments and one of us (N.J.G.) acknowledges the Australian Government for financial assistance through a Commonwealth Postgraduate Award.

[1] H.R. Lewis, Jr., Phys. Rev. Lett. **18**, 510, 636 (1967).
[2] H.R. Lewis, Jr., J. Math. Phys. **9**, 1976 (1968).
[3] H.R. Lewis, Jr. and W.B. Riesenfeld, J. Math. Phys. **10**, 1458 (1969).
[4] M. Kruskal, J. Math. Phys. **3**, 806 (1962).
[5] E. Pinney, Proc. Am. Math. Soc. **1**, 681 (1950).
[6] C.J. Eliezer and A. Gray, S.I.A.M. **30**, 463 (1976).
[7] D.M. Fradkin, Am. J. Phys. **33**, 207 (1965).
[8] D.M. Fradkin, Prog. Theor. Phys. **37**, 798 (1967).
[9] J.M. Jauch and E.L. Hill, Phys. Rev. **57**, 641 (1940).
[10] W.R. Hamilton, Proc. R. Irish Acad. **3**, 344 (1847); C. Runge, *Vektoranalysis* (Dutton, New York, 1919), Vol. 1; W. Lenz, Z. Phys. **24**, 197 (1924).
[11] The Runge–Lenz vector for the Schrödinger–Coulomb Hamiltonian, $H=\frac{1}{2}\mathbf{p}^2-Ze^2/r$,

$$\mathbf{R}=\tfrac{1}{2}(\mathbf{L}\times\mathbf{p}-\mathbf{p}\times\mathbf{L})+Ze^2\hat{r},$$

has magnitude $R=a(1-\epsilon)$ and lies along the line $0R$ from the center of force to the perihelion of the ellipse

$$\frac{1}{r}=\frac{1}{l(l+1)}\{R\cos\theta+1\},$$

where

e = electron charge,
a = length of the semimajor axis,
l = angular momentum quantum number = 0, 1, 2, ···,
θ = plane polar angle measured from $0R$,

and
ϵ = eccentricity of the ellipse.

The constancy of $\mathbf{R}([H,\mathbf{R}]=0)$, may be viewed as an expression of the constancy of the orientation of the axes of the ellipse. A similar interpretation applies to H_{mn}. The commutation relations for $H\equiv E<0$,

$$[L_k, L_m]=i\epsilon_{kmn}L_n,$$

$$[R_k, L_m]=i\epsilon_{kmn}R_n,$$

$$[R_k, R_m]=i\epsilon_{kmn}(-2HZ^2e^{-4})L_n,$$

are an algebraic statement that the hydrogen atom possesses the dynamical symmetry $O(4)=SO(3)\otimes SO(3)$.
[12] P. Camiz *et al.*, J. Math. Phys. **12**, 2040 (1971) and Refs. cited therein.
[13] D.C. Khandekar and S.V. Lawande, J. Math. Phys. **16**, 384 (1975).
[14] E.C.G. Sudarshan and N. Mukunda, *Classical Dynamics: A Modern Perspective* (Wiley, New York, 1974), p. 58.
[15] The identities (46) and (47) differ from their classical counterparts (15) and (12) respectively. The additional terms which appear have coefficients either of $\hbar$ or $\hbar^2$ and so the difference is due to quantum mechanical effects and not to any change in the structure of the tensor I_{mn}.
[16] *Non-Compact Groups in Particle Physics*, edited by Y. Chow (Benjamin, New York, 1966).
[17] N. Mukunda, L. O'Raifeartaigh, and E.C.G. Sudarshan, Phys. Rev. Lett. **15**, 1041 (1965).
[18] E.C.G. Sudarshan, N. Mukunda, and L. O'Raifeartaigh, Phys. Lett. **19**, 332 (1965).
[19] R. Gilmore, *Lie Groups, Lie Algebras and Some of Their Applications* (Wiley, New York, 1974).
[20] Yu.N. Demkov, Zh. Eksp. Teor. Fiz. **44**, 2007 (1963) [Sov. Phys. JETP **17**, 1349 (1963)].
[21] I. Vendramin, Nuovo Cimento A **54**, 190 (1968).

Lattices of effectively nonintegral dimensionality

Deepak Dhar*

George W. Downs Laboratory of Physics, California Institute of Technology, Pasadena, California 91125
(Received 2 September 1976)

We construct a class of lattice systems that have effectively nonintegral dimensionality. A reasonable definition of effective dimensionality applicable to lattice systems is proposed and the effective dimensionalities of these lattices are determined. The renormalization procedure is used to determine the critical behavior of the classical XY model and the Fortuin–Kasteleyn cluster model on the truncated tetrahedron lattice which is shown to have the effective dimensionality $2\log 3/\log 5$. It is found that no phase transition occurs at any finite temperature.

INTRODUCTION

In a recent paper,[1] Nelson and Fisher defined the truncated tetrahedron lattice. The lattice is planar, multiply connected, and has a coordination number 3. Although admittedly rather artificial, it is of interest in the theory of phase transitions because it provides us with a lattice of effectively nonintegral dimensionality. Besides providing a physical basis and testing grounds for such formal techniques as the ϵ expansion,[2] the study of critical behavior of model Hamiltonians on such lattices may be expected to improve our understanding of the influence of dimensionality in determining the nature of phase transitions in physical systems.

The plan of this paper is as follows: In Sec. I we define what we mean by effective nonintegral space dimensionality of an infinite lattice. The proposed definition is different from that assumed in Ref. 1. In particular the space dimensionality of the truncated tetrahedron lattice using this definition is found to be $2\ln_5 3 \approx 1.3651$ and not $\ln_2 3 \approx 1.5851$ as proposed by Nelson and Fisher. Arguments are presented in favor of our definition. In Sec. II two examples of lattice systems with nonintegral dimensionality are given. One is the generalization of the truncated tetrahedron lattice to the truncated n-simplex lattice. The other is a planar multiply connected lattice of coordination number 3 which may be obtained by deleting some bonds from a planar square lattice. The dimensionality of these lattices is determined in Sec. III by deriving a functional equation for their characteristic functions. In Sec. IV the behavior of the classical XY model and the cluster model on the truncated tetrahedron lattice is analyzed using the renormalization procedure.

I. DEFINITION OF THE SPACE DIMENSIONALITY OF AN INFINITE LATTICE

Consider an infinite lattice. For our purpose a lattice is specified by its graph consisting of lattice points and undirected lines joining them called bonds. For simplicity we consider only one kind of bond. Two sites that have a bond in common are called nearest neighbors. How do we assign a dimensionality to an arbitrary infinite lattice? (The effective dimensionality of a finite lattice may be defined to be zero.) Any proposed definition of effective dimensionality should satisfy some elementary properties. It should agree with the conventional integral value of dimensionality for "regular" lattices. It should depend only weakly on the lattice in the sense that introduction or deletion of a finite number of lattice points or bonds should not change its value. And preferably it should satisfy some scaling relations between critical exponents.[3]

It is instructive to look at the familiar case when the space dimensionality d is integral. We consider a "simple cubic" lattice in d dimensions and consider a model in which a scalar displacement χ_i is associated with each site i and nearest neighbors are connected by harmonic springs of equal spring constants. Introduction of normal mode coordinates converts this problem into one of independent simple harmonic oscillators. Each normal mode is characterized by a wavenumber κ which is a d-dimensional vector lying within the first Brillouin zone. We also know that:

(i) For small wavenumbers κ, the frequency ω of the mode with wave vector κ, is approximately proportional to the magnitude of κ, i.e., $\omega^2 = \kappa^2 C^2$ for small κ^2.

(ii) The number of modes with $|\kappa| < \kappa$ is proportional to κ^d for small κ. The two facts together imply that the fractional number of modes with frequency less than ω is proportional to ω^d for small ω.

In making a transition to more general lattices with possibly nonintegral dimensionality, we may again define a nearest neighbor harmonic interaction model. It is difficult to say just what meaning may be assigned to a "nonintegral dimensional vector κ". At best we may say that low values of $|\kappa|$ correspond to slowly space varying modes and high values of $|\kappa|$ correspond to the modes where the spatial variation is large. We necessarily assume that something like (ii) is correct. Compare this, for instance, with the scaling property assumed by Wilson [Eqs. (A3) and (A9) in Ref. 4]. The validity of (i) is rendered plausible by the observation that the equation $\omega^2 = \kappa^2 C^2$ is just the Fourier transform of the equation $\partial^2\chi/\partial t^2 = C^2\nabla^2\chi$. If we identify χ_i as the displacement at site i and replace $\nabla^2\chi$ by its discrete analog $\nabla^2\chi_i \approx \sum_j(\chi_j - \chi_i)$ where j are sites neighboring i, we see that the equation correctly gives the equation of motion of χ_i in our model.

It is thus reasonable to postulate that the fractional number of modes of frequency less than ω is proportional to ω^d, for small ω, even if d is nonintegral. We adopt this as the definition of effective dimensionality of a given infinite lattice.

In practice, the determination of the fractional num-

 Copyright © 1977 American Institute of Physics

ber of modes below a frequency ω involves starting with a finite lattice of size N and then letting N tend to infinity. For an arbitrary lattice such a limit obviously does not exist. In statistical mechanics, we shall be only concerned with lattices which are sufficiently regular and all such thermodynamic limits will be assumed to exist.

The analysis given in Sec. III shows that the effective dimensionality of the truncated tetrahaedron lattice is $2 \log_5 3$.

Nelson and Fisher gave a different argument to determine the effective dimensionality of the truncated tetrahedron lattice. They define the diameter of the lattice as the maximum distance between any two lattice points as measured along bonds. Then it is easy to see that the diameter R of a rth order truncated tetrahedron is proportional to 2^r, while its volume V, which is equal to the total number of lattice points, goes like 3^r. Identifying the effective dimensionality d by the relation $V \propto R^d$, one finds that $d = \log_2 3$.

The two definitions of effective dimensionality are clearly not equivalent. The important difference is that while in Nelson and Fisher's definition, one considers the lattice as a whole, our proposed definition looks only at the low spatial frequency modes. Since the phase transitions are governed by the long range correlations, i.e., low frequency modes, we believe that our proposed definition is more relevant to the study of phase transitions.

For example, the modified rectangular lattice, described in the next section, has an effective dimensionality $\frac{3}{2}$ using our definition. If we used Nelson and Fisher's definition, the effective dimensionality of the lattice would be 2. However the behavior of model Hamiltonians like the Gaussian model or the Ising model on this lattice is very different from that on a two-dimensional square lattice. In particular, it may be shown that on this lattice, the Ising model shows no spontaneous magnetization or phase transition. Thus we may expect the assigned effective dimensionality to be less than 2. Similar arguments may be made for the truncated n-simplex lattice.

We remark that the effective dimensionality of a lattice has been defined with respect to a specific model (nearest neighbor harmonic spring interaction model). We expect the dimensionality of the lattice to be model independent and different definitions of effective dimensionality using other model Hamiltonians should give an identical value of dimensionality when it is calculated correctly from the scaling relations in terms of independently determinable critical exponents of the model. The quadratic interaction was chosen due to its simplicity.

In particular, the dimensionality of a lattice would not change if we introduced a second nearest neighbor spring interaction or any quadratic interaction of finite range for that matter. This may be proved rigorously using exact renormalization equations but is most easily seen by the application of first order perturbation theory. Let the first and second nearest neighbor spring constant be J and J' respectively, $J' \ll J$. The change in ω_i^2, the frequency of the ith mode is given by

$$\frac{d\omega_i^2}{dJ'} = \alpha\omega_i^2 + \text{higher order terms in } \omega_i^2,$$

where α is some constant depending on the lattice. This implies that the power law dependence of the density of states in the low frequency region remains unchanged.

II. EXAMPLES OF LATTICE SYSTEMS WITH NONINTEGRAL DIMENSIONALITY

In this section we describe the truncated n-simplex lattice and the modified rectangular lattice as examples of lattices with nonintegral dimensionality. Other lattices of this type are easy to construct. These lattices are so defined that exact renormalization equations may be written down for these systems.

The truncated n-simplex lattice is a simple generalization of the truncated tetrahedron lattice. The lattice is defined recursively. The graph of the zeroth order truncated n-simplex lattice is a complete graph on $(n+1)$ points. (A complete graph is a graph in which there exists a bond between every pair of lattice points.) The graph for the $(r+1)$th order lattice is obtained by replacing each of the vertices of the rth order graph by a complete graph on n points. Each of the new n points is connected to one of the lines coming to the original vertex. For the thermodynamic limit we let r tend to infinity. In Fig. 1 we illustrate this for the case $n=4$.

The case $n=1$ corresponds to the uninteresting case of mutually disconnected pairs of points. $n=2$ corresponds to the linear chain. For $n=3$ we get the truncated tetrahedron lattice. For higher values of n the lattices are nonplanar. It is easy to see that the rth order truncated n simplex lattice has $(n+1)n^r$ vertices and $[(n+1)n^{r+1}]/2$ bonds. Each lattice point has a coordination number n. Though the lattice is multiply connected, it is still possible to disconnect an arbitrarily large set of points from the rest of the lattice by just deleting

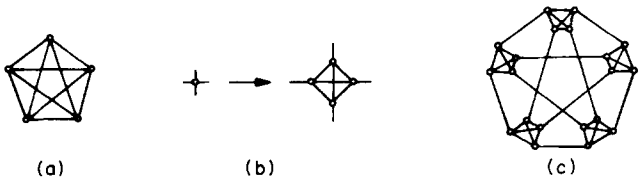


FIG. 1. (a) A complete graph on 5 points, zeroth order truncated 4-simplex lattice. (b) Replacement of single vertex by a complete 4 point graph. (c) Result of replacement on Fig. (a). First order truncated 4-simplex lattice.

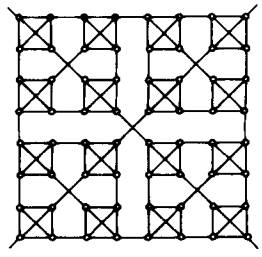


FIG. 2. A portion of the infinite truncated 4-simplex lattice.



FIG. 3. (a) Graph of a first order square. (b) Schematic representation of the graph of the $(r+1)$th order. The shaded squares denote the graph of rth order squares of which only the corner vertices are shown.

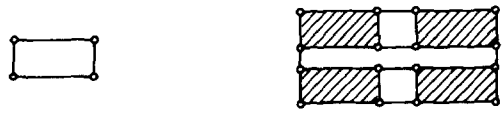


FIG. 4. (a) Graph of a first order rectangle. (b) Schematic representation of the graph of a $(r+1)$th order rectangle. The shaded rectangles denote graphs of rth order rectangles of which only the corner vertices are shown.

n bonds. We show later that the effective dimensionality of the truncated n-simplex lattice is $2\log n/\log(n+2)$.

The truncated 4-simplex lattice may be more conveniently drawn as in Fig. 2 as a square lattice with bonds connecting at most the "next nearest neighbor." The quotation marks indicate that the word is used in the Euclidean sense. By our definition all pairs of points that have a bond in common are nearest neighbors. The construction of the graph is explained in Fig. 3. We define the first order square as a complete graph on four points. An $(r+1)$th order square is obtained by joining four rth order squares together by bonds such that each rth order square is joined to the other rth order squares by bonds connecting corner points and each rth order square contributes one corner point to the $(r+1)$th order square. The graph in Fig. 2 is a third order square. It is easy to convince oneself that the graph of the infinite order square is topologically equivalent to the infinite truncated 4-simplex graph. Thus we may generate the truncated 4-simplex lattice from the graph of a two-dimensional infinite square lattice with "nearest neighbor" and "next nearest neighbor" bonds by selectively deleting some bonds. This deletion of bonds results in a change in the effective dimensionality of the lattice from 2 to $2\log 4/\log 6$.

We may similarly define a planar lattice to be called a modified rectangular lattice. This lattice has coordination number 3 and is multiply connected. The elementary cycles in the lattice are of length 4, 8, 16, 32, $\cdots$. The lattice is defined in Fig. 4. In this lattice the rth order rectangle does not posses the full permutation symmetry between the corner lattice points as did the rth order square in the previous case. This is because there is an asymmetry between the horizontal and vertical bonds and in the rth order rectangle $(r>1)$ there are more horizontal than vertical bonds. It is shown in Sec. III that the effective dimensionality of this lattice is $\frac{3}{2}$.

By varying n, in the truncated n-simplex lattice, we get lattices of effective dimensionality 0, 1, $2\ln_5 3$, $2\ln_6 4, \cdots$. However for any n the dimensionality of the lattice is less than 2. The same is true for the modified rectangular lattice or some other lattices which may be defined recursively. However it is easy to obtain lattices of higher dimensionality. Given two lattices L_1, L_2 we define their direct product lattice $L=L_1\times L_2$ as follows: For each ordered pair (l_1, l_2), where l_1 and l_2 are the lattice points in the lattices L_1 and L_2 respectively, we associate a unique lattice point $l\in L$. The point $l, l'\in L$, where $l=(l_1, l_2)$, $l'=(l_1', l_2')$, are nearest neighbors of each others in L iff

(i) $l_1=l_1'$ and (l_2 and l_2' are nearest neighbors in L_2) or

(ii) $l_2=l_2'$ and (l_1 and l_1' are nearest neighbors in L_1).

It is easy to see that effective dimensionality of L is the sum of dimensionalities of L_1 and L_2. For example, the direct product of a linear chain and a square lattice is a simple cubic lattice of dimensionality $3=1+2$.

This gives us a whole class of lattices with nonintegral dimensionality so that the effect of changing dimensionality on critical exponents etc., may be studied with much more detail than has been possible so far. By forming the direct product of the modified rectangular lattice with itself we get a lattice of effective dimensions 3. Verification that critical exponents for, say, the Ising model on this lattice are the same as for the simple cubic lattice would be an important test of the usefulness and relevance of our definition of dimensionality and of the strength of the universality hypothesis.

III. DETERMINATION OF THE EFFECTIVE DIMENSIONALITY

In this section we determine the effective dimensionality of lattices defined in the previous section by determining their frequency spectrum. The procedure is equivalent to determining the low temperature specific heat behavior of a Debye solid on the lattice. If the low temperature specific heat is proportional to T^d where T is the temperature, we identify d as the effective dimensionality of the lattice.

A. The truncated n simplex lattice

We define the spherical model on this lattice by assigning a continuous spin χ_i $(-\infty<\chi_i<\infty)$ to each of the sites i of the rth order truncated n-simplex lattice. The index i goes from 1 to $N_r=(n+1)n^r$. The nearest neighbor sites are assumed to have a ferromagnetic Ising interaction of strength J. The spherical constraint is taken care of by introducing a Lagrange parameter λ in the Hamiltonian so that the Hamiltonian of the system is

$$H=(J/2)\sum_{\text{nearest neighbors } ij}(\chi_i-\chi_j)^2+(\lambda J/2)\sum_{i=1}^{N_r}\chi_i^2. \qquad (1)$$

We get the partition function as

$$Z_r(\lambda)=\left(\prod_{i=1}^{N_r}\int_{-\infty}^{+\infty}d\chi_i\right)\exp(-\beta H) \tag{2}$$

where β is the inverse temperature. This is a Gaussian integral and may be done easily, giving

$$Z_r(\lambda)=\left(\frac{2\pi}{\beta J}\right)^{N_r/2}G_r(\lambda) \tag{3}$$

with

$$G_r(\lambda)=\prod_{i=1}^{N_r}(\lambda+\omega_i^2)^{-1/2} \tag{4}$$

where ω_i^2 are the normal mode frequencies of the model. The treatment of this model is quite parallel to that of the Ising model discussed in Ref. 1. In particular we may define the analog of star triangle and decoration transformations in the continuous spin case. The parameter λ acts like some external field on each spin. These transformations follow from the identities

$$\exp\left[-\frac{1}{2}\sum_{i,j=1}^{n}(x_i-x_j)^2\right]=n\int_{-\infty}^{+\infty}\frac{dy}{\sqrt{2\pi}}\times\exp\left[(-n/2)\sum_{i=1}^{n}(y-x_i)^2\right], \tag{5}$$

$$\int_{-\infty}^{+\infty}\frac{d\xi}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{d\eta}{\sqrt{2\pi}}\exp\left(-\frac{n}{2}(x_1-\xi)^2-\frac{\lambda}{2}\xi^2-\frac{1}{2}(\xi-\eta)^2-\frac{\lambda}{2}\eta^2-\frac{n}{2}(\eta-x_2)^2\right)$$

$$=(\lambda+n)^{-1/2}(\lambda+n+2)^{-1/2}\times\exp\left(-\frac{1}{2}\frac{n^2(x_1-x_2)^2}{(\lambda+n)(\lambda+n+2)}-\frac{1}{2}\frac{n\lambda(x_1^2+x_2^2)}{(\lambda+n)}\right). \tag{6}$$

These identities are applied as follows. We group the N_r spins into N_r/n groups of n spins each, such that each spin interacts with every other spin of the same group with interaction strength 1 in scaled units. Then we introduce a new spin y_j for each group j and couple it to each of the spins of its group with interaction strength n. The original $n(n-1)/2$ bonds within each group are deleted and replaced by n new bonds all connecting the new spins y_j to other spins (Fig. 5). Identity (5) guarantees that partition function remains unchanged.

The second identity corresponds to the dedecoration transformation of Fisher [Fig. 6]. We integrate over all the original spins χ_i. This leaves us with the new spins y_j which form an $(r-1)$th order truncated n-simplex lattice and all the spins y_j interact with their nearest neighbor spins on this lattice by quadratic interaction. By a scale transformation on the spin y_j this new interaction may be converted into an interaction of unit strength.

Applying these transformations we find that

$$G_r(\lambda)=[(\lambda+n)(\lambda+n+2)]^{-N_r/4+N_r/2n}G_{r-1}(\lambda^2+n\lambda+2\lambda). \tag{7}$$

Define

$$F(\lambda)=\lim_{r\to\infty}\ln G_r(\lambda)/N_r.$$

We get

$$F(\lambda)=\frac{n-2}{4n}\ln\frac{1}{(\lambda+n)(\lambda+n+2)}+\frac{1}{n}F(\lambda^2+n\lambda+2\lambda). \tag{8}$$

The frequency spectrum of the model may be determined from this functional equation for the characteristic function $F(\lambda)$. Details of the calculation are given in the Appendix. We note that $\lambda=0$ is a fixed point of the transformation; from Eq. (8), putting $\lambda=0$, we get

$$F(0)=\frac{n-2}{4(n-1)}\ln\frac{1}{n(n+2)}. \tag{9}$$

From Eq. (A1) of the Appendix, it is easily seen that for small $\lambda>0$, $F(\lambda)$ has an asymptotic expansion of the form $F(\lambda)=F(0)+A\lambda^{d/2}+$(higher order in λ), where d is the dimensionality of the lattice. Substituting in Eq. (8) we get

$$d=2\log_{n+2}n. \tag{10}$$

It is interesting to note that this result may be obtained without the complicated analysis of Eq. (8) given in the Appendix.

B. The modified rectangular lattice

For this lattice we wish to determine a functional equation of the same general form as Eq. (8) from which the dimensionality of the lattice may be deduced. The functional equation is however much more complicated and involves five variables instead of one.

We have

$$H=(J/2)\sum_{(ij)}(\chi_i-\chi_j)^2+(\lambda J/2)\sum_{i=1}^{N_r}\chi_i^2, \tag{1}$$

where the sum over (ij), extends over all the nearest neighbors. We consider a more general form of the Hamiltonian and allow different interaction strengths and more than nearest neighbor interaction between the spins belonging to the same first order rectangle. We write

$$H=\tfrac{1}{2}\sum_{(ij)}J_{ij}(\chi_i-\chi_j)^2+(\lambda J/2)\sum_i\chi_i^2, \tag{11}$$

where $J_{ij}=0$ unless i and j are nearest neighbor or they belong to the same first order rectangle. J_{ij} may take

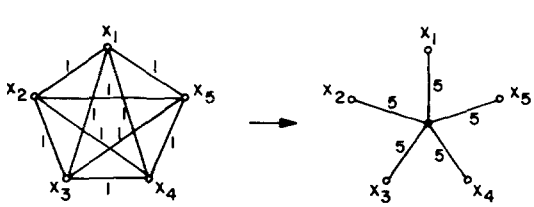


FIG. 5. An example of the generalized star triangle transformation with $n=5$.

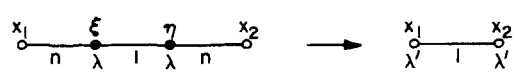


FIG. 6. The dedecoration transformation. Spins ξ and η are integrated over.

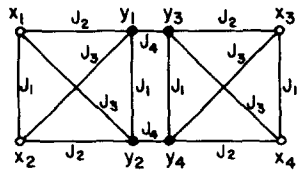


FIG. 7. Coupling constants for the modified rectangular lattice. The spins y_1, y_2, y_3, y_4 are integrated over.

four different values J_1, J_2, J_3, or J_4. These interactions are depicted in Fig. 7. J_1, J_2, and J_3 are respectively the vertical, horizontal, and diagonal bond strengths if i and j belong to the same first order rectangle. J_4 is the interaction strength if ij are nearest neighbors but do not belong to the same first order rectangle.

The renormalization transformation consists of integrating over the spins y_1, y_2, y_3, and y_4 and similar spins in other rectangles. This reduces the number of degrees of freedom in the Hamiltonian by a factor of 2. The integration over y_i $(i=1,4)$ gives us an effective interaction between x_i $(i=1$ to $4)$ and since the interaction is quadratic, it may again be expressed in the form (11) except for new values of the coupling constants J_1, J_2, J_3. The value of J_4 remains the same. The spins x_i $(i=1$ to $4)$ may be said to form a higher order spin block (second order). We note that an rth order rectangle is, by definition, a $(2r-1)$th order spin block. The most general form of the weight of the rth order spin block with corner spins x_1, x_2, y_1, y_2 and all other spins integrated over is

$$W^{(r)}(x_1,x_2;y_1,y_2) = C^{(r)}\exp\left[\begin{array}{l} -\frac{\beta J_1^{(r)}}{2}\{(x_1-x_2)^2+(y_1-y_2)^2\} - \frac{\beta J_2^{(r)}}{2}\{(x_1-y_1)^2+(x_2-y_2)^2\} \\ -\frac{\beta J_3^{(r)}}{2}\{(x_1-y_2)^2+(x_2-y_1)^2\} - \frac{\beta J\lambda^{(r)}}{2}\{x_1^2+x_2^2+y_1^2+y_2^2\} \end{array}\right]. \tag{12}$$

We also have the recursion relation (Fig. 7)

$$W^{(r+1)}(x_1,x_3;x_2,x_4) = \int_{-\infty}^{+\infty}\!\!\int\!\!\int\!\!\int\!\!\int dy_1\,dy_2\,dy_3\,dy_4\,W^{(r)}(x_1,x_2;y_1,y_2) \times W^{(r)}(y_3,y_4;x_3,x_4)\exp\left(-\frac{\beta J_4}{2}\{(y_1-y_3)^2+(y_2-y_4)^2\}\right). \tag{13}$$

The integrations in this equation are Gaussian and may be carried out easily. This gives us the equations for the new coupling constants

$$C^{(r+1)} = (C^{(r)})^2(2\pi)^2\beta^{-2}(\lambda^{(r)}J+J_2^{(r)}+J_3^{(r)})^{-1/2} \times(\lambda^{(r)}J+2J_1^{(r)}+J_2^{(r)}+J_3^{(r)})^{-1/2} \times(\lambda^{(r)}J+J_2^{(r)}+J_3^{(r)}+2J_4)^{-1/2} \times(\lambda^{(r)}J+2J_1^{(r)}+2J_4+J_2^{(r)}+J_3^{(r)})^{-1/2}, \tag{14a}$$

$$\lambda^{(r+1)} = \lambda^{(r)}(\lambda^{(r)}J+2J_2^{(r)}+2J_3^{(r)})(\lambda^{(r)}J+J_2^{(r)}+J_3^{(r)})^{-1}, \tag{14b}$$

$$\lambda^{(r+1)}J+2J_1^{(r+1)}+2J_3^{(r+1)} = \lambda^{(r)}J+J_2^{(r)}+J_3^{(r)}-(J_2^{(r)}+J_3^{(r)})^2(\lambda^{(r)}J+2J_4 +J_2^{(r)}+J_3^{(r)})^{-1}, \tag{14c}$$

$$\lambda^{(r+1)}J+2J_2^{(r+1)}+2J_3^{(r+1)} = (\lambda^{(r)}J+2J_1^{(r)}+2J_3^{(r)})(\lambda^{(r)}J+2J_1^{(r)}+2J_2^{(r)}) \times(\lambda^{(r)}J+2J_1^{(r)}+J_2^{(r)}+J_3^{(r)})^{-1}, \tag{14d}$$

$$\lambda^{(r+1)}J+2J_1^{(r+1)}+2J_2^{(r+1)} = [(\lambda^{(r)}J+2J_1^{(r)}+2J_3^{(r)})(\lambda^{(r)}J+2J_1^{(r)}+2J_2^{(r)}+2J_4 \times(\lambda^{(r)}J+2J_1^{(r)}+J_2^{(r)}+J_3^{(r)})] [\lambda^{(r)}J+2J_4+2J_1^{(r)}+J_2^{(r)}+J_3^{(r)}]^{-1}. \tag{14e}$$

If the free energy per spin of the Hamiltonian given by Eq. (11) is given by $F(J_1,J_2,J_3,J_4,\lambda)$ we see that the renormalization transformation gives us the equation

$$F(J_1,J_2,J_3,J_4,\lambda) = \frac{1}{8}\ln\frac{C^{(r+1)}}{(C^{(r)})^2}+\tfrac{1}{2}F(J_1',J_2',J_3',J_4',\lambda'), \tag{15}$$

where $J_1',J_2',J_3',J_4',\lambda'$ are given by Eqs. (14b)–(14e). To get the free energy per spin of the original modified rectangular model we put $J_1=J_2=J_4=J$, $J_3=0$. We note that $\lambda=0$ is a fixed point of the renormalization transformation. The free energy per spin is a singular function of λ near $\lambda=0$ and has an asymptotic expansion of the form $(\lambda>0)$.

$$F(J_1,J_2,J_3,J_4,\lambda) = F(J_1,J_2,J_3,J_4,0) + \lambda^{d/2}G(J_1,J_2,J_3,J_4) + \text{higher order terms in } \lambda. \tag{16}$$

We wish to determine the behavior of F for small λ. However, the value of λ increases with iteration approximately as $\lambda^{(r+1)}\approx 2\lambda^{(r)}$ as is evident from Eq. (14b). Also values of J_1, J_2, J_3 decrease with iteration. We choose initial λ sufficiently small and r sufficiently large so that

$$\lambda^{(r)}J \ll J_1^{(r)}, J_2^{(r)}, J_3^{(r)} \ll J_4. \tag{17}$$

Then the recursion equations (14) may be approximated as

$$\lambda^{(r+1)} \approx 2\lambda^{(r)}, \tag{18a}$$

$$J_1^{(r+1)}+J_3^{(r+1)} \approx \tfrac{1}{2}(J_2^{(r)}+J_3^{(r)}), \tag{18b}$$

$$J_2^{(r+1)}+J_3^{(r+1)} \approx 2(J_1^{(r)}+J_2^{(r)})(J_1^{(r)}+J_3^{(r)})(2J_1^{(r)}+J_2^{(r)}+J_3^{(r)})^{-1}, \tag{18c}$$

$$J_1^{(r+1)}+J_2^{(r+1)} \approx J_1^{(r)}+\tfrac{1}{2}J_2^{(r)}+\tfrac{1}{2}J_3^{(r)}. \tag{18d}$$

Now if we write $F^{\text{singular}}(J_1,J_2,J_3,J_4,\lambda)$ as the singular part of free energy near $\lambda=0$, we get from Eq. (15)

$$F^{\text{singular}}(J_1^{(r)}, J_2^{(r)}, J_3^{(r)}, J_4, \lambda^{(r)})$$

$$\approx \tfrac{1}{2} F^{\text{singular}}(J_1^{(r+1)}, J_2^{(r+1)}, J_3^{(r+1)}, J_4, \lambda^{(r+1)}). \tag{19}$$

It can be shown that Eqs. (18b)—(18d) imply that $J_2^{(r)}/J_1^{(r)}$ and $J_3^{(r)}/J_1^{(r)}$ tend to the limiting ratios $(2^{1/3}+2^{-1/3})$ and $(2^{-1/3})$ geometrically with r. Hence for large enough r the Eqs. (18a)—(18d) have the form

$$\lambda^{(r+1)} \approx 2\lambda^{(r)}, \tag{20a}$$

$$J_1^{(r+1)} \approx 2^{-1/3} J_1^{(r)}, \tag{20b}$$

$$J_2^{(r+1)} \approx 2^{-1/3} J_2^{(r)}, \tag{20c}$$

$$J_3^{(r+1)} \approx 2^{-1/3} J_3^{(r)}. \tag{20d}$$

Now since the variable $J_2^{(r)}$ and $J_3^{(r)}$ are asympotically proportional to $J_1^{(r)}$, these variables are unnecessary for our discussion. Any one of the variables $J_1^{(r)}, J_2^{(r)}, J_3^{(r)}$ is adequate for the discussion of critical properties of the model. These variables provide the relevant scale for our problem. We assume that $F^{\text{singular}}(J_1, J_2, J_3, J_4, \lambda) \approx A(\lambda J_4/J_1)^{d/2}$ where A is some constant. Substituting this in Eq. (19) and with the help of Eqs. (20a) and (20b) we find that $d = \frac{3}{2}$ which is the promised result.

IV. ANALYSIS OF THE CRITICAL BEHAVIOR OF SOME OTHER HAMILTONIANS

In the previous section the nearest neighbor harmonic interaction model (or equivalently the spherical model) on the truncated n-simplex lattice and the modified rectangular lattice was analyzed. Nelson and Fisher[1] have analyzed the behavior of the Ising model on the truncated tetrahedron lattice and determined the behavior of specific heat and susceptibility as a function of temperature. In the following we sketch the behavior of some other Hamiltonians on these lattices. Due to the special structure of these lattices, exact renormalization equations may be written down involving only a finite number of coupling constants. For definiteness we shall consider only the truncated 3-simplex lattice. Other lattices may be treated similarly and so long as $d<2$ the qualitative behavior of the lattices is quite similar and is characterized by the absence of phase transitions at any finite temperature. Lattices with $d>2$ are more interesting because they show phase transitions but the exact renormalization equations are much more complicated and are difficult to analyze. We confine ourselves to the nearest neighbor interactions for simplicity.

A. The classical XY model

In the classical XY model, a spin direction θ_i, $0 < \theta_i \leq 2\pi$, is assigned to each site i of the lattice. The Hamiltonian of the system is given by

$$H = -J \sum_{(ij)} \cos(\theta_i - \theta_j) - h \sum_i \cos\theta_i, \tag{21}$$

where the summation over ij extends over all pairs of nearest neighbor sites. The corresponding partition function is given by

$$Z = \prod_i \left(\int_0^{2\pi} \frac{d\theta_i}{2\pi} \right) \exp[-\beta H]. \tag{22}$$

First consider the case of zero magnetic field. We write

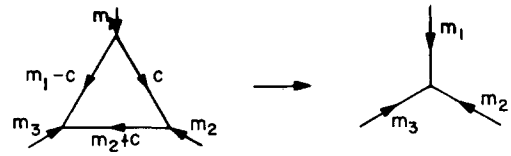


FIG. 8. The renormalization transformation for the XY model on the truncated 3-simplex lattice. The three vertices in the left figure are replaced by a single vertex.

$$\exp[\beta J \cos(\theta_i - \theta_j)] = \sum_{n_{ij}=-\infty}^{+\infty} I_{n_{ij}}(\beta J) \exp[i n_{ij}(\theta_i - \theta_j)], \tag{23}$$

where $I_m(x)$ is the modified Bessel function of x of order m and express Z as a summation over all possible values of (n_{ij}). With each term of this summation we associate an arrow configuration on the bonds of the lattice. The arrow configuration has n_{ij} arrows going from site i to site j for all nearest neighbor sites i and j with the convention $n_{ij} = -n_{ji}$. Thus the partition function is expressed as a sum over all the possible arrow configurations. The weight of a bond with r arrows is $I_r(\beta J)$. Integration over θ_i are trivial and show that only those configurations of arrows contribute to the partition sum in which the total number of arrows going into any site is zero.

Equivalently we may assign a weight $T(m_1, m_2, m_3)$ to a vertex with m_1, m_2, m_3 arrows going into it from its three nearest neighbors. If $m_1 + m_2 + m_3 \neq 0$ we set the corresponding vertex weight $T(m_1, m_2, m_3) = 0$.

The renormalization procedure for the truncated 3-simplex lattice consists of replacing a group of three vertices by a single vertex and is illustrated in Fig. 8. If $T^{(r)}(m_1, m_2, m_3)$ denotes the vertex weight after r iterations of the renormalization transformation, the renormalization equations are given by

$$T^{(r+1)}(m_1, m_2, m_3)$$

$$= \sum_{c=-\infty}^{+\infty} T^{(r)}(m_1, c - m_1, -c) T^{(r)}(m_2, m_1 - c, c + m_3)$$

$$\times T^{(r)}(m_3, c, -c - m_3). \tag{24}$$

The starting values of vertex weights are

$$T^{(0)}(m_1, m_2, -m_1 - m_2) = [I_{m_1}(\beta J) I_{m_2}(\beta J) I_{-m_1-m_2}(\beta J)]^{1/2}. \tag{25}$$

The free energy per site is given by

$$f(\beta) = -\frac{1}{\beta} \lim_{r\to\infty} \frac{1}{3^r} \ln T^{(r)}(0,0,0). \tag{26}$$

It is easy to verify that Eqs. (24) have only one attractive fixed point given by

$$T^{(r)}(m_1, m_2, m_3)/T^{(r)}(0,0,0) = \delta_{m_1,0}\delta_{m_2,0}\delta_{m_3,0}. \tag{27}$$

This corresponds to spin directions at large distances being uncorrelated and hence to the absence of spontaneous magnetization. If the temperature is very low, $\beta J \gg (m_1^2 + m_2^2 + m_3^3)$, we have

$$T^{(0)}(m_1, m_2, -m_1-m_2)$$

$$\approx A^{(0)} \exp[-\{m_1^2+m_2^2+(-m_1-m_2)^2\}/4\beta J]. \quad (28)$$

Then from Eq. (24) we get

$$T^{(1)}(m_1, m_2, -m_1-m_2)$$

$$\approx A^{(1)} \exp[-5\{m_1^2+m_2^2+(-m_1-m_2)^2\}/12\beta J], \quad (29)$$

where $A^{(0)}$ and $A^{(1)}$ are some constants. This shows that the renormalization equation for the temperature is

$$\beta' = \tfrac{3}{5}\beta. \quad (30)$$

Now let us introduce a small external field h, $\beta h \ll 1$. It is easy to see that the vertex weights $T(m_1, m_2, m_3)$ are no longer zero if $m_1+m_2+m_3 \neq 0$. To first order in βh, and for $\beta J \gg (m_1^2+m_2^2+m_3^2)$ the other nonzero vertex weights are

$$T^{(0)}(m_1, m_2, m_3) = \beta h A^{(0)} \exp[-\{m_1^2+m_2^2+m_3^2\}/4\beta J]$$

$$\text{for } m_1+m_2+m_3 = \pm 1. \quad (31)$$

After one renormalization transformation, these may be shown to transform to

$$T^{(1)}(m_1, m_2, m_3) = 3\beta h A^{(1)} \exp[-5\{m_1^2+m_2^2+m_3^2\}/12\beta J]$$

$$\text{for } m_1+m_2+m_3 = \pm 1. \quad (32)$$

Thus we see that to the lowest order in h, the external field is transformed according to the renormalization equation

$$\beta' h' = 3\beta h. \quad (33)$$

Using Eqs. (30) and (33), it is easy to determine the behavior of susceptibility as a function of temperature. Let $g(\beta, h)$ be the singular part of the logarithm of partition function per spin. By the renormalization transformation

$$g(\beta, h) = \tfrac{1}{3} g(\tfrac{3}{5}\beta, 5h), \quad (34)$$

and by the susceptibility

$$\chi(\beta) = -\frac{1}{\beta}\frac{\partial^2 g(\beta, h)}{\partial h^2}\bigg|_{h=0} = -\frac{5}{\beta'}\frac{\partial^2 g(\beta', h')}{\partial h'^2}\bigg|_{h'=0}$$

$$= 5\chi(\tfrac{3}{5}\beta). \quad (35)$$

Hence if $\chi(\beta) \sim \beta^{-\bar{\gamma}}$ we get

$$\bar{\gamma} = \ln 5/(\ln 5 - \ln 3). \quad (36)$$

Similar analysis of the truncated n-simplex lattice for arbitrary n shows that in general, for $d < 2$

$$\bar{\gamma} = 2/(2-d). \quad (37)$$

The specific heat tends to simple harmonic oscillator value $\frac{1}{2}K_B$ as the temperature tends to zero.

The exponent $\bar{\gamma}$ in Eq. (37) should not be identified with the critical exponent γ which appears in the scaling relations and specifies the divergence of susceptibility as a function of temperature slightly away from the critical point. The reason is the arbitrariness in the definition of critical exponents when the critical temperature is zero. The procedure of determining critical exponents in terms of the divergence of correlation length[1] is of doubtful validity here because the lattice is not translationally invariant and the "correlation length" is a function of position. Different definitions of the "averaged correlation length" may well give rise to different dependence on temperature. In general, it is difficult to find a useful definition of the correlation length for the spatially inhomogeneous and highly anisotropic lattices studied here. Consequently the critical exponents ν and η are undefined even if the transition temperature is finite. (To define them in terms of scaling relations would be begging the question.) The critical exponents like $\alpha, \beta, \gamma, \delta$, etc., which may be defined by differentiation of thermodynamic quantities are however unambiguously defined for nonzero transition temperatures.

B. The Fortuin-Kasteleyn cluster model

The cluster model was introduced by Fortuin and Kasteleyn.[5-7] It is defined in terms of a parameter κ and includes as special cases the percolation model ($\kappa = 1$), the Ising model ($\kappa = 2$), resistive networks ($\kappa = 0$), and the n-state Potts model ($\kappa = n$), etc.

The partition function of the cluster model for any given graph is given by

$$Z(p, \kappa) = \sum_C p^{r(C)}(1-p)^{N-r(C)}\kappa^{\gamma(C)}, \quad (38)$$

where N is the total number of bonds in the graph. The summation extends over all the possible configurations C of "occupied" or "unoccupied" states of bonds on the lattice. $r(C)$ is the number of occupied bonds in the configuration C and $\gamma(C)$ is the number of distinct clusters in the configuration C. Fortuin and Kastelyn have called this the random cluster model. However the model is completely specified by a Hamiltonian which has a term proportional to the number of occupied bonds and a term proportional to the total number of clusters. It is thus no more random than any other statistical mechanical system and the adjective "random" is unnecessary.

Due to its very general nature, the cluster model is of great interest in statistical physics. However, for arbitrary values of κ the model has not been solved even in two dimensions. We sketch below the renormalization group treatment of this model for the truncated tetrahedron lattice. It is shown that the model does not show any phase transition for any values of κ and p such that $\kappa > 0$, $0 \leq p \leq 1$.

We define the restricted partition functions $Z^{(r)}(\alpha|\beta|\gamma)$, $Z^{(r)}(\alpha\beta|\gamma)$, $Z^{(r)}(\beta\gamma|\alpha)$, $Z^{(r)}(\alpha\gamma|\beta)$, and $Z^{(r)}(\alpha\beta\gamma)$. Here $Z^{(r)}(\alpha|\beta|\gamma)$ is the partition function of the rth order triangle whose corner vertices are α, β, and γ and the summation corresponds to all configurations of edges within the triangle subject to the constraint that no sequence of occupied bonds within the rth order triangle connects any two of the vertices α, β, γ. $Z^{(r)}(\alpha\beta\gamma)$ is defined similarly except that in this case α, β, and γ are connected together by bonds lying within the rth order triangle. By symmetry between the corner vertices α, β, and γ we have

$$Z^{(r)}(\alpha\beta|\gamma) = Z^{(r)}(\beta\gamma|\alpha) = Z^{(r)}(\alpha\gamma|\beta) = Z_2^{(r)} \quad \text{(say)} \quad (39a)$$

and we write

$$Z^{(r)}(\alpha|\beta|\gamma) = Z_1^{(r)}, \tag{39b}$$

$$Z^{(r)}(\alpha\beta\gamma) = Z_3^{(r)}. \tag{39c}$$

It is quite straightforward though tedious to write down the expressions for $Z_1^{(r+1)}$, $Z_2^{(r+1)}$, $Z_3^{(r+1)}$ in terms of $Z_1^{(r)}$, $Z_2^{(r)}$, $Z_3^{(r)}$ by summing over the 2^3 possible states of the three bonds that connect the rth order triangles to form the $(r+1)$th order triangle and grouping the terms that correspond to the same connectivity structure of the $(r+1)$th order triangle together. The result is

$$\begin{aligned} Z_1^{(r+1)} &= q^3(Z_1 + 3Z_2 + Z_3)^3 + 3pq^2\kappa^{-1}(Z_1 + 3Z_2 + Z_3) \\ &\times (Z_1 + 2Z_2)(Z_1 + 4Z_2 + 2Z_3) + 3p^2q\kappa^{-2} \\ &\times \{(Z_1 + 2Z_2)^3 + 3(Z_1 + 2Z_2)^2(Z_2 + Z_3) + Z_1(Z_2 + Z_3)^2\} \\ &+ p^3\kappa^{-3}\{\kappa(3Z_3Z_2^2 + Z_2^3) + 14Z_2^3 + 12Z_1Z_2Z_3 \\ &+ 24Z_2^2Z_1 + 3Z_1^2Z_3 + 9Z_1^2Z_2 + Z_1^3\}, \end{aligned} \tag{40a}$$

$$\begin{aligned} Z_2^{(r+1)} &= q^2p\kappa^{-1}(Z_1 + 3Z_2 + Z_3)(Z_2 + Z_3)^2 + p^2q\kappa^{-2}(Z_2 + Z_3)^2 \\ &\times (2Z_1 + 7Z_2) + p^3\kappa^{-3}\{\kappa Z_3^2Z_2 + 7Z_2^2Z_3 + Z_3^2Z_1 \\ &+ 2Z_1Z_2Z_3 + 4Z_2^3 + Z_1^2Z_3\}, \end{aligned} \tag{40b}$$

$$Z_3^{(r+1)} = 3p^2q\kappa^{-2}(Z_2 + Z_3)^2Z_3 + p^3\kappa^{-3}(\kappa Z_3^3 + 6Z_2Z_3^2 + 3Z_2^2Z_3), \tag{40c}$$

where we have put $q = 1 - p$ and $Z_i^{(r)} = Z_i$, $i = 1, 2, 3$. If $q > 0$, analysis of these equations shows that there exists only one fixed point which is attractive and corresponds to $Z_2/Z_1 = Z_3/Z_1 = 0$. This implies absence of phase transitions at any finite temperature. For small q and for small r, $Z_3^{(r)}$ is much larger than $Z_2^{(r)}$ or $Z_1(r)$. If we renormalize the Z_i after each iteration so that $Z_1^{(r)} + 3Z_2^{(r)} + Z_3^{(r)} = 1$ we get to the lowest order in q

$$Z_1^{(r)} = \kappa^2q^3, \tag{41a}$$

$$Z_2^{(r)} = \kappa q^2 r, \tag{41b}$$

$$Z_3^{(r)} = 1 - 3q^2\kappa r. \tag{41c}$$

From this we may deduce that for small q the mean size of one cluster (which corresponds to the susceptibility for the Ising model) is approximately given by $3^{a/\kappa q^2}$ where a is some absolute constant. For the Ising model, this result has already been obtained by Nelson and Fisher. We need only note the correspondence $q = e^{-\beta J}/(e^{\beta J} + e^{-\beta J})$ and $\kappa = 2$ to specialize our result to the Ising model of interaction strength J.

To summarize, in this paper we have introduced a class of lattices of effectively nonintegral dimensionality. The lattices are not translationally invariant but sufficiently regular so that the existence of thermodynamic limit is assured for most physically interesting Hamiltonians. The procedure of determining critical exponents as a function of dimensionality was illustrated in Sec. IV where the classical XY model was analyzed for $d < 2$. For lattices with $d > 2$, exact analysis of thermodynamic properties is difficult; but approximate renormalization group equations or numerical extrapolation methods like series expansions may be used to determine the value of critical exponents. This seems to be a promising field for further investigation.

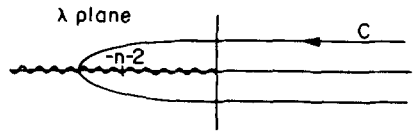


FIG. 9. The complex λ plane showing the curving C and the branch cut along the negative real axis.

APPENDIX

We have the equation

$$F(\lambda) = \frac{n-2}{4n}\ln\{(\lambda + n)(\lambda + n + 2)\}^{-1} + \frac{1}{n}F(n\lambda + 2\lambda + \lambda^2). \tag{8}$$

This equation is valid for all real $\lambda > 0$. As λ tends to infinity $F(\lambda)$ tends to $-\frac{1}{2}\ln\lambda$. This condition specifies $F(\lambda)$ completely for all real positive λ when combined with Eq. (8). From Eq. (4) we may write

$$F(\lambda) = -\tfrac{1}{2}\int_0^{\omega_{\max}^2} d\omega^2\, D(\omega^2)\ln(\omega^2 + \lambda), \tag{A1}$$

where $D(\omega^2)\,d\omega^2$ is the fractional number of modes in the frequency range ω^2 and $\omega^2 + d\omega^2$. From Eq. (A1) we may define $F(\lambda)$ as an analytic function in λ over the entire complex λ plane. The function has logarithmic branch points on the negative real axis and consists of many sheets. Consider the sheet in which $F(\lambda)$ is real for all real positive λ. We make the branch cut along the negative real line. Then it is easy to see that

$$H(\omega^2) \overset{\text{def}}{=} -\frac{2}{\pi}\,\mathrm{Im}F(-\omega^2 + i\epsilon), \quad \text{for all } \omega^2, \tag{A2}$$

$$= \int_0^{\omega^2} d\omega'^2\, D(\omega'^2), \qquad \text{for } \omega^2 > 0. \tag{A3}$$

Consider $\lambda = \lambda_R + i\epsilon$ where λ_R is real. As we vary λ_R from $+\infty$ to $-\infty$, the function $(\lambda^2 + n\lambda + 2\lambda)$ traces out the curve C shown in Fig. 9. For $\lambda_R < -(n+2)/2$ the curve C crosses over the negative real axis and goes into a different sheet. The value of the function $F(\lambda)$ on this sheet differs from its value on the original sheet by an additive imaginary constant. Taking the imaginary part of the Eq. (8) it is easy to show that

$$0 < \omega^2 < \frac{n+2}{2}, \quad H(\omega^2) = \frac{1}{n}H(n\omega^2 + 2\omega^2 - \omega^4), \tag{A4}$$

$$\frac{n+2}{2} < \omega^2 < \infty, \quad H(\omega^2) = \frac{n-2}{2n}\{E(\omega^2 - n) + E(\omega^2 - n - 2)\} + \frac{2}{n} - \frac{1}{n}H(n\omega^2 + 2\omega^2 - \omega^4), \tag{A5}$$

where $E(x)$ is the unit step function which is zero for negative arguments and $+1$ for positive arguments. From Eq. (A3) it is easy to see that $H(\omega^2)$ is a nonnegative monotonically increasing function of ω^2, and we have

$$H(\omega^2) = 0, \quad \text{for } -\infty < \omega^2 \leq 0, \tag{A6}$$

$$H(\omega^2) = 1, \quad \text{for } \infty > \omega^2 > n + 2. \tag{A7}$$

We shall determine the function $H(\omega^2)$ in the unknown range $0 < \omega^2 < (n+2)$ by repeated application of Eqs. (A4) and (A5).

Consider $\omega_1^2 < \omega^2 < n$ where $(n+2)\omega_1^2 - \omega_1^4 = (n+2)$. Then from Eq. (A7) we get $H(\omega^2(n+2) - \omega^4) = 1$ and hence by using Eqs. (A4) and (A5) we get

$$H(\omega^2) = \frac{1}{n} \quad \text{for } \omega_1^2 < \omega^2 < n. \tag{A8}$$

Consider again $n < \omega^2 < \omega_2^2$ where $(n+2)\omega_2^2 - \omega_2^4 = n+2$. Then again $H(n\omega^2 + 2\omega^2 - \omega^4)$ is equal to 1 by Eq. (A7), and using Eq. (A5) we get

$$H(\omega^2) = \tfrac{1}{2}, \quad \text{for } n < \omega^2 < \omega_2^2. \tag{A9}$$

Thus we have determined the function $H(\omega^2)$ in the interval $\omega_1^2 < \omega^2 < \omega_2^2$. This can be used to determine the function $H(\omega^2)$ in the intervals $\omega_3^2 < \omega^2 < \omega_4^2$ and $\omega_5^2 < \omega^2 < \omega_6^2$ where ω_3^2, ω_4^2, ω_5^2, and ω_6^2 are constants determined by the equations

$$(n+2)\omega_3^2 - \omega_3^4 = (n+2)\omega_6^2 - \omega_6^4 = \omega_1^2, \tag{A10}$$

$$(n+2)\omega_4^2 - \omega_4^4 = (n+2)\omega_5^2 - \omega_5^4 = \omega_2^2, \tag{A11}$$

and so on. Eventually $H(\omega^2)$ is determined in all the interval $0 < \omega^2 < n+2$ except for some small set of zero measure. Thus we see that $H(\omega^2)$ is constant everywhere except for an infinite but denumerable number of points where its value increases discontinuously. Thus $D(\omega^2)$ which is the derivative of $H(\omega^2)$ and is the spectral density of the system is a sum of an infinite number of delta functions. Also the points $\omega^2 = 0$ and $\omega^2 = n+2$ are clearly the points of accumulation of the delta functions and hence again by Eqs. (A4) and (A5) there are an infinite number of such points of accumulation. Clearly the spectral density $D(\omega^2)$ is a highly singular function of ω^2. However in thermodynamics we are usually interested only in integrals of $D(\omega^2)$ multiplied by a sufficiently smooth function of ω^2. Hence only some sort of smeared value of $D(\omega^2)$ is of interest. It is easy to prove that in the case there exists nonzero finite positive constants A and B such that

$$H(\omega^2) \leq A\omega^d, \quad \text{for } n+2 \geq \omega^2 \geq 0, \tag{A12}$$

$$H(\omega^2) \geq B\omega^d, \tag{A13}$$

where

$$d = 2\ln(n)/\ln(n+2). \tag{A14}$$

This result is sufficiently strong to permit us to identify d as the effective dimensionality of the lattice.

ACKNOWLEDGMENTS

I am deeply indebted to my advisor, Professor Jon Mathews, for his encouragement, discussions, and critical comments throughout the course of this investigation. The financial support of the H. Dudley Wright Foundation is gratefully acknowledged.

*Richard P. Feynman Fellow.

[1]D.R. Nelson and M.E. Fisher, Ann. Phys. (N.Y.) 91, 226 (1975).
[2]K. Wilson and J. Kogut, Phys. Rep. C 12, 75 (1974).
[3]M.E. Fisher, Rep. Prog. Phys. 30, 615 (1967).
[4]K. Wilson, Phys. Rev. D 7, 2924 (1973).
[5]C.M. Fortuin and P.W. Kastelyn, Physica 57, 536 (1972).
[6]C.M. Fortuin, Physica 58, 393 (1972).
[7]C.M. Fortuin, Physica 59, 545 (1972).

Stationary gravitational fields of a dually charged perfect fluid

A. Das and S. Kloster

Department of Mathematics, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada
(Received 25 October 1976)

Stationary field equations in the presence of a charged perfect fluid with both electric and monopole currents in isometric motion are studied. It is shown that the eight-parameter group of transformations which preserve the stationary electrovac equations can also be applied to dually charged sources. In the case of dually charged dust an equilibrium condition $\rho = (\sigma^*\sigma)^{1/2}$ implies a functional relationship between $\mathrm{Re}\Gamma$ and the complex potentials Φ and Φ^*. Furthermore, it is proved that when $\rho \neq 0$ and $\sigma \neq 0$, the additional assumption of an arbitrary linear relationship between Γ and Φ leads uniquely to the Israel–Spanos class of solutions.

1. INTRODUCTION

We investigate the case of a perfect fluid which is allowed to have both electric and monopole currents (dually charged). It is found that Bonnor's[1] transformation, which switches the electric and magnetic fields, can easily be extended to this case. Next the assumption of isometric motion is made, which is very useful because it allows the electromagnetic field tensor to be reduced to just two potentials (even when monopoles are present and the usual 4-potential does not exist). Two complex potentials Γ and Φ are introduced and the problem is reduced to three dimensions. In a beautiful paper, Kramer, Neugebauer, and Stephani[2] (KNS) have discovered an eight-parameter group of transformations which preserve the stationary electrovac equations. This group was found independently by Kinnersley.[3] In the present paper it is shown that the entire eight-parameter group of transformations can be applied to dually charged sources. (In the absence of a monopole current however, only a four-parameter subgroup can act.) One wonders if the classical models of elementary particles should be dually charged sources in a stationary universe where space and time reflections are violated[4] but an eight-parameter symmetry group is relevant.

In Sec. 4 the pressure term is set equal to zero and thus the case of dually charged dust is obtained. It is shown that the equation of motion implies that there must exist a functional relationship between g_{44} and the complex potentials Φ and Φ^*. This is a generalization of the result of De and Raychaudhuri[5] in the static case. Das[6] has shown that the equilibrium condition $\sigma = \pm\rho$ in the static case leads to the Weyl–Majumdar[7]–Papapetrou[8] (WMP) condition $g_{44} = (C + A_4)^2$. In the present setting of dually charged dust in isometric motion, the WMP condition has been generalized in Theorem 1 under the equilibrium condition $(\sigma^*\sigma)^{1/2} = \rho$.

Ernst[9] has shown in the axially symmetric stationary electrovac case that the only possible analytic functional relationship between the complex potentials Γ and Φ is a linear one. This result can be generalized by dropping the assumption of axial symmetry. By placing a restriction on the constants of the linear relationship, Perjes[10] and also Israel and Wilson[11] (IW) independently succeeded in reducing the field equations to one complex harmonic equation in Euclidean space (and thus generalized the Majumdar–Papapetrou class of static solutions). Israel and Spanos[12] (IS) have subsequently obtained a special class of solutions of stationary fields of dually charged dust in which the restricted linear relation is still maintained (and thus generalized the class of static solutions in Ref. 6). In Theorem 2 we have proved that, for a dually charged dust in isometric motion with $\rho \neq 0$ and $\sigma \neq 0$, the assumption of a linear relationship between Γ and Φ leads uniquely to the IS solution. Moreover, Corollary 3 shows that if the ratio of electric charge to monopole charge is constant, then these solutions must be static.

2. NOTATIONS AND FIELD EQUATIONS

The Einstein–Maxwell equations for a dually charged perfect fluid may be written as follows:

$$\begin{aligned} &Q_{ij} \equiv G_{ij} + \kappa[(\rho + p)u_i u_j - p\gamma_{ij} + \tfrac{1}{2}(F_{ik}F^k{}_j + F^*_{ik}F^{*k}{}_j)] \\ &\quad = 0, \\ &M^i \equiv (F^{ij} + iF^{*ij})_{;j} - \sqrt{4\pi}\sigma u^i = 0, \\ &n \equiv u^i u_i = 0, \end{aligned} \tag{2.1}$$

where $\kappa = 8\pi$ and the metric form $\gamma_{ij}dx^i dx^j$ has signature -2. F^{ij} is the electromagnetic field tensor, $\mathrm{Re}\sigma$ is the electric charge density, $\mathrm{Im}\sigma$ is the monopole charge density, ρ is the matter density, p is the pressure, and u^i is the 4-velocity.[13]

The number of unknown functions in (2.1) is 24 $= 10(g_{ij}) + 6(F_{ij}) + 4(u_i) + 2(\sigma) + 1(\rho) + 1(p)$. The number of equations is $23 = 10(Q_{ij}) + 8(M^i) + 1(n) + 4$ (coordinate conditions), leaving one degree of freedom.

Bonnor[1] has shown that electrovac fields may be transformed into magnetovac fields. This can easily be generalized to the fluid case, for we observe that the transformation $F^{ij} \leftrightarrow F^{*ij}$, $\mathrm{Re}\sigma \leftrightarrow \mathrm{Im}\sigma$ preserves (2.1).

3. ISOMETRIC MOTION

We now assume that the motion of the fluid is isometric, which means there is a timelike Killing vector field ξ^i which is parallel to the velocity u^i, and we assume that F_{ij} is invariant under this Killing motion. The coordinate system can be chosen so that $\xi^i = \delta^i{}_4$.

Copyright © 1977 American Institute of Physics

It can be shown[14] that there exist potential functions A and B such that $F_{4i} = -A_{,i}$ and $F^*_{4i} = B_{,i}$, and we can define $\Phi = A + iB$.

The metric can be put in the following form:

$$\gamma_{ij}\,dx^i\,dx^j = -e^{-\omega}(g_{\alpha\beta}dx^\alpha dx^\beta) + e^{\omega}(a_\alpha dx^\alpha + dx^4)^2, \quad (3.1)$$

where the Latin indices range from 1 to 4, the Greek indices range from 1 to 3, and the associated space V_3 with metric $g_{\alpha\beta}$ is positive definite. In this V_3 we define a vector $\boldsymbol{\tau} \equiv -e^{2\omega}\,\mathrm{curl}(\mathbf{a})$. The space–time will be static iff $\boldsymbol{\tau} = 0$. Since the fluid terms do not contribute to T^α_4, there exists[15] a "twist potential" χ such that $\chi_{,\alpha} = T_\alpha + \frac{1}{2}i\kappa(\Phi^*\Phi_{,\alpha} - \Phi\Phi^*_{,\alpha})$. We can now define[16] the complex function $\Gamma = e^{\omega} - \frac{1}{2}\kappa\Phi^*\Phi + i\chi$. The field equations (2.1) can be written as follows[17]:

$$\begin{aligned}\sigma_{\alpha\beta} \equiv{}& R_{\alpha\beta} - \kappa e^{-\omega}\,\mathrm{Re}(\Phi_{,\alpha}\Phi^*_{,\beta} + 2pg_{\alpha\beta}) \\ &+ \tfrac{1}{2}e^{-2\omega}\,\mathrm{Re}(\Gamma^*_{,\alpha} + \kappa\Phi\Phi^*_{,\alpha})(\Gamma_{,\beta} + \kappa\Phi^*\Phi_{,\beta}) \\ &= 0, \end{aligned} \quad (3.2)$$

$$\nu \equiv \Delta_2\Phi - e^{-\omega}\Phi_{,\alpha}(\Gamma^{,\alpha} + \kappa\Phi^*\Phi^{,\alpha}) - e^{-\omega/2}\sqrt{4\pi}\sigma^* = 0, \quad (3.3)$$

$$\begin{aligned}\mu \equiv{}& \Delta_2\Gamma - e^{-\omega}\Gamma_{,\alpha}(\Gamma^{,\alpha} + \kappa\Phi^*\Phi^{,\alpha}) \\ &+ \kappa(-3p - \rho + e^{-\omega/2}\sqrt{4\pi}\sigma^*\Phi^*) \\ &= 0, \end{aligned} \quad (3.4)$$

where $\Delta_2\Phi \equiv g^{-1/2}(g^{1/2}g^{\alpha\beta}\Phi_{,\alpha})_{,\beta}$, $g = \det g_{\alpha\beta}$, and indices are raised with the 3-metric $g^{\alpha\beta}$. The equation of motion is

$$(\rho + p)\omega_{,\mu} = \sqrt{4\pi}e^{-\omega/2}(\sigma\Phi_{,\mu} + \sigma^*\Phi^*_{,\mu}) + 2p_{,\mu}. \quad (3.5)$$

The field equations can be put into an invariant form. Triad labels are denoted by capital Roman indices which take the values 1, 2, 3 and the summation convention on these indices is adopted. An orthonormal triad field $\lambda^\alpha{}_A(x)$ in V_3 satisfies $\lambda^\alpha{}_A\lambda_{\beta\alpha} = \delta_{AB}$. Let γ_{ABC} denote the Ricci rotation coefficients in V_3, and define

$$\begin{aligned} &T_{AB\ldots} \equiv \lambda_{A\alpha}\lambda_{B\beta}\ldots T^{\alpha\beta\cdots}, \\ &T_{\ldots,A} \equiv \lambda^\alpha{}_A T_{\ldots,\alpha}, \\ &T_{AB[CD]} \equiv \tfrac{1}{2}(T_{ABCD} - T_{ABDC}). \end{aligned} \quad (3.6)$$

Then the invariant form of equations (3.2)–(3.5) is the following:

$$\begin{aligned}&2[\gamma_{MN[L,R]} + \gamma_{MNA}\gamma_{A[LR]} + \gamma_{AM[R}\gamma_{|AN|L]} \\ &\quad + 2\kappa e^{-\omega}\delta_{M[L}\,\mathrm{Re}(\Phi_{,R]}\Phi^*_{,N} + 2p\delta_{R]N}) \\ &\quad - e^{-2\omega}\delta_{M[L}\,\mathrm{Re}(\Gamma^*_{,R]} + \kappa\Phi\Phi^*_{,R]})(\Gamma_{,N} + \kappa\Phi^*\Phi_{,N}) \\ &\quad + 2\kappa e^{-\omega}\delta_{N[R}\,\mathrm{Re}(\Phi_{,L]}\Phi^*_{,M} + 2p\delta_{L]M}) \\ &\quad - e^{-2\omega}\delta_{N[R}\,\mathrm{Re}(\Gamma^*_{,L]} + \kappa\Phi\Phi^*_{,L]})(\Gamma_{,M} + \kappa\Phi^*\Phi_{,M}) \\ &\quad + [\kappa e^{-\omega}(\Phi_{,A}\Phi^*_{,A} + 6p) - \tfrac{1}{2}e^{-2\omega}(\Gamma^*_{,A} + \kappa\Phi\Phi^*_{,A}) \\ &\quad \times(\Gamma_{,A} + \kappa\Phi^*\Phi_{,A})]\delta_{M[R}\delta_{L]N} = 0. \end{aligned} \quad (3.2')$$

$$\begin{aligned}\nu \equiv{}& \Phi_{,AA} + \gamma_{CAA}\Phi_{,C} - e^{-\omega}\Phi_{,A}(\Gamma_{,A} + \kappa\Phi^*\Phi_{,A}) \\ &- e^{-\omega/2}\sqrt{4\pi}\sigma^* = 0, \end{aligned} \quad (3.3')$$

$$\begin{aligned}\mu \equiv{}& \Gamma_{,AA} + \gamma_{CAA}\Gamma_{,C} - e^{-\omega}\Gamma_{,A}(\Gamma_{,A} + \kappa\Phi^*\Phi_{,A}) \\ &+ \kappa(-3p - \rho + e^{-\omega/2}\sqrt{4\pi}\sigma^*\Phi^*) = 0, \end{aligned} \quad (3.4')$$

$$(\rho + p)\omega_{,A} = \sqrt{4\pi}e^{-\omega/2}(\sigma\Phi_{,A} + \sigma^*\Phi^*_{,A}) + 2p_{,A}. \quad (3.5')$$

TABLE I.

	duality	gravitational gauge	electromagnetic gauge
Γ'	$\lvert\alpha\rvert^2\Gamma$	$\Gamma + i\gamma$	$\Gamma - 2\kappa\delta\Phi - 2\kappa\lvert\delta\rvert^2$
Φ'	$\alpha\Phi$	Φ	$\Phi + 2\delta^*$
p'	$\lvert\alpha\rvert^2 p$	p	p
ρ'	$\lvert\alpha\rvert^2\rho$	ρ	ρ
σ'	$\alpha^*\lvert\alpha\rvert\sigma$	σ	σ

The number of unknowns in Eqs. (3.2)–(3.4) is $14 = 6(g_{\alpha\beta}) + 2(\Gamma) + 2(\Phi) + 2(\sigma) + 1(\rho) + 1(p)$, while the number of equations is $13 = 6(\sigma_{\alpha\beta}) + 2(\nu) + 2(\mu) + 3$(coordinate conditions).

There is an eight-parameter group of transformations which preserve the field equations (3.2)–(3.4) outside of sources. This group can be generated by five basic transformations.[18] It turns out that all[19] of these also preserve the equations with sources, where the effects on the sources are described below. Table I treats the first, second, and fourth basic transformations of KNS. Both of the remaining basic transformations can be generated by the discrete transformation[20] T defined by $T(\Gamma, \Phi) = (\Gamma^{-1}, \Phi\Gamma^{-1})$. This preserves the field equations outside of sources and is actually a proper element of the group. Defining $T_\gamma(\Gamma, \Phi) = (\Gamma + i\gamma, \Phi)$ (the second basic transformation), we find that the third transformation is just $T \circ T_\gamma \circ T$. The fifth transformation is generated in a similar way from the fourth and T. The conjugation transformation K defined by $K(\Gamma, \Phi) = (\Gamma^*, \Phi^*)$ also preserves the stationary electrovac equations.

T and K can be extended to sources as follows:

$$\begin{aligned} &T(p) = (\Gamma^*\Gamma)^{-1}p, \\ &T(\sigma^*) = \lvert\Gamma\rvert^{-1}\Gamma^{-2}(\Gamma + \kappa\lvert\Phi\rvert^2)\sigma^* \\ &\qquad - 2\sqrt{4\pi}e^{\omega/2}\lvert\Gamma\rvert^{-1}\Gamma^{-2}\Phi(3p + \rho), \\ &T(\rho) = (\Gamma^*\Gamma)^{-1}(-3p + e^{-\omega/2}\sqrt{4\pi}\sigma^*\Phi^*) \\ &\qquad + (\kappa\lvert\Phi\rvert^2\Gamma^{*-1}\Gamma^{-2} + \Gamma^2)(-3p - \rho + e^{-\omega/2}\sqrt{4\pi}\sigma^*\Phi^*), \\ &K(p) = p, \quad K(\sigma) = \sigma^*, \quad K(\rho) = \rho. \end{aligned} \quad (3.7)$$

In terms of the potentials Bonnor's transformation is $\Gamma' = \Gamma^*$, $\Phi' = -i\Phi^*$, $\sigma' = i\sigma^*$, so that this is just K followed by the α transformation with $\alpha = -i$. Thus we see that the eight-parameter group as well as T and K takes a dually charged perfect fluid solution to another such solution, and, naturally, dust to dust.

4. EXACT SOLUTIONS FOR DUALLY CHARGED DUST

In this section isometric motion is assumed throughout. In the case of dually charged dust ($p = 0$), the equation of motion (3.5) becomes

$$\rho\omega_{,\mu} = \sqrt{4\pi}e^{-\omega/2}(\sigma\Phi_{,\mu} + \sigma^*\Phi^*_{,\mu}). \quad (4.1)$$

If either ρ or σ is nonzero, then (4.1) implies that the vectors $\omega_{,\mu}, \Phi_{,\mu}, \Phi^*_{,\mu}$ are linearly dependent, so that the Jacobian $\partial(\omega, \Phi, \Phi^*)/\partial(x^1, x^2, x^3)$ is zero. Thus there must be a functional relationship (constraint) of the

form

$$G(\omega, \Phi, \Phi^*) = 0. \tag{4.2}$$

If no other constraints exist, then by comparing (4.1) and (4.2) we obtain

$$\rho = \lambda G_{,\omega}, \quad \sqrt{4\pi} e^{-\omega/2}\sigma = -\lambda G_{,\Phi},$$
$$\sqrt{4\pi} e^{-\omega/2}\sigma^* = -\lambda G_{,\Phi^*}, \tag{4.3}$$

where λ is a Lagrange multiplier. The condition

$$\rho = |\sigma| \equiv (\sigma^*\sigma)^{1/2} \tag{4.4}$$

can be regarded as defining an equilibrium.

Theorem 1: If $\rho \neq 0$ and A and B are functionally independent, then $\rho = |\sigma|$ iff

$$\pi^{-1/2} e^{\omega/2} = e^{i\theta}(\Phi + \gamma) + e^{-i\theta}(\Phi^* + \gamma^*) + V(e^{i\theta}), \tag{4.5}$$

where γ is a complex constant, V is an arbitrary function, and θ is a real function of Φ and Φ^* defined implicitly by

$$0 = (\Phi + \gamma) - e^{-2i\theta}(\Phi^* + \gamma^*) + V'(e^{i\theta}). \tag{4.6}$$

Proof: By using (4.3) and $\rho \neq 0$, the implicit function theorem shows that ω is a function of Φ and Φ^*. Thus $\omega_{,\mu} = \omega_{,\Phi}\Phi_{,\mu} + \omega_{,\Phi^*}\Phi^*_{,\mu}$. By (4.1) we find

$$\rho\omega_{,\Phi} = \sqrt{4\pi} e^{-\omega/2}\sigma. \tag{4.7}$$

Thus (4.4) becomes

$$(\pi^{-1/2} e^{\omega/2})_{,\Phi}(\pi^{-1/2} e^{\omega/2})_{,\Phi^*} = 1. \tag{4.8}$$

This is formally the same as Eq. (12) on p. 36 of Courant and Hilbert,[21] all solutions of which are developable surfaces. They give the general solution in Eqs. (9) and (10) on p. 10, and these reduce to (4.5) and (4.6) in our case. Q.E.D.

For axially symmetric electrovac fields, Ernst[9] has shown that the only possible analytic relationship between Γ and Φ is a linear one. This result holds even if the symmetry assumption is removed. The next theorem tells what happens if this linear relationship is continued inside matter.

Theorem 2: If $\Gamma = \gamma\Phi + \delta$ in a dually charged dust, where γ, δ are complex constants, $\rho \neq 0$, and A and B are functionally independent, then

(i) $e^{\omega} = |4^{-1}\pi^{-1/2}\gamma^* + \sqrt{4\pi}\Phi|^2$,

(ii) $\rho = |\sigma|$,

(iii) $\mathrm{Re}\delta = (16\pi)^{-1}|\gamma|^2$,

(iv) the associated V_3 is flat,

(v) $\sigma^* = -2\kappa^{-1}U^{-2}|U|^{-1}\nabla^2 U$,

where $U = \sqrt{16\pi}(\gamma^* + \kappa\Phi)^{-1}$, and hence the only solution is that of Israel and Spanos.[12]

Proof: Since $\Gamma_{,\alpha} = \gamma\Phi_{,\alpha}$, Eqs. (3.3) and (3.4) give $\kappa(-\rho + e^{-\omega/2}\sqrt{4\pi}\sigma^*\Phi^*) = -\gamma e^{-\omega/2}\sqrt{4\pi}\sigma^*$. Hence

$$\kappa\rho = (\gamma + \kappa\Phi^*)e^{-\omega/2}\sqrt{4\pi}\sigma^*. \tag{4.9}$$

By (4.7), $\omega_{,\Phi^*} = \kappa(\gamma + \kappa\Phi^*)^{-1}$ and $\omega = \ln(\gamma + \kappa\Phi^*) + \ln G(\Phi)$. Now $e^{\omega} = (\gamma + \kappa\Phi^*)G(\Phi) = (\gamma^* + \kappa\Phi)(G(\Phi))^*$, so that $G(\Phi) = c^2(\gamma^* + \kappa\Phi)$ and $e^{\omega} = c^2|\gamma^* + \kappa\Phi|^2$. But $c^2|\gamma^* + \kappa\Phi|^2 - \frac{1}{2}\kappa|\Phi|^2 = \mathrm{Re}\Gamma = \frac{1}{2}(\gamma\Phi + \gamma^*\Phi^*) + \mathrm{Re}\delta$. By comparing coefficients of Φ, Φ^* one obtains $c^2 = (16\pi)^{-1}$ and $\mathrm{Re}\delta = (16\pi)^{-1}|\gamma|^2$, which establishes (i) and (iii). (ii) follows from Theorem 1 with $V \equiv 0$. A direct calculation of Eq. (3.2) shows that $R_{\alpha\beta} = 0$, so that the associated V_3 is flat.[22] We define $U = \sqrt{16\pi}(\gamma^* + \kappa\Phi)^{-1}$ and use Eq. (3.3) to obtain (v). By (4.9), $\nabla^2 U = -4\pi\rho|U|^2 U$, which is Eq. (15) of IS with $\rho_{IS} = |U|^2\rho$. Q.E.D.

Note that the function U has to satisfy one real equation

$$U^{-1}\nabla^2 U = U^{*-1}\nabla^2 U^*, \tag{4.10}$$

because the equations

$$i\,\mathrm{curl}(\mathbf{a}) = U^*\nabla U - U\nabla U^* \tag{4.11}$$

must be solved. In case $\rho = 0$, we have $\nabla^2 U = 0$, which gives the PIW stationary electrovac class (cf. IS).

In the case of plane symmetry we may take U to be independent of x^3 and let $z \equiv x^1 + ix^2$. Then (4.10) becomes

$$U^{-1}U_{,zz^*} = U^{*-1}U^*_{,zz^*}. \tag{4.12}$$

One class of solutions is given by $U = \int A(\alpha) \times \exp[\alpha f(z) + \alpha^{-1}(f(z))^*]\,d\alpha$, where $A(\alpha)$ and $f(z)$ are arbitrary analytic functions.

Corollary 3: If the ratio of monopole-to-electric charge is constant, then the above metric must be static.

Proof: From (4.9) we see that σ, and hence $\gamma + \kappa\Phi^*$ and U, must have constant phase. By Eq. (4.11), $\mathrm{curl}(\mathbf{a}) = iU^2\nabla(U^*/U) = 0$, so the metric is static.

ACKNOWLEDGMENTS

The authors are indebted to the NRC Grant No. 3993. One of them (S. K.) is also indebted to NRC Grant No. A4831.

[1] W. B. Bonnor, Proc. Phys. Soc. A 67, 225 (1954).
[2] D. Kramer, G. Neugebauer, and H. Stephani, Forschr. Physik 20, Heft 1, 1 (1972).
[3] W. Kinnersley, J. Math. Phys. 14, 651 (1973). For the historical development of these symmetries, see, e.g., H.A. Buchdahl, Quart. J. Math., Oxford Ser. 5, 116 (1954), R.A. Matzner and C.W. Misner, Phys. Rev. 154, 1229 (1966), R. Geroch, J. Math. Phys. 12, 918 (1971), and S. Kloster, M. M. Som, and A. Das, J. Math. Phys. 15, 1096 (1974).
[4] N. D. Hari Dass, Phys. Rev. Lett. 36, 393 (1976).
[5] U. K. De and A. K. Raychaudhuri, Proc. Roy. Soc. Lond. A 303, 47 (1968).
[6] A. Das, Proc. Roy. Soc. Lond. A 267, 1 (1962).
[7] S. D. Majumdar, Phys. Rev. 72, 390 (1947).
[8] A. Papapetrou, Proc. Roy. Irish Acad. A 51, 191 (1947).
[9] F. J. Ernst, Phys. Rev. 168, 1415 (1968).

[10]Z. Perjes, Phys. Rev. Lett. **27**, 1668 (1971).
[11]W. Israel and G.A. Wilson, J. Math. Phys. **13**, 865 (1962).
[12]W. Israel and J.T.J. Spanos, Lett. Nuovo Cimento **7**, 245 (1973).
[13]Equations (2.1) reduce to (C.28) and (C.29) of KNS by letting $\sqrt{4\pi}\,\sigma=\rho$ and $F_{ij}=-H_{ij}$. Note that KNS use signature +2, which changes the signs of the last three terms of Q_{ij}. We define the dual of a tensor so that $F^*_{14}=\sqrt{-\det\gamma_{ij}}F^{23}$, as in KNS, while some authors use the opposite sign.
[14]Cf. KNS Eq. (C.11). They use ψ for A and χ for $-B$. Our $\Phi=2\Phi^*_{KNS}=-(4\pi)^{-1/2}\Psi_{IW}$.
[15]See the derivation of Eq. (17) in IW.
[16]$\Gamma=\mathcal{E}_{IW}=-4\Gamma^*_{KNS}$.
[17]These equations reduce to (C.42) of KNS. In their equations (b) and (c), the right sides should be $\frac{1}{4}F^{-1}j_4$ and $\frac{1}{4}F^{-1}T_4$.
[18]Table 7 of KNS. See also Kinnersley, Ref. 3.
[19]Some of these were applied to sources by D. Kramer and G. Neugebauer, Ann. Physik **7**, Bd. 27, 129 (1971).
[20]This is a generalization of a transformation given by Buchdahl in Ref. 3.
[21]R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1966), Vol. II.
[22]Cf. sec 4 of IW.

Renormalization group structure for translationally invariant ferromagnets*

George A. Baker, Jr.

Theoretical Division, University of California, Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545

Samuel Krinsky

Brookhaven National Laboratory, Upton, New York 11973
(Received 18 June 1976)

We introduce a correlation function description of the renormalization group approach to critical phenomena. Our work is based on treating the normalization group operator as a *linear mapping* on the set of Ursell functions, rather than as a nonlinear mapping on the space of Hamiltonians. We mainly consider the "mean-spin" renormalization group, but the closely related "decimation" transformation is also considered. Using this approach, we demonstrate for a suitable class of systems that the spectrum of the renormalization group operator is bounded and countably infinitely degenerate. We give counterexamples to the notion that there must be convergence to a renormalization group fixed point. Our formulation of the renormalization group is sufficiently general so that convergence to a fixed point does not necessarily imply hyperscaling, i.e., the vanishing of the anomalous dimension of the vacuum, ω^*. In the case of convergence to a fixed point we find $\delta=(d+\sigma-\omega^*)/(d-\sigma-\omega^*)$, with $\omega^*\geq 0$ necessarily. Above the critical temperature we are able to prove that convergence is obtained to the "infinite temperature fixed point," which result generalizes the central limit theorem to ferromagnetically correlated variables.

1. INTRODUCTION AND SUMMARY

Over the past few years the renormalization group introduced by Wilson[1] into problems of critical phenomena has provided both physical insight and computational methods of fair accuracy.[2,3] As is usual when a fruitful new theoretical method is introduced, the computational aspects have advanced far more than inquiries into the foundations of the theory.[4] In this paper, we undertake a study of certain basic properties of the renormalization group approach. In order to proceed, we have adopted a new formulation of the renormalization group which had its inspiration in a talk given by Høye[5] on spin-cell transformations applied to the Gaussian and spherical models. There are many important foundational questions which are as yet unanswered. For example, the most accurate numerical work[6] on the spin-1/2, three-dimensional Ising model fails to obey hyperscaling, which property is inherent in the renormalization group approach as currently developed. We show that the convergence of the renormalization group procedure to a fixed point does not necessarily require the validity of hyperscaling.

In the theory of critical phenomena, the renormalization group operator is usually thought of as a prescription mapping the space of Hamiltonians into itself. The motivation for this mapping lies in Kadanoff's[7] suggestion that near the critical point, when the correlation length is very large, neighboring spins are very strongly correlated so that the mean of a block of spins should not behave very differently than a single spin. One starts with a space of Hamiltonians, $\mathcal{S}_N$. Since no metric is normally specified, the nature of the space is not normally made explicit. Let us have the spatially uniform Hamiltonians, H_N, depend explicitly on N "spins" interacting in a volume V in d-dimensional space. For each element $H^{(N)}$ belonging to $\mathcal{S}_N$ we suppose there exists a spatial uniform extension to all larger size systems of the same density. In order to define the statistical mechanics for such a system, we need the partition function

$$Z\{H^{(N)}\}=\mathrm{Tr}_N^{\psi}\{\exp(H^{(N)})\} \tag{1.1}$$

where the factor minus $1/kT$ is absorbed into the Hamiltonian and the superscript ψ denotes the set of state vectors over which the trace is taken. The free energy per spin functional is then

$$f_N\{H^{(N)}\}=(1/N)\ln Z\{H^{(N)}\}. \tag{1.2}$$

The renormalization group operators are defined by the following procedure. Extend the original Hamiltonian to a larger volume

$$V'=Vb^d, \quad N'=Nb^d \tag{1.3}$$

and the number of spins as well. Make a change of variables

$$\psi(\mathbf{x})\to\hat{\psi}(\mathbf{y}) \tag{1.4}$$

and define

$$\exp(H_1^{(N)})=\mathrm{Tr}'\{\exp(H^{(N')})\}, \tag{1.5}$$

where Tr' is the partial trace over $(N'-N)$ of the variables $\hat{\psi}$ defining $H^{(N')}$. As an example, consider a one-dimensional set of spins σ_j, $j=1,\ldots,N$. Extend this set to σ_j, $j=1,\ldots,2N$. Define $\hat{\psi}(j)=\sigma_{2j}$, $j=1,\ldots,N$, $\hat{\psi}(j+N)=\sigma_{2j-1}$, $j=1,\ldots,N$. Then the partial trace would consist of summing over the allowed states of $\hat{\psi}(j)$, $j=N+1,\ldots,N'$, $N'=2N$ which leaves a system of N spins (see Nelson and Fisher, and Krinsky and Furman[8]). We then use this procedure to define

$$R_b(H^{(N)})=H_1^{(N)}. \tag{1.6}$$

If we repeat the process, we can define

Copyright © 1977 American Institute of Physics

$$H_2^{(N)} = R_b(H_1^{(N)}) = R_b[R_b(H^{(Nb^{2d})})]$$
$$= \ln \mathrm{Tr}'\mathrm{Tr}''\{\exp(H^{(Nb^{2d})})\}. \tag{1.7}$$

The ambiguity concerning the spatially uniform extention required of $H_1^{(N)}$ is resolved by the assumed extendability of $H^{(N)}$ itself through

$$H_1^{(Nb^d)} = R_b(H^{(Nb^d)}). \tag{1.8}$$

Thus we have

$$H_n^{(N)} = R_b(H_{n-1}^{(N)}) = R_{b^n}(H^{(Nb^{(n-1)d})}). \tag{1.9}$$

If the Hamiltonian family $H^{(N)}$ $(N \to \infty)$ has a thermodynamic limit, then, as

$$\lim_{N\to\infty} f_N = f_\infty, \tag{1.10}$$

we must have

$$\lim_{N\to\infty}[f_N\{H_n^{(N)}\}/b^{nd}] = f_\infty. \tag{1.11}$$

Since, if, as $n \to \infty$, we had the asymptotic relation

$$H_n^{(N)} \sim b^{nd} H_*^{(N)}, \tag{1.12}$$

then by the method of steepest descents we would have a limit like (1.11), workers in the area have been led to introduce a spin rescaling factor c so that $\hat{\psi}(\mathbf{y}) \to c\hat{\psi}(\mathbf{y})$ in order to preserve the basic spin fluctuation magnitude. If the operator $R_{b,c}$ includes this spin rescaling as well as the transformation of R_b, then what is sought is a fixed point of

$$R_{b,c}(H_*^{(N)}) = H_*^{(N)}, \tag{1.13}$$

which would be approached by $H_n^{(N)}$ defined by

$$R_{b,c}(H_n^{(N)}) = H_{n+1}^{(N)}, \quad H_0^{(N)} = H^{(N)}, \tag{1.14}$$

where the behavior of $H_n^{(N)}$ is easily deduced in terms of $H_n^{(N)}$ and $H_*^{(N)}$. Since $R_{b,c}$ is a nonlinear operator, a discussion of the convergence of $H_n^{(N)}$ to a fixed point Hamiltonian $H_*^{(N)}$ involves flows in the Hamiltonian space, S_N. These flows are usually discussed in terms of special models in terms of the behavior of an infinite set of interaction parameters,[3] the thermodynamic significance of which is not completely transparent. This approach has led, among numerous other approaches, to approximate, nonlinear integral recursion relations[1] for H_n, and the expansions[2] about four dimensions in powers of $\epsilon = 4 - d$ for the critical exponents.

To our knowledge, no natural metric has been proposed for the space of Hamiltonians, so at present no definition of convergence exists in a form useful for the proof of theorems. Even if one defined a metric, the nonlinearity of the renormalization group operator specified in (1.5) and (1.6) suggests that an analysis would be difficult. With these remarks in mind, instead of using a Hamiltonian to describe the physical system under consideration, we will describe it in terms of a complete set of correlation functions for an infinite system. We will consider the action of the renormalization group operator on the complete set of Ursell functions, which are defined in terms of the correlation functions by

$$u_0 = \langle 1\rangle, \quad u_1(j) = \langle \sigma_j\rangle,$$
$$u_2(j,k) = \langle \sigma_j\sigma_k\rangle - \langle \sigma_j\rangle\langle \sigma_k\rangle \tag{1.15}$$
$$u_3(j,k,l) = \langle \sigma_j\sigma_k\sigma_l\rangle - \langle \sigma_j\rangle\langle \sigma_k\sigma_l\rangle - \langle \sigma_k\rangle\langle \sigma_j\sigma_l\rangle$$
$$- \langle \sigma_l\rangle\langle \sigma_j\sigma_k\rangle + 2\langle \sigma_j\rangle\langle \sigma_k\rangle\langle \sigma_l\rangle \cdots.$$

In defining the expectation values $\langle\ \rangle$, we suppose that the thermodynamic limit has been taken, i.e., $\langle O\rangle = \lim_{\Lambda\to\infty} \mathrm{Tr} O \exp(H_{\Lambda_0}) \mathrm{Tr}\exp(H_\Lambda)$, where H_Λ is the Hamiltonian for a finite system of size Λ. As Ruelle[9] has shown, the set of all Ursell functions provides a complete description of the thermodynamics of the system.

If we define the new "block spins" as linear combinations of the original spins,

$$s_j = \sum_k A(j,k)\sigma_k, \tag{1.16}$$

then the Ursell functions u_n' of the block spins are given in terms of the original Ursell functions by

$$u_0' = u_0,$$
$$u_1'(J) = \sum_j A(J,j)u_1(j), \tag{1.17}$$
$$u_2'(J,K) = \sum_{j,k} A(J,j)A(K,k)u_2(j,k).$$

In this formulation the renormalization group operator takes the form of a *linear mapping* on the set of correlation functions. The nonlinearity inherent in the previous formulation now resides in the transcription from Ursell functions to Hamiltonians. Since the theory of linear transformations is much more tractable than that of nonlinear ones, we believe this alternate approach will be useful in the study of the general structure of the renormalization group.

We may conveniently write (1.17) as

$$U' = R_A U, \tag{1.18}$$

where $U = \{1, u_1(j), u_2(j,k), \cdots\}$ and the subscript A denotes the dependence of the renormalization group operator on the transformation (1.16). It follows from the definition that

$$R_A R_B = R_{AB}, \tag{1.19}$$

where the subscript AB indicates a matrix product

$$s_j = \sum_{k,l} A(j,k)B(k,l)\sigma_l. \tag{1.20}$$

Clearly, if

$$U_{n+1} = R_A U_n, \tag{1.21}$$

then

$$U_{n+1} = R_A^n U_1 = R_{A^n} U_1. \tag{1.22}$$

The question of convergence to a fixed point becomes the question of whether or not there exists $\lambda_A \neq 0$ such that

$$\lim_{n\to\infty} \lambda_A^{-n} R_{A^n} U_1 = U^* \neq 0, \tag{1.23}$$

where $R_A U^* = \lambda_A U^*$. In Sec. 5, we discuss how to define a metric on the space of correlation vectors $\{U\}$ such that one obtains a reflexive Banach space. Once one has a metric the notion of convergence is then well defined.

At various points in our discussion, we will assume

that the system under consideration satisfies one or more of the following properties:

(1) Translational invariance on a hypercubic lattice.

(2) The spins have finite magnitude in the sense that $|\langle\sigma^n\rangle| \leq M_n(H) < \infty$ in zero or finite magnetic field H.

(3) The Ursell correlation functions can be given a norm in terms of which to discuss convergence in a reflexive Banach space.

(4) The Ursell or truncated correlation functions are asymptotically monotonic for large separations, i.e., $u_n(\lambda\mathbf{i}, \lambda\mathbf{j}, \ldots, \lambda\mathbf{k})$ is monotonic as $\lambda \to \infty$.

(5) The correlation functions have reflection symmetry in the principal planes of the lattice.

(6) The system is ferromagnetic, i.e., the Griffiths[10] and Newman[11] inequalities hold.

(7) The magnetization is continuous at the critical point, that is $\lim_{\epsilon \to 0} M(T_c - \epsilon, H = 0+) = 0$.

These assumptions allow us to make considerable progress investigating the general structure of the renormalization group transformations. However, in Sec. 5, we present a counterexample showing that they are not sufficient to imply convergence to a fixed point. In order to exclude the counterexample, a separate assumption or demonstration for the particular case at hand is required. The crucial aspects involve the asymptotic behavior of the Ursell functions $u_n(\lambda\mathbf{i}, \lambda\mathbf{j}, \ldots, \lambda\mathbf{k})$ and $u_n(\lambda\mathbf{i}, \lambda\mathbf{i} + \delta, \lambda\mathbf{j}, \ldots, \lambda\mathbf{k} + \delta')$ as $\lambda \to \infty$. Although the required smoothness is certainly expected to hold for systems with short-range potentials or smooth long-range potentials, we are not aware of any proof of the required smoothness. This lack indicates that there exist important technical problems which must be overcome in the future in order to establish convergence to a fixed point on a rigorous basis.

Even upon assuming convergence to a fixed point, the renormalization group as developed in this paper does *not* imply hyperscaling, i.e., the satisfaction of critical exponent scaling relations involving the spatial dimensionality d. Violation of hyperscaling has been extensively discussed by Stell[12] in the context of his weak scaling theory. Since the most accurate numerical exponents for the three-dimensional spin-1/2 Ising model apparently violate hyperscaling, the generality of our approach may prove to be of importance.

This paper is organized as follows. In the second section we consider the "decimation" transformation

$$s_{\mathbf{j}} = q_N \sigma_{N\mathbf{j}}, \tag{1.24}$$

and in the third section we study the more interesting "mean-spin" transformation

$$s_{\mathbf{j}} = \frac{q_N}{N^{d/2}} \sum_{\mathbf{r}=0}^{N-1} \sigma_{N\mathbf{j}+\mathbf{r}}, \tag{1.25}$$

or, as some may prefer to call it, the "mean-block-spin." We discuss the close connection between the structure of the mean spin transformation and the decimation transformation, and it turns out that much of the analysis developed in Sec. 2 can be applied in Sec. 3. We are primarily interested in translationally invariant systems, for which

$$u_m(\mathbf{j}_1, \ldots, \mathbf{j}_m) = \psi_{m-1}(\mathbf{j}_1 - \mathbf{j}_2, \ldots, \mathbf{j}_{m-1} - \mathbf{j}_m). \tag{1.26}$$

In Secs. 2 and 3, we study the eigenvalue problem

$$R_N \psi_1^* = \lambda_N \psi_1^*, \tag{1.27}$$

for the decimation and mean-spin transformation, respectively. For the ferromagnetic systems we are considering, the relevant spectrum is found to be bounded and countably degenerate. Our analysis uses the important observation that both transformations preserve certain positivity and monotonicity properties. Considering the action of the renormalization group upon the higher-order Ursell functions, we find that the eigenvalue problem

$$R_N \psi_m^* = \lambda_N \psi_m^* \quad (m \geq 1) \tag{1.28}$$

has simultaneous solutions for all N, with nonzero components ψ_m^* for at least two values of m, only if the spin rescaling factor is restricted to the form,

$$q_N = N^{(\omega^* - \sigma)/2}, \tag{1.29}$$

where ω^* is the anomalous dimension of the vacuum (see Fisher[12]) and σ is defined by the decay of the two-point correlation function as $\langle\sigma_0\sigma_{\mathbf{r}}\rangle \sim r^{\sigma - d}$. The corresponding eigenvalue is

$$\lambda_N = N^{\omega^*}, \tag{1.30}$$

and the eigenvectors behave like

$$\psi_m^*(\lambda\mathbf{j}_1, \ldots, \lambda\mathbf{j}_m) \sim \lambda^{\omega^* - (m+1)(d-\sigma+\omega^*)/2} \quad (\lambda \to \infty). \tag{1.31}$$

We give a representation of the Fourier transform of ψ_m^* in terms of free homogeneous functions Φ_m of specified degree. If the anomalous dimension of the vacuum $\omega^* \neq 0$, then hyperscaling is violated (see Sec. 7).

In Sec. 4 and throughout the rest of the paper, we almost exclusively study the mean-spin transformation. We show that Newman's[11] inequalities imply

$$\omega^* \geq 0. \tag{1.32}$$

When $\omega^* > 0$, there exists a rescaling which leads to a Gaussian fixed point, $\psi_m^* = 0$ $(m > 1)$; however, it does *not* necessarily follow that the critical exponents must be Gaussian, e.g., $\gamma = 1$.

It is natural to define

$$T_N = N^{-\omega^*} R_N, \tag{1.33}$$

so that the eigenvectors of R_N become *fixed points* of T_N, i.e.,

$$T_N \psi_m^* = \psi_m^*. \tag{1.34}$$

In the fifth section we introduce a norm $\|\psi\|_\theta$ on the Ursell functions, parametrized by $\theta \geq 0$. For a given θ, we define a reflexive Banach space $\mathcal{B}_\theta$ by $\|\psi\|_\theta < \infty$. We prove that

$$\|T_N \psi\|_\theta \leq N^\theta \|\psi\|_\theta, \tag{1.35}$$

so that $N^{-\theta} T_N$ is a contraction operator on $\mathcal{B}_\theta$. Unfortunately, the nonzero fixed points of T_N do not belong to $\mathcal{B}_0$, and we show that $\mathcal{B}_0$ cannot be extended to a reflexive Banach space containing nonzero fixed points of T_N in such a manner as to keep T_N a contraction. There-

fore, the powerful convergence theorems[13] available for contraction mappings on reflexive, Banach spaces are insufficient to the application to the renormalization group. We present counterexamples proving that one does not always have convergence to a fixed point.

In the sixth section we use the GKS[10] and Newman[11] inequalities to prove that above the critical temperature, where the susceptibility is finite and the spontaneous magnetization is zero, one has convergence to the "infinite temperature fixed point" under repeated application of the renormalization group operator, as expected in noncritical statistical mechanics. This result is a generalization of the central limit theorem to ferromagnetically correlated variables. The relation of the renormalization group to the central limit theorem is mentioned by Baker[14] in his paper on the hierarchial model[15] and is discussed by Jona-Lasinio.[16]

In the seventh section, we discuss some conclusions which can be drawn if one *assumes* convergence to a critical fixed point. We find

$$\delta = \frac{d+\sigma-\omega^*}{d-\sigma-\omega^*}, \tag{1.36}$$

where $M \sim H^{1/\delta}$ on the critical isotherm, d is the spatial dimensionality, and $\langle \sigma_0 \sigma_r \rangle \sim r^{\sigma-d}$ $(r \to \infty)$. We note that if one carries out Wilson's[1] procedure for calculating the susceptibility index γ in terms of the growth properties of χ^{-1} and $d(\chi^{-1})/dT$, then

$$1 \leq \gamma \leq \gamma_w = \sigma/\zeta_w, \tag{1.37}$$

where γ_w is the result of Wilson's calculation of γ and ζ_w is Wilson's related parameter.

2. THE DECIMATION TRANSFORMATION

In this section we will study a special case of the linear block spin transformation (1.16), commonly called the "decimation" transformation. To begin, let us consider one dimension,

$$s_j = q_2 \sigma_{2j}. \tag{2.1}$$

The renormalization group transformation consists of taking the partial trace over all the odd spin variables and rescaling the even ones. Clearly, the result of n applications is, by (1.22), just

$$u_m^{(n)}(j_1, j_2, \ldots, j_m) = q_2^{mn} u_m(2^n j_1, 2^n j_2, \ldots, 2^n j_m), \tag{2.2}$$

so that the limiting behavior for large n is determined by the asymptotic behavior of the Ursell functions $u_m(\lambda j_i, \lambda j_2, \ldots, \lambda j_m)$ as $\lambda \to \infty$. It is instructive the consider the Fourier transform version of (2.2). We define

$$\underset{\sim}{u}_m(p_1, p_2, \ldots, p_m) = \sum_{j_1=-\infty}^{\infty} \cdots \sum_{j_m=-\infty}^{\infty} u_m(j_1, j_2, \ldots, j_m) \times \exp[2\pi i(p_1 j_1 + \cdots + p_m j_m)], \tag{2.3}$$

where the momenta $p_1, \ldots, p_m$ run over a single period, i.e., $0 \leq p_i < 1$. For $n=1$, (2.2) becomes[5]

$$\underset{\sim}{u}_m^{(1)}(p_1, p_2, \ldots, p_m) = (q_2/2)^m \sum_{\delta_1=0}^{1} \cdots \sum_{\delta_1=0}^{1} \underset{\sim}{u}_m(\tfrac{1}{2}p_1 + \tfrac{1}{2}\delta_1, \ldots, \tfrac{1}{2}p_m + \tfrac{1}{2}\delta_m), \tag{2.4}$$

where use was made of the identity $1 + \exp(i\pi) = 0$.

We can easily generalize the above discussion to refer to the transformation

$$s_j = q_N \sigma_{Nj}, \tag{2.5}$$

instead of (2.1). One obtains

$$\underset{\sim}{u}_m^{(1)}(p_1, p_2, \ldots, p_m) = \left(\frac{q_N}{N}\right)^m \sum_{\delta_1=0}^{N-1} \cdots \sum_{\delta_m=0}^{N-1} \underset{\sim}{u}_m\left(\frac{1}{N}(p_1+\delta_1), \ldots, \frac{1}{N}(p_m+\delta_m)\right), \tag{2.6}$$

using the identity, $(1/N)\sum_{\delta=0}^{N-1} \exp(2\pi i j/N) = 1$ or 0, as j is divisible by N or not.

By a suitable reinterpretation, the above equations [(2.1)–(2.6)] immediately extend to d dimensions. We treat the hypercubic lattices (plane square, simple cubic etc.) and now consider j as a vector ranging over all integer lattice points $\mathbf{j} = (j_1, \ldots, j_d)$. The unit period for the vector momenta becomes $0 \leq \mathbf{p} \cdot \mathbf{e}_\alpha < 1$, where $\mathbf{e}_\alpha$ is a unit vector in the αth direction. Each sum on delta becomes a multiple sum over each component separately. Thus we generalize (2.6) to be

$$\underset{\sim}{u}_m^{(1)}(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m) = \left(\frac{q_N}{N^d}\right)^m \sum_{\delta_1=0}^{N-1} \cdots \sum_{\delta_m=0}^{N-1} \underset{\sim}{u}_m\left(\frac{1}{N}(\mathbf{p}_1+\delta_1), \ldots, \frac{1}{N}(\mathbf{p}_m+\delta_m)\right), \tag{2.7}$$

and, of course, (2.2) for $n=1$ to be

$$u_m^{(1)}(\mathbf{j}_1, \mathbf{j}_2, \ldots, \mathbf{j}_m) = q_N^m u_m(N\mathbf{j}_1, N\mathbf{j}_2, \ldots, N\mathbf{j}_m). \tag{2.8}$$

If we now assume translational invariance, certain simplifications can be made in the Fourier transform versions. First define

$$\mathbf{k}_1 = \mathbf{p}_1, \quad \mathbf{k}_2 = \mathbf{p}_1 + \mathbf{p}_2, \quad \cdots, \quad \mathbf{k}_m = \mathbf{p}_1 + \mathbf{p}_2 + \cdots + \mathbf{p}_m. \tag{2.9}$$

Then the conservation of total momentum implies that $\underset{\sim}{u}_m^{(n)}$ has a factor of $\delta(\mathbf{k}_m)$. Of course, $\underset{\sim}{u}_1^{(n)}$ is simply proportional to $\delta(\mathbf{k}_1)$. It is convenient to define the reduced functions ψ and Ω by

$$\underset{\sim}{u}_m^{(n)}(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m) = \delta(\mathbf{k}_m)\Omega_{m-1}^{(n)}(\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_{m-1}), \tag{2.10}$$

$$u_m^{(n)}(\mathbf{j}_1, \mathbf{j}_2, \ldots, \mathbf{j}_m) = \psi_{m-1}^{(n)}(\mathbf{j}_2 - \mathbf{j}_1, \mathbf{j}_3 - \mathbf{j}_2, \ldots, \mathbf{j}_m - \mathbf{j}_{m-1}). \tag{2.11}$$

Hence, for the translationally invariant system (2.7) and (2.8) become

$$\psi_m^{(1)}(\mathbf{r}_1, \ldots, \mathbf{r}_m) = q_N^{m+1} \psi_m(N\mathbf{r}_1, \ldots, N\mathbf{r}_m), \tag{2.12}$$

$$\Omega_m^{(1)}(\mathbf{k}_1, \ldots, \mathbf{k}_m) = q_N \left(\frac{q_N}{N^d}\right)^m \sum_{\delta_1=0}^{N-1} \cdots \sum_{\delta_m=0}^{N-1} \Omega_m\left(\frac{1}{N}(\mathbf{k}_1+\delta_1), \ldots, \frac{1}{N}(\mathbf{k}_m+\delta_m)\right), \tag{2.13}$$

where use has been made of the periodicity in the momenta and the δ's have been relabelled.

As was observed by Gallavotti and Knops[17] in the context of the hierarchical model, the renormalization group preserves certain properties. One can verify from our equations that if for some n and any m one or more of the conditions

$$u_m^{(n)} \geq 0, \quad \psi_m^{(n)} \geq 0, \quad \underset{\sim}{u}_m^{(n)} \geq 0, \quad \Omega_n^{(n)} \geq 0 \tag{2.14}$$

holds, then the same condition or conditions hold for all higher values of n. Positivity is preserved both in the real space and the Fourier space versions.

Another important property is that monotonicity is preserved. We state this result for the two-spin correlations in one dimension. If

$$\psi_1^{(n)}(j+1) \leq \psi_1^{(n)}(j), \tag{2.15}$$

then the same is true for all higher n. The preservation of this monotonicity property follows directly from (1.22), where we see that higher iterates are just scaled subsets of the variables that appear in lower iterates. Various extentions to multipoint correlations and higher dimensions can be given. One use of conservation laws is that they can be used to rule out convergence to certain fixed points given information about the initial correlation vector.

Let us now consider the eigenvalue problem associated with the decimation transformation. We first consider a one-dimensional translationally invariant system and study (2.5) with $N=2$. We denote this renormalization group transformation by R_2, and determine its spectrum in the invariant subspace formed by the two-point functions. The eigenvalue equation is

$$R_2\psi_1^* = \lambda_2\psi_1^*, \tag{2.16}$$

and, from (2.12), this becomes

$$q_2^2\psi_1^*(2j) = \lambda_2\psi_1^*(j). \tag{2.17}$$

Equation (2.17) has directly the solutions

$$\begin{aligned}
&\psi_1^*(0) = \lambda_2 q_2^{-2}\psi_1^*(0),\\
&\psi_1^*(2) = \lambda_2 q_2^{-2}\psi_1^*(1), \quad \psi_1^*(4) = \lambda_2^2 q_2^{-4}\psi_1^*(1),\\
&\qquad\qquad \psi_1^*(8) = \lambda_2^3 q_2^{-6}\psi_1^*(1), \cdots,\\
&\psi_1^*(6) = \lambda_2 q_2^{-2}\psi_1^*(3), \quad \psi_1^*(12) = \lambda_2^2 q_2^{-4}\psi_1^*(3),\\
&\qquad\qquad \psi_1^*(24) = \lambda_2^3 q_2^{-6}\psi_1^*(3), \cdots,\\
&\psi_1^*(10) = \lambda_2 q_2^{-2}\psi_1^*(5), \quad \psi_1^*(20) = \lambda_2^2 q_2^{-4}\psi_1^*(5),\\
&\qquad\qquad \psi_1^*(40) = \lambda_2^3 q_2^{-6}\psi_1^*(5), \cdots.
\end{aligned} \tag{2.18}$$

These solutions are completely parameterized in terms of the scale factor $\lambda_2 q_2^{-2}$, and the components $\psi_1^*(2j-1)$ of the vector at odd separations. The restriction

$$|\psi_1(j)| \leq C, \tag{2.19}$$

which follows for physical spins of finite size, implies that if $\psi_1^* \neq 0$, then $|\lambda_2| \leq q_2^2$ for a physical fixed point of (2.17) to exist. Since the physical spins being treated are real, and we choose q_2^2 to be real, λ_2 is necessarily real, and so in this special case the spectrum of R_2 is

$$-q_2^2 \leq \lambda_2 \leq q_2^2.$$

For each point in this spectrum there is a countable infinity of independent eigenvectors indexed by an *odd* integer J. The complete set of such eigenvectors is given by

$$\begin{aligned}
f^{(J)}(j) &= (\lambda_2 q_2^{-2})^\pi \quad \text{for } j = J2^\pi (\pi = 0, 1, 2, \cdots)\\
&= 0 \quad \text{otherwise},
\end{aligned} \tag{2.20}$$

where J is an odd integer. The $j=0$ equation of (2.18) admits a nontrivial solution only at the end point of the spectrum, $\lambda_2 = q_2^2$. This solution is

$$f^{(0)}(j) = \begin{cases} 1, & j=0,\\ 0, & \text{otherwise}. \end{cases} \tag{2.21}$$

Thus, as any solution of (2.17) can be expanded as

$$\psi_1^*(j) = \psi_1^*(0) f^{(0)}(j) + \sum_{J=0}^{\infty} f^{(2J+1)}(j)\psi_1^*(2J+1), \tag{2.22}$$

the set of vectors $f^{(J)}(j)$ given by (2.20) and (2.21) are complete. The values of the ψ_1^* in (2.22) are arbitrary, except $\psi_1^*(0) = 0$ unless $\lambda_2 = q_2^2$.

Clearly the solutions (2.20) are not monotonically decreasing, but linear combinations of them may be. We give two examples of monotonically decreasing solutions of (2.17) to show that this condition does not uniquely determine an eigenvector. Let $E(j) = (\ln j)/(\ln 2)$; then

$$f_1(j) = (\lambda_2 q_2^{-2})^{E(j)}$$

and

$$f_2(j) = (\lambda_2 q_2^{-2})^{[E(j)]}$$

are both monotonic solutions of (2.17), where $[x]$ denotes the Gauss greatest integer function.

If we consider as well as R_2, R_p where p is a prime number, then by (1.19)

$$R_2 R_p = R_p R_2 = R_{2p}. \tag{2.23}$$

Thus, if ψ^* is an eigenfunction of R_2, then, by commutivity, so too is $R_p\psi^*$. We seek therefore to construct, from among the degenerate eigenvectors of R_2, an eigenvector which is also an eigenvector of R_p. The equation would be, from (2.12),

$$\lambda_p\psi_1^*(j) = q_p^2\psi_1^*(pj), \tag{2.24}$$

which has the solutions

$$\begin{aligned}
&\psi_1^*(p) = (\lambda_p q_p^{-2})\psi_1^*(1), \quad \psi_1^*(p^2) = (\lambda_p q_p^{-2})^2\psi_1^*(1), \cdots,\\
&\psi_1^*(2p) = (\lambda_p q_p^{-2})\psi_1^*(2), \quad \psi_1^*(2p^2) = (\lambda_p q_p^{-2})^2\psi_1^*(2), \cdots.
\end{aligned} \tag{2.25}$$

In particular, if for example $p=3$, the following sets of ψ^*'s are linked together

$$\begin{aligned}
&0,\\
&1, 2, 3, 4, 6, 8, 9, 12, \cdots,\\
&5, 10, 15, 20, 30, 40, \cdots,\\
&7, 14, 21, 28, \cdots,\\
&\cdots
\end{aligned} \tag{2.26}$$

Thus, we can construct a denumerably infinite set of simultaneous eigenvectors of R_2 and R_3. By continuing this procedure, we find that for

$$-q_p^2 \leq \lambda_p \leq q_p^2 \tag{2.27}$$

we can construct an eigenvector of all R_p, p a prime perpendicular to the $j=0$ component, and if $\lambda_p = q_p^2$ for all p, the $j=0$ component is also an eigenvector. If $\pi(J,p)$ is the number of times p divides J, then for this common eigenvector we have

$$\psi_1^*(J) = \left[\prod_{p \text{ prime}} (\lambda_p q_p^{-2})^{\pi(J,p)}\right]\psi_1^*(1), \quad \psi_1^*(0) = 0. \tag{2.28}$$

Because of the appearance of the number theoretic function $\pi(J,p)$ in (2.28) the behavior $\psi_1^*(J)$ as a function of J can be very erratic. In the case of a simple ferromagnetic system, one expects a smooth monotonic decay with J, for large J. As we observed earlier in (2.15), if the original spin–spin correlation function has a monotonic decay property, then this property is preserved under iteration of the renormalization group. Therefore, the physically relevant fixed point should be monotonically decreasing, and as we will now show, this property suffices to determine all $(\lambda_p q_p^{-2})$ in terms of $(\lambda_2 q_2^{-2})$.

The proof of this uniqueness is as follows. Given any $\epsilon > 0$, there exist integers A, B, C, D such that

$$-\epsilon < \frac{C}{D} - \frac{\ln p}{\ln 2} < 0 < \frac{A}{B} - \frac{\ln p}{\ln 2} < \epsilon. \tag{2.29}$$

Since $2^{BC} < p^{BD} < 2^{AD}$, the monotonicity of the solution $\psi_1^*(j)$ implies

$$\psi_1^*(2^{BC}) \geq \psi_1^*(p^{BD}) \geq \psi_1^*(2^{AD}). \tag{2.30}$$

Using $\psi_1^*(2j) = \lambda_2 q_2^{-2}\psi_1^*(j)$ and $\psi_1^*(pj) = \lambda_p q_p^{-2}\psi_1^*(j)$, we find

$$(\lambda_2 q_2^{-2})^{BC} \geq (\lambda_p q_p^{-2})^{BD} \geq (\lambda_2 q_2^{-2})^{AD}. \tag{2.31}$$

It follows from (2.29) and (2.31) that

$$\left| \frac{\ln(\lambda_p q_p^{-2})}{\ln(\lambda_2 q_2^{-2})} - \frac{\ln p}{\ln 2} \right| < \epsilon,$$

and since this inequality holds for any $\epsilon > 0$, we deduce

$$\lambda_p q_p^{-2} = (\lambda_2 q_2^{-2})^{\ln p/\ln 2} = p^{-(1-\sigma)}, \tag{2.32}$$

where we define

$$\sigma = \ln(2\lambda_2 q_2^{-2})/\ln 2 \leq 1, \tag{2.33}$$

by (2.27).

Consequently, normalizing $\psi_1^*(1) = 1$, we have

$$\begin{aligned} &\psi_1^*(j) = j^{-(1-\sigma)}, \quad j = 1, 2, 3, \cdots, \\ &\psi_1^*(0) = 0. \end{aligned} \tag{2.34}$$

When $\sigma = 1$, there is the additional eigenvector $\psi_1^*(0) = 1$ and $\psi_1^*(j) = 0$ for $j \neq 0$.

The spectrum is restricted to

$$0 \leq \lambda_p \leq q_p^2 \tag{2.35}$$

by monotonicity.

Since any R_N can be decomposed as

$$R_N = \prod_{p=\text{prime}} R_p^{\tau(N,p)},$$

the fixed point ψ_1^* defined by (2.34) is the unique monotonic fixed point (given $\sigma < 1$) of the entire renormalization group $\{R_N\}$ for a one-dimensional translationally invariant lattice.

Let us now extend the above analysis to higher dimensions, still considering the translationally invariant two-spin correlation function. From (2.8), we obtain the eigenvalue equation $(N = 2)$

$$q_2^2\psi_1^*(2\mathbf{j}) = \lambda_2\psi_1^*(\mathbf{j}). \tag{2.36}$$

We may write out explicitly for $d = 2$ the solutions to (2.36); they are

$$\begin{aligned} &\psi_1^*(0,0) = \lambda_2 q_2^{-2}\psi_1^*(0,0), \\ &\psi_1^*(2,0) = \lambda_2 q_2^{-2}\psi_1^*(1,0), \quad \psi_1^*(4,0) = (\lambda_2 q_2^{-2})^2\psi_1^*(1,0), \cdots, \\ &\psi_1^*(0,2) = \lambda_2 q_2^{-2}\psi_1^*(0,1), \quad \psi_1^*(0,4) = (\lambda_2 q_2^{-2})^2\psi_1^*(0,1), \cdots, \\ &\psi_1^*(2,2) = \lambda_2 q_2^{-2}\psi_1^*(1,1), \quad \psi_1^*(4,4) = (\lambda_2 q_2^{-2})^2\psi_1^*(1,1), \cdots, \\ &\psi_1^*(6,0) = \lambda_2 q_2^{-2}\psi_1^*(3,0), \quad \psi_1^*(12,0) = (\lambda_2 q_2^{-2})^2\psi_1^*(3,0), \cdots, \\ &\psi_1^*(4,2) = \lambda_2 q_2^{-2}\psi_1^*(2,1), \quad \psi_1^*(8,4) = (\lambda_2 q_2^{-2})^2\psi_1^*(2,1), \cdots. \end{aligned} \tag{2.37}$$

These solutions are completely parametrized in terms of $\lambda_2 q_2^{-2}$ and certain of the independent amplitudes. Along the lattice lines these are just $\psi_1^*(2j-1, 0)$ and $\psi_1^*(0, 2j-1)$ as before. The prescription along any ray is similar. For example, on the ray through (3, 1) the independent amplitudes are $\psi_1^*(6j-3, 2j-1)$, and so on for the other cases. The imposition of the condition of monotonicity along each ray, at least for asymptotically large separations, together with the requirement that we have a simultaneous eigenvector of all the R_p (and hence all R_N) reduces the degeneracy to one free amplitude for each ray. When $\lambda_p = q_p^2$, we again have the extra eigenvector $\psi_1^*(0,0) = 1$, $\psi_1^*(\mathbf{j}) = 0$ for $\mathbf{j} \neq 0$. For dimensions higher than two the analysis follows that for $d = 2$, and we find again that the spectrum is restricted to

$$0 \leq \lambda_p \leq q_p^2, \tag{2.38}$$

and

$$\psi_1^*(\mathbf{j}) = F(\mathbf{j}\cdot\mathbf{e}_1, \ldots, \mathbf{j}\cdot\mathbf{e}_d), \tag{2.39}$$

where F is a homogeneous function of its d arguments of degree $\sigma - 1$. The $\mathbf{e}_\alpha$ are unit vectors along the lattice lines and σ is defined by (2.33). The eigenvector of (2.39) satisfies

$$\psi_1^*(\lambda\mathbf{j}) \sim \lambda^{-(1-\sigma)} \quad (\lambda \to \infty), \tag{2.40}$$

and the freedom of the homogeneous function is needed to allow for the possibly different amplitude along each ray. We remark that translational invariance automatically implies $\psi_1^*(\mathbf{j}) = \psi_1^*(-\mathbf{j})$. This property restricts the allowed F's in (2.39). For a ferromagnetic system Griffiths' inequalities[10] imply $F \geq 0$.

Let us close this section by emphasizing a weakness of the decimation transformation pointed out by Høye.[5] It follows from general considerations that the Fourier transform $\Omega_1^*(p_1)$ must be positive. Since as observed in (2.14) positivity is preserved upon iteration, the physically relevant fixed point $\Omega_1^*(p_1)$ must be positive. Since by (2.34) $\psi_1^*(0) = 0$, it follows that

$$\int_0^1 \Omega_1^*(p_1)\, dp_1 = 0, \tag{2.41}$$

so that the fixed points for the decimation transformation do not obey the necessary positivity condition.

3. THE MEAN-SPIN TRANSFORMATION

In this section, and in the rest of the paper, we will mainly consider a renormalization group based on a block-spin variable proportional to the mean spin of the block. For a one-dimensional system, we select

$$s_j = \frac{q_N}{N^{1/2}} \sum_{r=0}^{N-1} \sigma_{Nj+r}. \tag{3.1}$$

The transformation induced on the Ursell function is

$$u_m^{(1)}(j_1, \ldots, j_m) = \left(\frac{q_N}{N^{1/2}}\right)^m \sum_{r_1=0}^{N-1} \cdots \sum_{r_m=0}^{N-1} u_m(Nj_1 + r_1, \ldots, Nj_m + r_m). \tag{3.2}$$

The Fourier transform becomes

$$\underset{\sim}{u}_m^{(1)}(p_1, \ldots, p_m) = \left(\frac{q_N}{N^{3/2}}\right)^m \sum_{\delta_1=0}^{N-1} \cdots \sum_{\delta_1=0}^{N-1} u_m\left(\frac{p_1+\delta_1}{N}, \ldots, \frac{p_m+\delta_m}{N}\right) \times \prod_{\lambda=1}^{m} \frac{\exp(-2\pi i p_\lambda) - 1}{\exp(-2\pi i(p_\lambda + \delta_\lambda)/N) - 1}, \tag{3.3}$$

where use has been made of Van der Monde determinants and identities involving the roots of unity in deriving the above result. As pointed out by Høye[5] for the special case $m=2$, $N=2$ on a translationally invariant lattice, one can make a simple transformation reducing (3.3) to the form (2.6) found for the decimation renormalization group. Defining

$$\underset{\sim}{u}_m^{(n)}(p_1, \ldots, p_m) = \underset{\sim}{Q}_m^{(n)}(p_1, \ldots, p_m) \prod_{\lambda=1}^{m} [\exp(-2\pi i p_\lambda) - 1], \tag{3.4a}$$

we find

$$\underset{\sim}{Q}_m^{(1)}(p_1, \ldots, p_m) = \left(\frac{q_N}{N^{3/2}}\right)^m \sum_{\delta_1=0}^{N-1} \cdots \sum_{\delta_m=0}^{N-1} \underset{\sim}{Q}_m\left(\frac{p_1+\delta_1}{N}, \ldots, \frac{p_m+\delta_m}{N}\right). \tag{3.4b}$$

Comparing (3.4b) with (2.6), we see the close relationship between the mean spin and decimation transformations. Further insight is gained by defining the coordinate space function

$$Q_m^{(n)}(j_1, \ldots, j_m) = \sum_{l_1=0}^{j_1-1} \cdots \sum_{l_m=0}^{j_m-1} u_m^{(n)}(l_1, \ldots, l_m), \tag{3.5a}$$

for $j_\alpha \geq 1$. Then using (3.2) we find

$$Q_m^{(1)}(j_1, \ldots, j_m) = \left(\frac{q_N}{N^{1/2}}\right)^m Q_m(Nj_1, \ldots, Nj_m), \tag{3.5b}$$

which has the form of a decimation transformation (2.8). Let us note that if for a given m all the Ursell functions u_m have the same sign, the Q_m will have certain monotonicity properties which will be preserved under iteration of (3.5b). The transformation (3.5a) can be inverted by taking partial differences. Let us illustrate this point for $m=2$:

$$u_2(j,k) = Q_2(j+1, k+1) - Q_2(j, k+1) - Q_2(j+1, k) + Q_2(j, k) \quad (j, k \geq 1).$$

In considering a translationally invariant system, we use (2.9) and (2.10) together with (3.3) to obtain

$$\Omega_1^{(1)}(k_1, \ldots, k_m) = N\left(\frac{q_N}{N^{3/2}}\right)^{m+1} \sum_{\delta_1=0}^{N-1} \cdots \sum_{\delta_m=0}^{N-1} \Omega_m\left(\frac{k_1+\delta_1}{N}, \frac{k_1+\delta_1+\delta_2}{N}, \ldots, \frac{k_m+\delta_1+\cdots+\delta_m}{N}\right) \frac{\sin\pi k_1}{\sin[\pi(k_1+\delta_1)/N]} \times \frac{\sin\pi(k_2-k_1)\cdots\sin\pi(k_m-k_{m-1})\sin\pi k_m}{\sin[\pi(k_2-k_1+\delta_2)/N]\cdots\sin[\pi(k_m-k_{m-1}+\delta_m)/N]\sin[\pi(k_m+\delta_1+\cdots+\delta_m)/N]}, \tag{3.6}$$

where use has been made of momentum conservation on both the left- and right-hand sides of (3.3) in the derivation. Note should be made that the product term is explicitly nonnegative in the case $m=1$. If we make the simple transformation,

$$\Omega_m^{(n)}(k_1, \ldots, k_m) = Q_m^{(n)}(k_1, \ldots, k_m) \sin\pi k_1 \sin\pi(k_2-k_1)\cdots \times \sin\pi(k_m - k_{m-1}) \sin\pi k_m, \tag{3.7}$$

then the transformation (3.6) is reduced to the form of a decimation transformation, except the scale factor is now $N(q_N/N^{3/2})^{m+1}$.

For the special case $m=1$, $N=2$, Eq. (3.6) becomes

$$\Omega_1^{(1)}(k) = \tfrac{1}{2} q_2^2 [\Omega_1(k/2)(1+\cos\pi k) + \Omega_1((k+1)/2) \times (1-\cos\pi k)], \tag{3.8}$$

which agrees with Høye[5] upon noting that his spin rescaling factor $q_{2H} = q_2\sqrt{2}$. For $m=1$, Eq. (3.7) is

$$\Omega_1^{(n)}(k) = Q_1^{(n)}(k)(1-\cos 2\pi k)/2, \tag{3.9}$$

which is the transformation used by Høye.[5]

Let us now consider a translationally invariant ferromagnetic system and characterize the two-spin correlation functions which can be fixed points of the mean-spin renormalization group. The action of the renormalization group on the real-space reduced two-point function defined in (2.11) is given by

$$\psi_1^{(1)}(j) = q_n^2 \sum_{r=1-N}^{N-1} \left(1 - \frac{|r|}{N}\right) \psi_1(Nj + r). \tag{3.10}$$

We will now consider the structure of the mean-spin renormalization group. In order to conveniently study this group, we make the substitution

$$\begin{aligned} \psi_1(0) &= 2t(1), \\ \psi_1(1) &= t(2) - 2t(1), \\ \psi_1(j) &= t(j+1) - 2t(j) + t(j-1), \quad j \geq 2. \end{aligned} \tag{3.11}$$

Since translation invariance implies $\psi_1(j) = \psi_1(-j)$, we need consider only nonnegative j. The inverse of (3.11) is

$$\begin{aligned} t(1) &= \tfrac{1}{2}\psi_1(0), \\ t(j) &= \tfrac{1}{2} j \psi_1(0) + \sum_{r=1}^{j-1} (j-r)\psi_1(r). \end{aligned} \tag{3.12}$$

Partial sums of the set of equations (3.10) yield the set

$$t^{(1)}(j) = (q_N^2/N) t(Nj), \quad j \geq 1, \tag{3.13}$$

which is identical to the decimation transformation (2.12) with $m=1$, except q_N^2 is now replaced by q_N^2/N.

From (3.11) we find

$$t(j+1)-t(j)=\tfrac{1}{2}\psi_1(0)+\sum_{r=1}^{j}\psi_1(r)\geq 0, \tag{3.14}$$

where the positivity follows from Griffiths' inequality[10] $\psi_1(r)\geq 0$. This equation shows that, for a ferromagnet, $t(j)$ is a *monotonically increasing* function of j, and since monotonicity is preserved by iteration of (3.13), it must hold for any resultant fixed point. Moreover, for a reasonable model, as $\psi_1(0)$ is the mean square of the spin size, it is not zero, so it follows from (3.12) that $t(j)$ should increase at least linearly with j. Also, the requirement (2.19) of finite spin size, $|\psi(j)|\leq C$, implies by (3.12) that $t(j)$ should not increase more rapidly than j^2. It now follows that the fixed point equation,

$$\lambda_N t^*(j)=(q_N^2/N)t^*(Nj), \tag{3.15}$$

has the spectrum

$$q_N^2\leq\lambda_N\leq Nq_N^2. \tag{3.16}$$

An argument like that leading to (2.32) shows that the derived monotonicity of $t^*(j)$ implies that the only fixed points of the entire renormalization group [i.e., simultaneous eigenvector solution of (3.15) for all N] are

$$t^*(j)=j^{1+\sigma}, \tag{3.17}$$

$$0\leq\sigma=\ln(\lambda_2 q_2^{-2})/\ln 2=\ln(\lambda_N q_N^{-2})/\ln N\leq 1, \tag{3.18}$$

where inequalities $0\leq\sigma\leq 1$ follow from (3.16). Using (3.11), we find

$$\begin{aligned}&\psi_1^*(0)=2,\\ &\psi_1^*(1)=2^{1+\sigma}-2,\\ &\psi_1^*(l)=(j+1)^{1+\sigma}-2j^{1+\sigma}+(j-1)^{1+\sigma},\quad j\geq 2.\end{aligned} \tag{3.19}$$

Note that $\psi_1^*(j)$ is positive and monotonically decreasing, with the asymptotic behavior

$$\psi_1^*(j)\sim j^{\sigma-1}\quad\text{for } j\to\infty\quad(0<\sigma\leq 1), \tag{3.20}$$

just as was the case for the decimation transformation. When $\sigma\to 0$, it follows directly from (3.19) that

$$\begin{aligned}&\psi_1^*(0)=2\to 2,\\ &\psi_1^*(1)\approx\sigma 2\ln 2\to 0,\\ &\psi_1^*(j)\approx\sigma(j)^{\sigma-1}\to 0.\end{aligned}$$

When $\sigma=1$, $\psi^*(j)=2$.

As mentioned at the end of Sec. 2, it follows from general considerations that the Fourier transform $\Omega_1(k)\geq 0$. We also saw that this positivity is preserved under iteration of (3.6), so that, for a physically relevant fixed point, $\Omega_1^*(k)$ must be positive. It follows from (2.41) that the fixed points of the decimation transformation were not positive. The direct Fourier transform of (3.19)

$$\Omega_1^*(k)=2+2(2^\sigma-1)\cos 2\pi k+\sum_{j=2}^{\infty}[(j+1)^{1+\sigma}-2j^{1+\sigma}+(j-1)^{1+\sigma}]\cos(2\pi kj) \tag{3.21a}$$

can be recast as

$$\Omega_1^*(k)=A\sum_{n=-\infty}^{\infty}\frac{1-\cos(2\pi k)}{|k+n|^{2+\sigma}}, \tag{3.21b}$$

which is manifestly positive. The normalization constant is given by

$$A^{-1}=2\int_0^\infty\frac{1-\cos(2\pi x)}{x^{2+\sigma}}dx, \tag{3.22}$$

where we remember that $0\leq\sigma\leq 1$ for this fixed point. In order to verify (3.21b), we need only compute its Fourier expansion coefficients and compare them with (3.21a).

Let us now proceed to consider a d-dimensional system. The block-spin variable is taken to be

$$S_{\mathbf{j}}=\frac{q_N}{N^{d/2}}\sum_{\mathbf{r}}S_{N\mathbf{j}+\mathbf{r}}. \tag{3.23}$$

The direct generalization of (3.10) is

$$\psi_1(\mathbf{j})=q_N^2\sum_{r_1=1-N}^{N-1}\cdots\sum_{r_d=1-N}^{N-1}\left(1-\frac{|r_1|}{N}\right)\cdots\left(1-\frac{|r_d|}{N}\right)\times\psi_1(N\mathbf{j}+\mathbf{r}). \tag{3.24}$$

When we constructed the transformation (3.11)–(3.12) for the one-dimensional system, the symmetry $\psi_1(j)=\psi_1(-j)$ implicit in a translationally invariant system was used to provide a starting point for the equations. By restricting our attention to systems with reflection invariance about lattice hyperplanes,

$$\mathbf{e}_\alpha\cdot\mathbf{j}\to-\mathbf{e}_\alpha\cdot\mathbf{j},\quad\alpha=1,\ldots,d, \tag{3.25}$$

we may construct a similar transformation in d dimensions. We illustrate this with $d=2$:

$$\begin{aligned}
&\psi_1(0,0)=4t(1,1),\\
&\psi_1(0,1)=2t(1,2)-4t(1,1),\\
&\psi_1(0,k)=2t(1,k+1)-4t(1,k)+2t(1,k-1),\quad k\geq 2,\\
&\psi_1(1,0)=2t(2,1)-4t(1,1),\\
&\psi_1(j,0)=2t(j+1,1)-4t(j,1)+2t(j-1,1),\quad j\geq 2,\\
&\psi_1(1,1)=t(2,2)-2t(1,2)-2t(2,1)+4t(1,1),\\
&\psi_1(1,k)=t(2,k+1)-2t(2,k)+t(2,k-1)\\
&\qquad-2t(1,k+1)+4t(1,k)-2t(1,k-1),\quad k\geq 2,\\
&\psi_1(j,1)=t(j+1,2)-2t(j,2)+t(j-1,2)\\
&\qquad-2t(j+1,1)+4t(j,1)-2t(j-1,1),\quad j\geq 2,\\
&\psi_1(j,k)=t(j+1,k+1)-2t(j+1,k)+t(j+1,k-1)\\
&\qquad-2t(j,k+1)+4t(j,k)-2t(j,k-1)\\
&\qquad+t(j-1,k+1)-2t(j-1,k)+t(j-1,k-1),\\
&j,k\geq 2.
\end{aligned} \tag{3.26}$$

The inverse of this transformation is

$$\begin{aligned}
&t(1,1)=\tfrac{1}{4}\psi_1(0,0),\\
&t(1,k)=\tfrac{1}{4}k\psi_1(0,0)+\tfrac{1}{2}\sum_{l=1}^{k-1}(k-l)\psi_1(0,l),\\
&t(j,1)=\tfrac{1}{4}j\psi_1(0,0)+\tfrac{1}{2}\sum_{m=1}^{j-1}(j-m)\psi_1(m,0),\\
&t(j,k)=\tfrac{1}{4}jk\,\psi_1(0,0)+\tfrac{1}{2}\sum_{l=1}^{k-1}j(k-l)\psi_1(0,l)
\end{aligned}$$

$$+\tfrac{1}{2}\sum_{m=1}^{j-1} k(j-m)\psi_1(m,0)$$
$$+\sum_{m=1}^{j-1}\sum_{l=1}^{k-1}(j-m)(k-l)\psi_1(m,l). \tag{3.27}$$

Under these transformations, as extended to d dimensions, the system of Eqs. (3.24) become

$$t^{(1)}(\mathbf{j}) = (q_N^2/N^d)t(N\mathbf{j}). \tag{3.28}$$

From the consideration of (3.27) and its generalization, we conclude that, as $\lambda\to\infty$, for ferromagnetic models there exist $A, B > 0$ such that

$$A\lambda^d \le t(\lambda\mathbf{j}) \le B\lambda^{2d}, \tag{3.29}$$

where the left-hand inequality follows from $\psi_1(0,0) > 0$ and the right-hand inequality from $|\psi_1(\mathbf{j})| \le C$. If we introduce the eigenvalue equation

$$\lambda_N t^*(\mathbf{j}) = (q_N^2/N^d)t^*(N\mathbf{j}), \tag{3.30}$$

then inequality (3.29) implies the restriction

$$q_N^2 \le \lambda_N \le N^d q_N^2, \tag{3.31}$$

which reduces to (3.16) for the special case $d=1$. Griffiths' inequalities[10] imply $t(\mathbf{j})$ is monotonically increasing, which together with (3.31) suffices to establish the form

$$t^*(\mathbf{j}) = T(\mathbf{j}\cdot\mathbf{e}_1, \ldots, \mathbf{j}\cdot\mathbf{e}_d), \tag{3.32}$$

where T is a homogeneous function of degree $d+\sigma$, with

$$0 \le \sigma = \ln(\lambda_2 q_2^{-2})/\ln 2 = \ln(\lambda_N q_N^{-2})/\ln N \le d. \tag{3.33}$$

Equations (3.26), extended to d dimensions, can be used to show the asymptotic behavior for large $|\mathbf{j}|$,

$$\psi_1^*(\mathbf{j}) \sim |j|^{\sigma-d} \quad (0 \le \sigma \le d). \tag{3.34}$$

We can extend (3.21b) to d dimensions and obtain

$$\Omega_1^*(\mathbf{k}) = \sum_{m_1=-\infty}^{\infty}\cdots\sum_{m_d=-\infty}^{\infty}\Phi(\mathbf{k}\cdot\mathbf{e}_1+m_1, \ldots, \mathbf{k}\cdot\mathbf{e}_d+m_d)$$
$$\times\prod_{\alpha=1}^{d}[1-\cos(2\pi\mathbf{k}\cdot\mathbf{e}_\alpha)], \tag{3.35}$$

where $\Phi \ge 0$ is a homogeneous function of degree $-(\sigma+2d)$ and Φ is even in each argument separately. This form shows the simplification of the infinite degeneracy of the eigenvectors resulting for a translation invariant ferromagnet with lattice hyperplane symmetry. It is interesting to note that the additional condition of asymptotic rotational invariance, i.e., $\Omega_1^*(\mathbf{k})$ depends only on $|\mathbf{k}|$ as $|\mathbf{k}|\to 0$, uniquely specifies Φ up to a normalization constant. The corresponding fixed point function is

$$\Omega_1^*(\mathbf{k}) =$$
$$A\sum_{m_1=-\infty}^{\infty}\cdots\sum_{m_d=-\infty}^{\infty}\frac{\prod_{\alpha=1}^{d}[1-\cos(2\pi\mathbf{k}\cdot\mathbf{e}_\alpha)]}{[\prod_{\alpha=1}^{d}(\mathbf{k}\cdot\mathbf{e}_\alpha+m_\alpha)^2][\sum_{\alpha=1}^{d}(\mathbf{k}\cdot\mathbf{e}_\alpha+m_\alpha)^2]^{\sigma/2}}. \tag{3.36}$$

We now turn to the analysis of the higher-order, Ursell-function portion of our correlation-function state-vector. From the structure of the equations so far developed, it is clear that the renormalization group formalism differs very little for extension to higher order Ursell functions from that for higher dimensions. The fixed point equation (3.30) generalizes directly to

$$\lambda_N t_m^*(\mathbf{j}_1, \ldots, \mathbf{j}_m) = (q_N/N^{d/2})^{m+1} t_m^*(N\mathbf{j}_1, \ldots, N\mathbf{j}_m). \tag{3.37}$$

This result can be seen most easily from the extension of (3.6) and (3.7) to higher dimensions, and (2.7). The reexpression of the d-dimensional version of (3.7) in terms of coordinate space yields a transformation analogous to (3.26).

Following analogous analysis to that given previously, we can deduce that the eigenvector solutions to (3.37) behave asymptotically as $\lambda\to\infty$ as

$$\psi_m^*(\lambda\mathbf{j}_1, \ldots, \lambda\mathbf{j}_m) \propto \lambda^{\omega^*-(m+1)(d-\sigma+\omega^*)/2}, \tag{3.38}$$

where we have used the definition of ω^*

$$-\tfrac{1}{2}d \le \tfrac{1}{2}(\omega^*-\sigma) \equiv \ln q_2/\ln 2 = \ln q_N/\ln N. \tag{3.39}$$

We have used assumption (4) of asymptotic monotonicity along rays to force the last equality in (3.39). This monotonicity implies the u's are of fixed sign which in turn implies that the t's are monotonic. The first inequality follows because monotonicity is preserved on iteration of (3.2), and if the inequality were violated, we would have for some m sufficiently large and sufficiently large λ, an implied violation of assumption (2), finite-spin magnitude. Equation (3.38) follows directly from (3.37) where λ_N and q_N are written as explicit functions of N by use of (3.33) and (3.39), and N is relabeled λ.

The ω^* so defined by (3.39) is the anomalous dimension of the vacuum. In the special case

$$\omega^* = 0 \tag{3.40}$$

we have

$$q_N = N^{-\sigma/2}, \quad \lambda_N = 1 \tag{3.41}$$

and (3.38) reduces to

$$\psi_m^*(\lambda\mathbf{j}_1, \ldots, \lambda\mathbf{j}_m) \propto \lambda^{-(m+1)(d-\sigma)/2} \tag{3.42}$$

in accord with the results of Wilson.[2] The structure of the renormalization group, as we have developed it, does not require the result $\omega^*=0$. It may happen, for example, that the Ursell functions, rather than being proportional to powers of ψ_1^*, exhibit additional anomalous cancellation and so require (3.38) rather than (3.42).

We may directly extend (3.35) to higher order Ursell functions. Thus

$$\Omega_m^*(\mathbf{k}_1, \ldots, \mathbf{k}_m) = \sum_{\mathbf{n}_1=-\infty}^{+\infty}\cdots\sum_{\mathbf{n}_m=-\infty}^{+\infty}\Phi_m(\mathbf{k}_1+\mathbf{n}_1, \ldots, \mathbf{k}_m+\mathbf{n}_m)$$
$$\times\prod_{\alpha=1}^{d}\left(\sin(\pi\mathbf{k}_m\cdot\mathbf{e}_\alpha)\prod_{i=1}^{m}\sin[\pi(\mathbf{k}_i-\mathbf{k}_{i-1})\ \mathbf{e}_\alpha]\right), \tag{3.43}$$

where Φ_m is homogeneous of degree $-\sigma-\frac{1}{2}(m-1)(d-\omega^*+\sigma)-(m+1)d$ and $\mathbf{k}_0=0$. The following scaling behavior follows:

$$\Omega_m^*(q\mathbf{k}_1, \ldots, q\mathbf{k}_m) \sim q^{-\sigma-(m-1)(d-\omega^*+\sigma)/2} \tag{3.44}$$

The conclusion of this section is that for a translationally invariant, ferromagnetic lattice system with symmetry about the lattice hyperplanes there exists

fixed points of the "mean-spin" renormalization group $N\{R_N, N=2, 3, \cdots\}$ defined by

$$s_j = N^{(\omega^* - d - \sigma)/2} \sum_{k=0}^{N-1} \sigma_{Nj+k} \tag{3.45}$$

subject to (3.39). The fixed points have the eigenvalues

$$\lambda_N = N^{\omega^*}. \tag{3.46}$$

The fixed points are degenerate to the extent represented by the set of homogeneous function Φ_m of (3.43). The physical conditions imposed amount to an asymptotic monotonicity of the Φ_m functions and serve to substantially limit the degeneracy of each fixed point to a "smooth" function.

It is to be noted that the nature of the fixed points depends on the asymptotic behavior of the original correlation vector. The "block-average" asymptotic behavior of ψ_m must be majorized by form (3.38). For those Ursell functions for which the bound is not obtained, in the sense that the exponent in form (3.38) is wrong, the corresponding amplitude for ψ_m^* would vanish. In fact, $\psi_m^* \equiv 0$, $m > 1$, for the Gaussian model by explicit calculation. As expected, the nature of the physical fixed points of the renormalization group is controlled by the long-range asymptotic behavior of the correlation vector, or, put more abstractly, by the definition of the space on which the renormalization group operators act.

4. ANOMALOUS DIMENSION OF THE VACUUM AND THE LIMITING SPIN FIELD

In this section we will discuss the possible values of the anomalous dimension of the vacuum ω^*, introduced in Eq. (3.39), and the effect on the limiting Hamiltonian associated with the fixed point of renormalization group. From (3.39) we already have the restriction

$$-(d-\sigma) \leq \omega^*. \tag{4.1}$$

For ferromagnetic Ising models, we will now improve (5.1) to

$$0 \leq \omega^*. \tag{4.2}$$

To do so, we make use of Newman's inequalities to show a contradiction to the assumption $\omega^* < 0$. Newman's inequalities hold for variables that can be represented as a finite or infinite sum of spin-1/2 Ising spins with ferromagnetic coupling. This class includes the higher spin Ising models and continuous-spin, Ising models with a spin distribution proportional to $\exp(-as^2 - s^4)$. The inequalities are (mean spin equals zero for this class of systems):

$$\langle \sigma_{i_1} \cdots \sigma_{i_{2n}} \rangle \leq \sum_{\pi} \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2n-1}} \sigma_{\pi_{2n}} \rangle, \tag{4.3}$$

where the sum over π extends over all $(2n)!/2^n n!$ partitions of the indices $i_1, \ldots, i_{2n}$ into pairs. Alternately, Newman also shows the inequality where the summation is over any family of partitions that has all pair partitions as a refinement. For example, for $n=2$, the Newman inequalities (4.3) become the GHS inequalities[18]

$$\langle \sigma_1 \sigma_2 \sigma_3 \sigma_4 \rangle \leq \langle \sigma_1 \sigma_2 \rangle \langle \sigma_3 \sigma_4 \rangle + \langle \sigma_1 \sigma_3 \rangle \langle \sigma_2 \sigma_4 \rangle + \langle \sigma_1 \sigma_4 \rangle \langle \sigma_2 \sigma_3 \rangle. \tag{4.4}$$

Now if we select, by looking at (3.2) for $j=0$,

$$V_m = \sum_{r_1=0}^{N-1} \cdots \sum_{r_{2m}=0}^{N-1} u_{2m}(r_1, \ldots, r_{2m}), \tag{4.5}$$

we may compute from (3.38) the asymptotic behavior for $N \to \infty$

$$V_m = O(N^{m(d+\sigma-\omega^*)+\omega^*}). \tag{4.6}$$

If $\omega^* < 0$, then the exponent of N in (4.6) is necessarily positive. From the definition (1.15) of the Ursell functions it follows that

$$\langle \sigma_1 \cdots \sigma_{2m} \rangle = \sum_{\pi} \prod_{\tau \in \pi} u(\tau), \tag{4.7}$$

where the summation is over all partitions of $1, \ldots, 2m$. For example, when $\langle \sigma_i^{2j+1} \rangle = 0$, we have, for $m=2$ and 1,

$$\begin{aligned} \langle \sigma_1 \sigma_2 \sigma_3 \sigma_4 \rangle &= u_4(1,2,3,4) + u_2(1,2)u_2(3,4) \\ &\quad + u_2(1,3)u_2(2,4) + u_2(1,4)u_2(2,3), \\ \langle \sigma_1 \sigma_2 \rangle &= u_2(1,2). \end{aligned} \tag{4.8}$$

If we apply (4.4) to (4.8) we can deduce by (4.5) and Griffiths' inequality[10] that

$$0 \leq V_4 + 3V_2^2 \leq 3V_2^2 \quad \text{or} \quad |V_4| \leq 3V_2^2. \tag{4.9}$$

Thus, by (4.6) we must have, for $\omega^* < 0$,

$$O(N^{2d+2\sigma-\omega^*}) \leq O(N^{2d+2\sigma}), \tag{4.10}$$

which is a contradiction. Thus (4.2) holds. A similar result has been obtained in the context of Boson field theories by Glimm and Jaffee[19] for $d \leq 4$.

Let us suppose that under the block spin transformation (3.45), there is convergence to a fixed point with $\omega^* > 0$. Instead of using the transformation (3.45), let us define

$$\Sigma_j = N^{-(d+\sigma)/2} \sum_{r=0}^{N-1} \sigma_{Nj+r}. \tag{4.11}$$

Then it follows that

$$\langle \Sigma_{j_1} \cdots \Sigma_{j_{2m}} \rangle_c = O(N^{-(m-1)\omega^*}), \tag{4.12}$$

where the subscript c means cumulant, or Ursell function. This result shows that all the higher order Ursell functions converge to zero under (4.11) as $N \to \infty$. (See also Kawasaki and Gunton[20]). This behavior is characteristic of a Gaussian spin-distribution and the Hamiltonian corresponding to the fixed point is

$$H^* = \tfrac{1}{2} \int_0^1 d\mathbf{p} [\Omega_1^*(\mathbf{p})]^{-1} \sigma_{\mathbf{p}} \sigma_{-\mathbf{p}}, \tag{4.13}$$

where Ω_1^* is given by (3.35) and $\sigma_{\mathbf{p}}$ is the usual momentum space version of σ_j. Thus, when $\omega^* > 0$, there is convergence to a Gaussian fixed point under (4.11). As shown later in Sec. 7, this convergence to a Gaussian fixed point does *not* necessarily imply Gaussian exponents. For example, it does not follow that $\gamma = 1$. When $\omega^* = 0$, then there is no rescaling which leads to a Gaussian fixed point.

We remark that if the three-dimensional spin-$\frac{1}{2}$ Ising series results,[6] $\omega^* \approx 0.06$, are correct, then the relevant fixed point would be of Gaussian type. The argu-

ments of field theoretic origin suggesting that $\omega^* = 0$ implicitly assume the existence of a nontrivial field theory, i.e., one in which the scattering amplitude does not vanish. Except in two dimensions, this assumption is not proven.

5. APPROACH TO FIXED POINTS

Before we can discuss the approach of the correlation vectors to a fixed point under the action of the renormalization group, it is necessary to introduce a norm. In order to introduce a sensible norm, it is desirable to first consider the properties of these vectors. According to Duneau, *et al.*,[21] it is physically reasonable to assume a strong cluster property of the form,

$$|\psi_m(\mathbf{j}_1, \ldots, \mathbf{j}_m)| \leq \sum_{T} \prod_{l \in T} w(d_l), \tag{5.1}$$

where the sum runs over all connected trees (graphs without closed loops) of m lines joining the $m+1$ points, $0, \mathbf{j}_1, \mathbf{j}_1+\mathbf{j}_2, \ldots, \mathbf{j}_1+\mathbf{j}_2+\cdots+\mathbf{j}_m$. The product is over the lines l of the tree with d_l the length of the line l. The function $w(x)$ is a positive function which goes to zero at infinity and from (5.1) with $m=1$,

$$|\psi_1(\mathbf{j})| \leq w(|\mathbf{j}|). \tag{5.2}$$

For our purposes it is convenient to note that a weaker form of (5.1) is

$$|\psi_m(j_1, \ldots, j_m)| \leq (m+1)^{m-1} \prod_{l \in T^*} w(d_l), \tag{5.3}$$

where T^* is the tree that gives the maximum term in (5.1), and the coefficient is Cayley's formula[22] for the number of labeled trees of $m+1$ vertices.

For translationally invariant ferromagnetic lattice systems Duneau, *et al.*[21] have *proven* a stronger result than (5.1) at points sufficiently far in the H, T plane from the critical point and the phase boundary. Their results are valid both for short range potentials and potentials which fall off like r^{-s}, $s > d$. In the case of the long range potentials, they prove inequality (5.1) using

$$w(x) = C(1+\alpha|x|)^{-(s-d-\epsilon)/2}, \tag{5.4}$$

where $\alpha, \epsilon > 0$ are arbitrary, but C is a constant depending on T, H, α, ϵ, and the interactions.

If for a particular case we have, in a notation consonant with Sec. 3,

$$w(x) \sim |x|^{-(d-\sigma-\omega^*)/2} \quad (x \to \infty), \tag{5.5}$$

then it would follow from (5.3) that

$$\int |\psi_m(r_1, \ldots, r_m)|^p \prod_{i=1}^{m} dr_i < \infty, \tag{5.6}$$

provided that

$$p > 2d/(d-\sigma-\omega^*). \tag{5.7}$$

With this introduction, we are prepared to introduce a norm on the space of correlation functions tailored just to include a preselected largest eigenvalue of the renormalization group. We define

$$\|\psi\| = \left[\sum_{m=1}^{\infty} N_m \left\{\sum_{\text{lattice}} |\psi_m|^{p_m}\right\}^{(1+\delta)/p_m}\right]^{1/(1+\delta)}, \tag{5.8}$$

where

$$p_m = \frac{2d}{(d-\sigma+\omega^*) + (d-\sigma-\omega^*-2\Theta)/m},$$
$$N_m = (m+1)^{-(1+\delta+\epsilon)(m-1)}, \quad \Theta, \delta, \epsilon > 0, \tag{5.9}$$

insures that the eigenvector corresponding to (3.44)–(3.45) has finite length. We only define (5.8) when we can select $\delta > 0$ such that $p_m > 1+2\delta$, which limits $\omega^* < d+\sigma$. It then follows that the space $\mathcal{B}_\Theta$ of ψ with $\|\psi\| < \infty$ is a separable, reflexive Banach space. In a space of this type there exist theorems on the approach of iterates of operators to a fixed point.[13]

Let us introduce the operator

$$T_N = N^{-\omega^*} R_N, \tag{5.10}$$

where R_N is the mean-spin renormalization group operator defined by (3.45). From (3.46), we see that the operator T_N has the eigenvectors defined in Sec. 3 as invariant elements, $T_N f^* = f^*$. We will now show that the operator $N^{-\Theta}T_N$ is a *contraction* on $\mathcal{B}_\Theta$. By a contraction we mean that, for all f in $\mathcal{B}_\Theta$,

$$\|N^{-\Theta}T_N f\| \leq \|f\|. \tag{5.11}$$

To prove (5.11), we use the mean value inequality,[23]

$$M^{-1}\sum_{j=1}^{M} |x_j|^p \geq \left[M^{-1}\sum_{j=1}^{M} |x_j|\right]^p, \quad p \geq 1. \tag{5.12}$$

This inequality is strict unless all the x_j are equal. Suppose we apply R_N to the ψ_m subspace. Then we compose the result as a sum of $N^{(m+1)d}$ terms. These terms are just a regrouping of the original ψ_m's except that there are N^d repetitions of each original term due to translational invariance. If we apply (5.12) to each element of $N^{-\Theta}T_N\psi$ using the appropriate p in each case, we find from (3.45) that we obtain (5.11).

Let us illustrate this argument by considering the action of T_2 on ψ_1 in $d=1$. Defining $\psi_1' = T_2\psi_1$, we find

$$\sum_r \psi_1'(r)^p = 2^{(1-\sigma)p}\sum_r [\tfrac{1}{2}\psi_1(2r) + \tfrac{1}{4}\psi_1(2r-1) + \tfrac{1}{4}\psi_1(2r+1)]^p$$
$$\leq 2^{(1-\sigma)p}\sum_r [\tfrac{1}{2}\psi_1(2r)^p + \tfrac{1}{4}\psi_1(2r-1)^p + \tfrac{1}{4}\psi_1(2r+1)^p].$$

Since $p = 1/(1-\sigma-\Theta)$, this becomes

$$\left[\sum_r \psi_1'(r)^p\right]^{1/p} \leq 2^{\Theta}\left[\sum_r \psi_1(r)^p\right]^{1/p}.$$

At first sight, one might think we should let $\Theta \to 0$ and establish T_N itself as a contraction on $\mathcal{B}_0$. Unfortunately, the fixed point vectors ψ^*, discussed in Sec. 3, do not belong to $\mathcal{B}_0$. On the other hand, one might imagine we could enlarge $\mathcal{B}_0$ to include ψ^* in a manner such that T_N remains a contraction. However, it will be shown that it is not possible to define a reflexive Banach space containing nonzero fixed points of T_N, on which T_N is a contraction.

Lorch[13] gives the following relevant theorem on the approach to fixed points. We quote only a special case directly related to our problem.

Theorem: Let $N^{-\Theta}T_N$ be a contraction on $\mathcal{B}_\Theta$, then the sequence $\{U_n\}$

$$U_n=\frac{1}{n}\sum_{j=0}^{n-1}N^{-\Theta j}T_N^j \tag{5.13}$$

converges strongly to a limiting transformation P, $\|P\| \leq 1$, which is a projection. The relation $P\mathbf{f}=\mathbf{f}$ holds for precisely those elements for which $N^{-\Theta}T_N\mathbf{f}=\mathbf{f}$. The relation $P\mathbf{f}=0$ holds precisely for the elements of the closed linear manifold spanned by elements of the form $N^{-\Theta}T_N\mathbf{g}-\mathbf{g}$, where $\mathbf{g}$ arbitrary in $\mathcal{B}_\Theta$.

Since Lorch also shows that the Banach space $\mathcal{B}_\Theta$ is the direct sum of the closed linear manifold of f's such that $N^{-\Theta}T_N\mathbf{f}=\mathbf{f}$ and the disjoint (except for zero) closed linear manifold of f's such that $\mathbf{f}=N^{-\Theta}T_N\mathbf{g}-\mathbf{g}$ and their limits, the theorem describes the complete action of P on $\mathcal{B}_\Theta$. Unfortunately, since the fixed point vectors, discussed in Sec. 3, satisfy $T_N\psi^*=\psi^*\neq N^{\Theta}\psi^*$, it follows that $P\psi^*=0$, so although we have proved convergence to a fixed point, it is the trivial one.

Let us present a counterexample which shows that general convergence theorems, like that of Lorch, cannot be expected to apply to the renormalization group operator. Consider R_2 for the mean-spin transformation acting on the two-point function in one dimension. Let

$$\psi_1(r)=b_{l+1}/2^{(1-\sigma)(l+1)},\quad l=[\log_2(r)], \tag{5.14}$$

where $[x]$ is the greatest integer less than or equal to x for $x\geq 0$ and -1 otherwise. We compute only

$$\psi_1^{(L)}(0)=(1+2^{-L-1})u^{(L)}-\tfrac{3}{2}v^{(L)} \tag{5.15a}$$

where

$$\begin{aligned} u^{(L)}&=b_L+2^{-\sigma}u^{(L-1)},\\ v^{(L)}&=\tfrac{1}{2}b_L+2^{-(1+\sigma)}v^{(L-1)}, \end{aligned} \tag{5.15b}$$

and $\psi_1^{(L)}$ indicates T_2 has been applied L times. For ψ to go to a limit, it is necessary that $\psi_1^{(L)}(0)$ goes to a limit as $L\to\infty$. However, if we select b_l randomly in the interval

$$2^{\sigma-1}\leq b_l\leq 1, \tag{5.16}$$

then, since by (5.15) $\psi_1^{(L)}(0)\approx\frac{1}{4}b_L+a_L$ where a_L is independent of b_L, it is evident that $\psi_1^{(L)}(0)$ does not tend to a limit. If we select b_l in the upper quarter or lower quarter of the range (5.16) as $[\log_2(l)]$ is even or odd, then even the $\Theta=0$ version of (5.13),

$$\frac{1}{n}\sum_{l=0}^{n-1}\psi_1^{(l)(0)}, \tag{5.17}$$

does not tend to a limit.

We point out that although the behavior of the correlation function (5.14) is not as usually expected, it is monotonically decreasing, and we can construct a Gaussian model exhibiting that behavior. Integrating formally over the spins, we obtain the partition function

$$Z\propto\exp\left[\tfrac{1}{2}\sum_{ij}h_i\psi_1(i-j)h_j\right], \tag{5.18}$$

where the h_l are site-dependent magnetic fields, and $\langle\sigma_i\sigma_j\rangle=\psi_1(i-j)=(\partial^2/\partial h_i\partial h_j)\ln Z|_{h=0}$.

An additional model which can be readily shown not to tend to a limit under the renormalization group transformation is the hierarchical model[14] with random bond strengths. We select as a Hamiltonian

$$H=J\sum_{l=0}^{L-1}b_l 2^{-l\sigma}\sum_{m=1}^{2^{L-1-l}}(s_{m,l})^2, \tag{5.19}$$

where

$$\left.\begin{aligned} s_{m,l+1}&=(1/\sqrt{2})(\hat{s}_{2m-1,l}-\hat{s}_{2m,l})\\ \hat{s}_{m,l+1}&=(1/\sqrt{2})(\hat{s}_{2m-1,l}+\hat{s}_{2m,l}) \end{aligned}\right\},\quad m=1,\ldots,2^{L-1} \tag{5.20}$$

with $\hat{s}_{m,-1}$ the fundamental spins. The factors b_l are taken to be random on $0<b_l\leq 1$. Following the analysis of Baker,[14] we derive the recursion relations

$$\begin{aligned} I_l(x)&=\int_{-\infty}^{\infty}dy\,\exp[-Kb_l y^2-\tfrac{1}{2}Q_l(x+y)-\tfrac{1}{2}Q_l(x-y)],\\ Q_{l+1}(x)&=-2\ln[I_l(2^{(\sigma-1)/2}x)/I_l(0)]. \end{aligned} \tag{5.21}$$

The Q_l do not approach a limit as long as the b_l vary randomly.

Let us consider the following simple example. Define

$$\begin{aligned} \phi_m(r)&=\psi_1^*(r),\quad r\geq 2^M,\\ &=0,\qquad r<2^M, \end{aligned} \tag{5.22}$$

where $\psi_1^*(r)$ is a totally positive fixed point of T_2. Then by construction

$$\begin{aligned} &(T_2\phi_M)(r)=\psi_1^*(r),\quad r\geq 2^{M-1}+1,\\ &(T_2\phi_M)(2^{M-1})>0,\\ &(T_2\phi_M)(r)=0,\quad r<2^{M-1}. \end{aligned} \tag{5.23}$$

Since $\psi_1^*(r)>0$, τ_M (defined below) will be positive and "orthogonal" to ϕ_M. One then expects the inequality

$$\|T_2\phi_M\|=\|\phi_M+\tau_M\|>\|\phi_M\| \tag{5.24}$$

for any sensible norm of a Banach space which includes ψ^* and these amputations of ψ^*. Therefore, T_2 is *not* a contraction under these circumstances.

In order to insure the convergence of $T_N^n\psi$ to a fixed point, we must have more restrictive hypotheses on the initial vector ψ. In particular, if we can expand ψ about infinite separation in terms of pure or fractional powers of $1/r$, then there is a gap in the eigenspectral resolution of ψ and there will be geometric convergence of $T_N^n\psi$ to the fixed point as $n\to\infty$. The example of two identical interpenetrating self-interacting sublattices A and B, where $\langle\sigma_\alpha\sigma_\beta\rangle=0$, $\alpha\in A$ and $\beta\in B$, can converge to a fixed point, and shows that the expansion requirement just stated is stronger than necessary. If we can only resolve ψ in terms of a continuous spectral representation, as, for example,

$$\psi_1(r)=\int_{d-\sigma}^{\infty}\frac{ds}{r^s}=\frac{1}{r^{d-\sigma}\ln r}, \tag{5.25}$$

direct calculation (using the decimation transformation for ease of exposition) yields

$$\psi_1^{(n)}(r)=T_N^n\psi_1(r)=1/r^{d-\sigma}(n\ln N+\ln r), \tag{5.26}$$

which tends to zero in the sense of the norm (5.8). We can rescale $\psi_1^{(n)}$ as

$$(n\ln N)\psi_1^{(n)}\to 1/r^{d-\sigma}+O(1/n). \tag{5.27}$$

There is now convergence, but it is no longer geometric.

We remark that in the case of the hierarchical model, where R_2 can be applied exactly, the usual expansion analysis of the spin-weight function, together with the smoothing effects of the integration procedure, shows that the only relevant part of the spectrum is a discrete subset. Therefore, there is a gap in the spectrum, and geometric convergence is to be expected. Numerical work confirms the geometric convergence, and shows it to be quite rapid in practice.

6. A CENTRAL LIMIT THEOREM FOR HIGH TEMPERATURES

In the previous sections we have considered quite general situations and obtained correspondingly general results. In this section we will show that when the mean-spin transformation is repeatedly applied to a ferromagnet above its critical temperature, where the susceptibility is finite and the spontaneous magnetization is zero, that the "infinite temperature fixed point" is approached. The limiting infinite temperature fixed point Hamiltonian consists of a lattice of independent, normally distributed spins. For ease of exposition we will consider only the one-dimensional case, but the corresponding results can be obtained in higher dimension by direct extension of our discussion. Let a one-dimensional Ising model be described by the Hamiltonian

$$H = -\tfrac{1}{2}\sum_{i,j} J(i-j)\sigma_i\sigma_j, \tag{6.1}$$

with purely ferromagnetic interactions $J(i-j) \geq 0$. The interactions need not be explicitly specified, but we assume that they are chosen such that the thermodynamic limit exists, and that in this limit the correlation functions are translationally invariant. The spins may be discrete or continuous, for in what follows we require only that the GKS [10] and Newman[11] inequalities are applicable. If the potential behaves for $r \to \infty$ like $J(r) \sim r^{-1-\sigma}$ $(0 < \sigma \leq 1)$, then the system will have long-range order at sufficiently low temperatures.[15]

For $T > T_c$, let us define a sequence of spin distributions as follows. For each integer $N \geq 1$, we define the moments

$$\mu_p(N) = \left\langle \left(\frac{1}{N^{1/2}} \sum_{j=1}^{N} \sigma_j \right)^p \right\rangle, \quad p = 1, 2, 3, \cdots. \tag{6.2}$$

The brackets $\langle\ \rangle$ indicate the expectation value after the thermodynamic limit has been taken. The spin distribution function $f_N(z) \geq 0$ is defined as the substantially unique nonnegative solution of the Hamburger moment problem[24]

$$\int_{-\infty}^{\infty} f_N(z)\, dz = 1, \tag{6.3a}$$

$$\int_{-\infty}^{\infty} z^p f_N(z)\, dz = \mu_p(N), \quad p = 1, 2, 3, \cdots. \tag{6.3b}$$

This problem is determinate since for any integers $0 < A < B$,

$$\sum_{s=A}^{B}\sum_{t=A}^{B} \mu_{s+t}(N)\, y_s y_t = \left\langle \left[\sum_{s=A}^{B} \left(\frac{1}{N^{1/2}} \sum_{j=1}^{N} \sigma_j \right)^s y_s \right]^2 \right\rangle \geq 0, \tag{6.4a}$$

and we will demonstrate below that the Carleman condition for two-sided distributions

$$\sum_{r} [\mu_p(N)]^{-1/p} = \infty \tag{6.4b}$$

is satisfied. Equations (6.4a), (6.4b) show the solution to (6.3) exists and is substantially unique. If we write

$$f_N(z) = \exp[-Q_N(z)], \tag{6.5}$$

then the $Q_N(z)$ with $N = 2^l$ are the appropriate analogue for the Ising model of the $Q_l(z)$ introduced by Wilson[1] in his approximate renormalization group recursion relations and by Baker[14] in his exact renormalization group treatment of the hierarchical model. The recursion relations considered by Wilson and by Baker have the property that for $T > T_c$, the spin distributions $\exp[-Q_l(z)]$ approach a Gaussian as $l \to \infty$. That is, under repeated application of the renormalization group they approach an "infinite temperature fixed point." This behavior is a generalization of the central limit theorem. We might mention that a discussion of the behavior of spin distributions for the two-dimensional Ising model has been given by McCoy and Wu[25] and that Jona-Lasinio[16] has recently stressed the relationship of Wilson's[1] ideas to the central limit theorems of probability theory.

We shall now use the GKS [10] and Newman[11] inequalities to prove that the spin distribution functions $f_N(z)$ defined in (6.3) approach a Gaussian limit as $N \to \infty$, when $T > T_c$. Specifically, we prove that the *finiteness* of the limit

$$\lim_{N\to\infty} \sum_{j=0}^{N} \langle \sigma_0 \sigma_j \rangle = A/2, \tag{6.6}$$

implies that, for all $z_1 < z_2$,

$$\lim_{N\to\infty} \int_{z_1}^{z_2} dz\, f_N(z) = \frac{1}{(2\pi\chi)^{1/2}} \int_{z_1}^{z_2} dz \exp(-z^2/2\chi), \tag{6.7}$$

where $f_N(z)$ is defined by (6.3) and

$$\chi = \lim_{N\to\infty} \mu_2(N) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N}\sum_{j=1}^{N} \langle \sigma_i \sigma_j \rangle. \tag{6.8}$$

That χ is finite follows from the finiteness of the limit in (6.6). To see this result, we use Griffiths' inequality, $\langle \sigma_i \sigma_j \rangle \geq 0$, together with

$$\sum_{i=0}^{N} \langle \sigma_0 \sigma_i \rangle \leq A/2, \tag{6.9}$$

and as by (3.10)

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} \langle \sigma_i \sigma_j \rangle = \langle \sigma_0 \sigma_0 \rangle + 2\sum_{j=1}^{N} \left(1 - \frac{j}{N}\right) \langle \sigma_0 \sigma_j \rangle \leq A. \tag{6.10}$$

The proof of (6.7) proceeds in two steps. We first show that the moments $\mu_p(N)$ defined in (6.2) converge to the Gaussian values

$$\lim_{N\to\infty} \mu_{2p}(N) = \frac{(2p)!}{2^p p!} \chi^p, \quad p = 1, 2, 3, \cdots, \tag{6.11}$$

and that all odd moments are zero. Next we use the Tchebycheff inequalities[24] to show that the convergence of the moments in (6.11) implies the convergence of the distribution stated in (6.7).

To prove that the moments converge to their Gaussian values, we use the GKS inequalities[10]

$$\langle \sigma_{i_1} \cdots \sigma_{i_n} \rangle \geq 0, \tag{6.12a}$$

$$\langle \sigma_{i_1} \cdots \sigma_{i_m} \sigma_{i_{m+1}} \cdots \sigma_{i_n} \rangle \geq \langle \sigma_{i_1} \cdots \sigma_{i_m} \rangle \langle \sigma_{i_{m+1}} \cdots \sigma_{i_n} \rangle, \tag{6.12b}$$

and the Newman inequalities[11] (4.3), i.e.,

$$\langle \sigma_{i_1} \cdots \sigma_{i_{2n}} \rangle \leq \sum_{\pi} \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2n-1}} \sigma_{\pi_{2n}} \rangle, \tag{6.13}$$

where the sum over π extends over all $(2n)!$ partitions of the indices $i_1, \ldots, i_{2n}$. We obtain

$$\begin{aligned} N^{-p} \sum_{i_1=1}^{N} \cdots \sum_{i_{2p}=1}^{N} & \max_{\pi} \{ \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2p-1}} \sigma_{\pi_{2p}} \rangle \} \\ & \leq \mu_{2p}(N) \\ & \leq N^{-p} \sum_{i_1=1}^{N} \cdots \sum_{i_{2p}=1}^{N} \left[\sum_{\pi} \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2n-1}} \sigma_{\pi_{2n}} \rangle \right] \\ & = [(2p)!/2^p p!] \mu_2(N)^p. \end{aligned} \tag{6.14}$$

The lower bound follows from the GKS inequality (6.12b) while the upper bound follows from Newman's inequality (6.13). We shall show that when $N \to \infty$, the difference between the upper and lower bounds in (6.14) vanishes, thus establishing (6.11).

The convergence of the sum in (6.6) implies that given any $\epsilon > 0$, there exists integer N^* such that

$$\sum_{j=N^*}^{\infty} \langle \sigma_0 \sigma_j \rangle < \epsilon. \tag{6.15}$$

We will show that

$$\begin{aligned} 0 \leq N^{-p} \sum_{i_1=1}^{N} \cdots \sum_{i_{2p}=1}^{N} & \left[\sum_{\pi} (\langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2p-1}} \sigma_{\pi_{2p}} \rangle) \right. \\ & \left. - \max_{\pi} \{ \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2p-1}} \sigma_{\pi_{2p}} \rangle \} \right] \\ & \leq \frac{(2p)!}{2^p p!} \left[(2p) A^{p-1} \epsilon + 2p(p-1) A^p \frac{2N^*}{N} \right], \end{aligned} \tag{6.16}$$

where A is defined in (6.9). Since the right-hand side can be made arbitrarily small by first choosing $\epsilon > 0$, then an appropriate N^*, and finally choosing N sufficiently large, this proves the convergence of the moments given in (6.11).

In order to establish (6.16), we note that (6.15) implies

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ |i-j| \geq N^*}}^{N} \langle \sigma_i \sigma_j \rangle \leq \epsilon, \tag{6.17}$$

and (6.9) and (6.10) imply

$$\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{\substack{l=1 \\ |j-k| < N^*}}^{N} \langle \sigma_i \sigma_j \rangle \langle \sigma_k \sigma_l \rangle \leq A^2 \frac{2N^*}{N}. \tag{6.18}$$

Let us define $R = \{i_1, \ldots, i_{2p} = 1, \ldots, N\}$. We then write $R = R_1 \cup R_2 \cup R_3$, where

R_1 = subset of R with less than p pairs of indices satisfying $|i-j| < N^*$,

R_2 = subset of R with exactly p pairs of indices satisfying $|i-j| < N^*$,

R_3 = subset of R with more than p pairs of indices satisfying $|i-j| < N^*$.

Let us denote by $I(R)$ the sum in (6.16) which we wish to bound from above. We note that $I(R) = I(R_1) + I(R_2) + I(R_3)$, where $I(R_j)$ is the sum restricted to R_j $(j = 1, 2, 3)$.

We first consider $I(R_1)$ and note that it is not decreased if we do not subtract off the maximum term,

$$I(R_1) \leq \sum_{R_1} \left[\sum_{\pi} \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2p-1}} \sigma_{\pi_{2p}} \rangle \right]. \tag{6.19}$$

The first term in the sum over partitions is $\langle \sigma_{i_1} \sigma_{i_2} \rangle \cdots \langle \sigma_{i_{2p-1}} \sigma_{i_{2p}} \rangle$. There must be at least one correlation function $\langle \sigma_{i_{2a-1}} \sigma_{i_{2a}} \rangle$ in this product with $|i_{2a-1} - i_{2a}| \geq N^*$. The sum over R_1 of all terms in which $|i_1 - i_2| \geq N^*$ is bounded by $A^{p-1}\epsilon$, where we use (6.10) and (6.17). The sum over R_1 of all terms in which $|i_3 - i_4| \geq N^*$ is again bounded by $A^{p-1}\epsilon$. Hence, we bound the sum of the first term by $pA^{p-1}\epsilon$. Repeating this analysis for the other $(2p)!/2^p p!$ terms in the sum over partitions, we obtain

$$I(R_1) \leq \frac{(2p)!}{2^p p!} p A^{p-1} \epsilon. \tag{6.20}$$

Now consider

$$\begin{aligned} I(R_2) = \sum_{R_2} \Big[& \sum_{\pi} \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2p-1}} \sigma_{\pi_{2p}} \rangle \\ & - \max_{\pi} \{ \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2p-1}} \sigma_{\pi_{2p}} \rangle \} \Big]. \end{aligned}$$

In the sum over partitions there is exactly one term with

$$|i_{\pi_1} - i_{\pi_2}| < N^*, \ldots, |i_{\pi_{2p-1}} - i_{\pi_{2p}}| < N^*.$$

We obtain something larger by replacing the $\max_{\pi}$ by this term, thus removing it. In all the remaining terms at least one pair of indices satisfies $|i_{\pi_{2a-1}} - i_{\pi_{2a}}| \geq N^*$. The argument used to bound $I(R_1)$ now implies

$$I(R_2) \leq [(2p)!/2^p p!] p A^{p-1} \epsilon. \tag{6.21}$$

By not subtracting off the maximum term, we see that

$$I(R_3) \leq \sum_{R_3} \left[\sum_{\pi} \langle \sigma_{\pi_1} \sigma_{\pi_2} \rangle \cdots \langle \sigma_{\pi_{2p-1}} \sigma_{\pi_{2p}} \rangle \right].$$

Consider the first term $\langle \sigma_{i_1} \sigma_{i_2} \rangle \cdots \langle \sigma_{i_{2p-1}} \sigma_{i_{2p}} \rangle$ in the sum over partitions. There must exist at least two terms in the product $\cdots \langle \sigma_i \sigma_j \rangle \langle \sigma_k \sigma_l \rangle \cdots$ such that $|i-k| < N^*$, $|i-l| < N^*$, $|j-l| < N^*$, or $|j-k| < N^*$. First consider all terms in which $\langle \sigma_{i_1} \sigma_{i_2} \rangle \langle \sigma_{i_3} \sigma_{i_4} \rangle$ have two indices within N^* of each other. We can bound the sum over R_3 of these terms by $4A^p 2N^*/N$, where we have used (6.10) and (6.18). The factor of 4 arises since there are four ways of choosing which pair of indices is within N^*. There are $p(p-1)/2$ ways to choose which terms have two indices within N^*, so

$$I(R_3) \leq [(2p)!/2^p p!] 2p(p-1) A^p 2N^*/N. \tag{6.22}$$

Combining (6.20)–(6.22), we obtain the desired result

(6.16); thus the moments $\mu_p(N)$ converge to the Gaussian moments (6.11) as $N \to \infty$.

Using the upper bound in (6.14), together with (6.10) which shows that $\mu_2(N) \leq A$ uniformly in N, we can verify that the Carleman[24] condition (6.4b) holds, since $\Sigma(1/p) = \infty$. Thus we conclude (i) that every distribution

$$F_N(z) = \int_{-\infty}^{z} f_N(z')\, dz' \tag{6.23}$$

is unique in the usual sense of distributions and (ii) by the method of Tchebycheff inequalities[24] we may use the convergence of the first p moments for $N > N_0$ to confine

$$|F_N(z) - G(z)| < \epsilon'(p, \epsilon), \tag{6.24}$$

for $N > N_0$, where $G(z)$ is the cumulative Gaussian distribution. By choosing p large enough and ϵ small enough, we can make ϵ' as small as desired. Since the individual moments converge, we can select an N_0 great enough to satisfy (6.24) with $\epsilon' > 0$ arbitrarily small. Therefore,

$$\lim_{N \to \infty} F_N(z) = G(z). \tag{6.25}$$

To complete the discussion, consider the block spin variables s_j defined in (3.1) with $q_N = 1$,

$$s_j = \frac{1}{N^{1/2}} \sum_{k=0}^{N-1} \sigma_{Nj+k}. \tag{6.26}$$

It follows from translational invariance that, in the limit $N \to \infty$, each spin s_j is normally distributed according to (6.7). It remains to prove that the s_j become independent variables when $N \to \infty$. To show this, consider the expectation value

$$\phi = \langle s_{j_1}^{n_1} s_{j_2}^{n_2} \cdots s_{j_m}^{n_m} \rangle, \tag{6.27}$$

where $\Sigma n_i = 2\nu$ and $j_1 < j_2 \cdots < j_m$. First consider the case when all $n_i = 2\nu_i$ are even. Bounding ϕ from below by the GKS inequalities, we have

$$\begin{aligned} 0 \leq \prod_{i=1}^{m} \langle s_{j_1}^{n_i} \rangle \leq \phi \\ \leq \sum_{L} \prod_{i=1}^{m} \left[N^{-\nu_i} \sum_{\Pi_i} \langle \sigma_{\Pi_1} \sigma_{\Pi_2} \rangle \cdots \langle \sigma_{\Pi_{2\nu_i - 1}} \sigma_{\Pi_{2\nu_i}} \rangle \right] \\ + N^{-\nu} \sum_{L} \sum_{\Pi} \langle \sigma_{\Pi_1} \sigma_{\Pi_2} \rangle \cdots \langle \sigma_{\Pi_{2\nu-1}} \sigma_{\Pi_{2\nu}} \rangle, \end{aligned} \tag{6.28}$$

where $\sum_L$ is the sum over the $N^{2\nu}$ individual σ's, the partitions Π_i contain only pairs of σ's from the same $s_{j_i}^{n_i}$, and the remaining partitions Π are characterized by having at least two pairs containing σ's from different $s_{j_i}^{n_i}$. Suppose these two pairs are denoted $\langle s_a s_b \rangle \langle \sigma_c \sigma_d \rangle$. It follows that $|a-b| \leq N^*$ and $|c-d| \leq N^*$ implies that either $|a-c| \leq N^*$ or $|b-d| \leq N^*$ or both. Using arguments like those leading to (6.20) and (6.22), we see that any one term in the last summation over partitions can be bounded by $(\nu A^{\nu-1}\epsilon + 2\nu(\nu-1)N^* A^\nu/N)$. Since there are fewer than $(2\nu)!/(2^\nu \nu!)$ partitions Π, we can make the second term on the right-hand side of (6.28) as small as desired by selecting ϵ which implies N^* and then choosing N sufficiently large. The first term on the right-hand side of (6.28) becomes equal to the GKS lower bound in the limit $N \to \infty$, as shown earlier in this section. Therefore, we conclude that if all n_i are even,

$$\lim_{N \to \infty} \phi = \lim_{N \to \infty} \prod_{i=1}^{m} \mu_{n_i}(N) = \prod_{i=1}^{m} \left[\frac{(2\nu_i)!}{2^{\nu_i}(\nu_i)!} \chi^{\nu_i} \right]. \tag{6.29}$$

If Σn_i is odd, then ϕ vanishes by the assumed spin symmetry. If Σn_i is even, but any of the n_i are odd, then there must be at least one term in every partition which links two different $s_{j_i}^{n_i}$. Therefore, the argument leading to (6.29) now shows that ϕ vanishes in the limit as $N \to \infty$. Thus we have shown

$$\lim_{N \to \infty} \langle s_{j_1}^{n_1} \cdots s_{j_i}^{n_m} \rangle = \langle s_{j_1}^{n_1} \rangle \cdots \langle s_{j_m}^{n_m} \rangle. \tag{6.30}$$

But this condition is characteristic of independent variables, so the fixed point Hamiltonian is ($\beta = 1$)

$$H^* = \tfrac{1}{2} \sum_{j=-\infty}^{\infty} s_j^2/\chi. \tag{6.31}$$

It is to be noted that the form anticipated on general grounds

$$Q_N(z) \approx N^d q(z/N^{d/2}),$$

where Q_N is defined in (6.5), contains higher order corrections to the fixed point Hamiltonian (6.31).

7. CRITICAL EXPONENT RELATIONS

One of the chief reasons for the current interest in the renormalization group lies in its provision of a framework within which to study the dependence of the thermodynamic functions upon temperature and magnetic field near the critical temperature. In the previous sections we have been exclusively concerned with the action of the renormalization group upon a system in zero magnetic field and at fixed temperature. We will now show how the properties of the renormalization group at arbitrary magnetic field can be related to those in zero field. We will also consider the relation of fixed points at different temperatures, and we will find that *nonlinear* operations intrude in the study of the temperature variation. These considerations will lead to a discussion of critical indices and some relations between them. The general methods we employ are intrinsically similar to those of Wilson,[1] but they appear here in the context of our different approach.

First, let us consider a nonzero magnetic field H. The partition function Z in the presence of a magnetic field is

$$Z/Z_0 = \langle \exp(H \Sigma_i \sigma_i) \rangle, \tag{7.1}$$

where Z_0 is the partition function in zero field and (7.1) is merely symbolic since the sum over "i" is infinite. Using translational invariance, we find that the magnetization per site is

$$M = \langle \sigma_0 \exp(H \Sigma_i \sigma_i) \rangle / \langle \exp(H \Sigma_i \sigma_i) \rangle. \tag{7.2a}$$

Introducing the "mean spin" block variables s_j defined in (3.45), we obtain

$$M = \frac{N^{(-d+\sigma-\omega^*)/2} \langle s_0 \exp(H N^{(d+\sigma-\omega^*)/2} \Sigma_j s_j) \rangle}{\langle \exp(H N^{(d+\sigma-\omega^*)/2} \Sigma_j s_j) \rangle}. \tag{7.2b}$$

A lower bound $M_-(N)$ is obtained by keeping only the $j = 0$ term in the sum in the exponential, since from Griffiths' inequalities[10] we get something smaller by removing ferromagnetic interactions.

$$M \geq M_-(N) = \frac{N^{(-d+\sigma-\omega^*)/2}\langle s_0 \exp(HN^{(d+\sigma-\omega^*)/2} s_0)\rangle}{\langle \exp(HN^{(d+\sigma-\omega^*)/2} s_0)\rangle}. \tag{7.3}$$

Since s_0 is a "mean spin" over an $N\times N\times\cdots\times N$ block, we expect for a reasonable system that $M_-(N)$ will closely approximate M if N is large enough. If we assume $M(T_c-\epsilon, H=0)\to 0$ as $\epsilon\to 0$, then there is a unique equilibrium state,[26] so we obtain convergence $M_-(N)\to M$ as $N\to\infty$. The expression (7.3) only involves the moment-generating function and its derivative for the distribution of s_0:

$$f(x) = \langle \exp(x s_0)\rangle = \sum_{n=0}^{\infty} \frac{x^n}{n!}\langle s_0^n\rangle = \exp\left[\sum_{m=1}^{\infty}\frac{x^m}{m!}\langle s_0^m\rangle_c\right], \tag{7.4}$$

where $\langle s_0^m\rangle_c$ is the mth cumulant. If we assume convergence to a fixed point, then from (3.38) and (3.45) we may compute directly for N sufficiently large that

$$\langle s_0^m\rangle_c = O(N^{\omega^*}). \tag{7.5}$$

Of course, the cumulants vanish for m odd. Then we have

$$f(HN^{(d+\sigma-\omega^*)/2}) = \exp\left[\sum_{m=1}^{\infty}\frac{H^m}{m!}O(N^{\omega^*+(d+\sigma-\omega^*)m/2})\right] = \exp[N^{\omega^*}g(HN^{(d+\sigma-\omega^*)/2})]. \tag{7.6}$$

To obtain the second equality, we have made an assumption in addition to that of convergence to a fixed point. From (7.3) we find

$$M_-(N) = N^{-d}\frac{\partial}{\partial H}\ln f = N^{(-d+\sigma+\omega^*)/2} g'(HN^{(d+\sigma-\omega^*)/2}). \tag{7.7}$$

Since we require a finite magnetization for all H in the limit as $N\to\infty$, Eq. (7.7) implies (provided $d-\sigma > \omega^*$)

$$g(x) \propto x^{(\delta+1)/\delta} \quad (x\to\infty), \tag{7.8}$$

where

$$\delta = \frac{d+\sigma-\omega^*}{d-\sigma-\omega^*}. \tag{7.9}$$

We now conclude that, as $M_-\to M$ as $N\to\infty$,

$$M \propto H^{1/\delta} \quad (H\to 0), \tag{7.10}$$

when there is convergence to a fixed point described by d, σ, and ω^*. A nonzero value of the anomalous dimension of the vacuum breaks hyperscaling. The inequality (4.2), $\omega^*\geq 0$, is seen to correspond to the Buckingham–Gunton[27] inequality $\delta \geq (d+\sigma)/(d-\sigma)$. For the infinite-temperature fixed point of Sec. 6, direct calculation shows $\delta = 1$. Let us remark that the analysis leading to (7.9) depends on the dominant behavior in N in each order of the expansion (7.6). The subdominant terms would be expected to lead to corrections to (7.10) with larger values of $(1/\delta)$, which become important as one moves away from $H=0$.

Let us now discuss the critical exponent γ describing the divergence of the magnetic susceptibility as the critical temperature is approached from above. Suppose the susceptibility per lattice site of the infinite system, χ, is finite for $\beta<\beta_c$, and diverges as $\beta\to\beta_c^-$, where $\beta = 1/kT$. Under repeated application of the renormalization group, the infinite temperature fixed point is approached for $\beta<\beta_c$, by the work of Sec. 6, and we *assume* a critical fixed point is approached for $\beta=\beta_c$. We wish to define a sequence, χ_N, which satisfies $\chi_N\to\chi$ as $N\to\infty$. A possible choice is

$$\tilde{\chi}_N = N^{-d}\sum_{\mathbf{i},\mathbf{j}=0}^{N-1}\langle\sigma_{\mathbf{i}}\sigma_{\mathbf{j}}\rangle, \tag{7.11}$$

but this is not useful for our purposes, because the spin-correlation functions cannot be presumed analytic[25] at $\beta=\beta_c$. To define a sequence which is analytic, let H denote the infinite system Hamiltonian and H_N the the truncation

$$H_N = \sum_{\mathbf{i},\mathbf{j}=0}^{N-1} J(\mathbf{i}-\mathbf{j})\sigma_{\mathbf{i}}\sigma_{\mathbf{j}}. \tag{7.12}$$

Define

$$\chi_N(\beta) = \frac{\mathrm{Tr}\{(N^{-d}\sum_{\mathbf{i},\mathbf{j}=0}^{N-1}\sigma_{\mathbf{i}}\sigma_{\mathbf{j}})\exp[-\beta H_N - \beta_c(H-H_N)]\}}{\mathrm{Tr}\exp[-\beta H_N - \beta_c(H-H_N)]}. \tag{7.13}$$

For convenience we confine our discussion to nearest neighbor interaction although extension of these results is clearly possible. Let us next show that $\chi_N\to\chi$ as $N\to\infty$. Select a large $N\times N\times\cdots\times N$ hypercube of sites in the infinite system. Griffiths' inequalities show that a lower bound on all the correlation functions is obtained by decoupling all bonds that cross the boundaries of the hypercube, and an upper bound is obtained by aligning all boundary spins. Both operations decouple the interior of the hypercube from the infinite exterior, and lead to a finite problem. Define

$$\chi_{N\pm} = N^{-d}\sum_{\mathbf{i},\mathbf{j}=0}^{N-1}\langle\sigma_{\mathbf{i}}\sigma_{\mathbf{j}}\rangle_{\pm}, \tag{7.14}$$

where the subscripts $-$ and $+$ refer to the boundary conditions leading to lower and upper bounds, respectively. Griffiths' inequalities show that, for $\beta\leq\beta_c$,

$$\chi_{N-}(\beta) \leq \chi_N(\beta) \leq \chi_{N+}(\beta). \tag{7.15}$$

Since χ_N^{-1} is finite for $\beta\leq\beta_c$, and we have assumed $M(T_c-\epsilon, H=0)\to 0$ as $\epsilon\to 0$, then it follows as before[25] that $\chi_{N\pm}^{-1}$ converges to χ^{-1} to within a preselected accuracy $\epsilon>0$, over a preselected range $\beta_1\leq\beta\leq\beta_c$, for $N\geq N(\epsilon,\beta_1)$ and so too does χ_N^{-1} converge to χ^{-1}, by (7.15).

Let us now study the temperature dependence of χ_N. We introduce the derivative operation $D/D\beta$ by

$$\frac{D\langle\sigma_A\rangle}{D\beta} = \langle H_N\sigma_A\rangle - \langle H_N\rangle\langle\sigma_A\rangle. \tag{7.16}$$

It is to be noted that (7.16) is a *nonlinear* operation on the correlation vectors. If we repeatedly apply (7.16), we find

$$\frac{D^n\langle\sigma_A\rangle}{D\beta^n} = \langle U_{n+1}(\sigma_A, H_N, H_N, \ldots, H_N)\rangle, \tag{7.17}$$

where $\langle U_n\rangle$ denotes the nth Ursell function. From the definition (7.13), we see that

$$\chi_N(\beta)=\sum_{n=0}^{\infty}\frac{(\beta_c-\beta)^n}{n!}\frac{D^n}{D\beta^n}\left\langle N^{-d}\sum_{i,j=0}^{N-1}\sigma_i\sigma_j\right\rangle\Big|_{\beta=\beta_c}. \tag{7.18}$$

It will be more convenient for us to consider $\chi_N^{-1}(\beta)$, where, for β near β_c,

$$\chi_N^{-1}(\beta)=\chi_N^{-1}\{1-(\chi_N'/\chi_N)(\beta-\beta_c)+\tfrac{1}{2}[((\chi_N')^2-\chi_N\chi_N'')/\chi_N^2]\times(\beta-\beta_c)^2+\tfrac{1}{6}[(6\chi_N\chi_N'\chi_N''-6(\chi_N')^3-\chi_N^2\chi_N''')/\chi_N^3]\times(\beta-\beta_c)^3+\cdots\}. \tag{7.19}$$

For the ordinary Gaussian model, $\chi_N^{-1}(\beta)$ is linear in β. If we reexpress the coefficients in (7.19) in terms of Ursell functions of the individual spins, then the only numerator to depend only on u_2 is that of the linear term; all higher-order terms depend on u_4, u_6, u_2u_4, etc., and no terms like Πu_2. Since $u_{2n}=0$, $n>1$, this is consistent with direct calculation on the Gaussian model.

The pairs of indices entering into H_N are restricted to be nearest neighbors, so we cannot compute the asymptotic behavior in N of (7.19) evaluated at $\beta=\beta_c$ directly from (4.8). To proceed, we use the result of Baker[28] that

$$0\le\chi_N^{-2}\frac{D\chi_N}{D\beta}\le C, \tag{7.20}$$

where C is independent of β. This shows $D(\chi_N^{-1})/D\beta$ is uniformly bounded and since we know from (4.8) that $\chi_N^{-1}(\beta_c)=O(N^{-\sigma})$, we can bound $\chi_N^{-1}(\beta)$ in the neighborhood $\beta_1\le\beta\le\beta_c$.

Furthermore, we can show that (7.19) can be bounded term-by-term by

$$\chi_N^{-1}\sum_{m=0}^{\infty}a_m(N)^{pm}(\beta_c-\beta)^m \tag{7.21}$$

where $p\le d$. To see this result we use Baker's[29] inequality (A23) once on (7.17) to obtain one factor of N^{σ}. The rest of the correlations are bounded by assumption (2). The inequality (A23) was derived from Lebowitz's inequalities, but it follows *a fortiori* from Newman's[11] inequalities. If we select $\zeta=\min(p)$ for which (7.21) holds as a term-by-term bound in the large N limit, then we find in a manner similar to (7.6) that

$$\chi_N^{-1}(\beta)=N^{-\sigma}g(N^{\zeta}(\beta_c-\beta)). \tag{7.22}$$

Since $\chi_N^{-1}(\beta)$ is bounded and β_c is the smallest value of β for which $\chi^{-1}(\beta)=0$, we see that

$$g(x)\sim x^{\sigma/\zeta}\quad(x\to\infty), \tag{7.23}$$

so that

$$\chi^{-1}(\beta)\sim(\beta_c-\beta)^{\sigma/\zeta}\quad(\beta\to\beta_c^-). \tag{7.24}$$

Subdominant terms in (7.23), such as

$$\chi_N^{-1}(\beta)\sim N^{-\sigma}[g(N^{\zeta}(\beta_c-\beta))+g_1(N^{\zeta'}(\beta_c-\beta))], \tag{7.25}$$

with $\zeta'<\zeta$, contribute terms like

$$\chi_N^{-1}(\beta)\sim(\beta_c-\beta)^{\sigma/\zeta}+(\beta_c-\beta)^{\sigma/\zeta'}, \tag{7.26}$$

which are negligible compared to the dominant one for $\beta\to\beta_c^-$. Since (7.24) implies

$$\gamma=\sigma/\zeta, \tag{7.27}$$

and we know[28,30] $\gamma\ge1$, we must have $\zeta\le\sigma$. The index ζ is related to the scaling behavior of the energy-density correlations. Namely, if

$$\langle U_{m+1}(\sigma_{\lambda_1}\sigma_{\lambda_1},\sigma_{\lambda_1+\delta},\ldots,\sigma_{\lambda_k+\delta'})\rangle\le A\lambda^{-n(d-\zeta)-(d-\sigma)}. \tag{7.28}$$

as $\lambda\to\infty$, with the equality holding for at least one $n>0$, then (7.22) holds with perhaps some of the coefficients in its expansion zero.

It is instructive to compare the analysis just presented with Wilson's[1] procedure of computing γ by linearizing the spin distribution about a fixed point. Let us briefly review Wilson's method. Define $Q_N(z)$ by

$$\left\langle\left(N^{-(d+\sigma)/2}\sum_{j=0}^{N-1}\sigma_j\right)^p\right\rangle=\int_{-\infty}^{\infty}dz\,z^p\exp[-Q_N(z)]. \tag{7.29}$$

Assume that when $\beta=\beta_c$,

$$Q_N(z)\to Q^*(z),\quad\text{as } N\to\infty, \tag{7.30}$$

and when $(\beta_c-\beta)$ is small, and N is large but not too large,

$$Q_N(z)\sim Q^*(z)+(\beta_c-\beta)N^{\zeta}Q'(z), \tag{7.31}$$

so that

$$Q_{MN}(z,\beta_c-\beta)=Q_M(z,(\beta_c-\beta)N^{\zeta}). \tag{7.32}$$

If the "translational invariance" of (7.32) remains valid outside the linear region, where from Sec. 6 we know

$$Q_N(z)\sim z^2N^{\sigma}/2\chi_N(\beta_c-\beta), \tag{7.33}$$

then

$$\chi_N^{-1}(\beta_c-\beta)\sim N^{-\sigma}g[(\beta_c-\beta)N^{\zeta}]. \tag{7.34}$$

It follows that

$$\chi^{-1}(\beta_c-\beta)\sim(\beta_c-\beta)^{\sigma/\zeta}, \tag{7.35}$$

so that $\gamma=\sigma/\zeta$. Implicit in this approach is the assumption that the value of ζ so obtained will do for the whole series or that

$$\chi_N\frac{d^n}{d\beta^n}(\chi_N^{-1})\Big/\left(\frac{\chi_N'}{\chi_N}\right)^n\le a_n \tag{7.36}$$

uniformly in N. Relative to the procedure we have used, it is clear that Wilson obtains (if carried through exactly) a lower bound on ζ and hence an upper bound on γ.

We remark that we may compute directly from the definitions

$$\eta=2-\sigma$$

and from Fisher's formula[31] and the Glimm–Jaffe[30] and Baker's[28] inequalities

$$\nu=\gamma/(2-\eta)=1/\zeta\ge\max(2/(d+1),\tfrac{1}{2}). \tag{7.37}$$

The variation of the physical quantities with respect to other parameters, such as a staggered magnetic field, for example, can be studied by similar methods. The limiting validity of the long-distance scaling formulas like (3.38) and (7.28) are fundamental to validity of this approach.

*Work performed under the auspices of the USERDA.

[1]K.G. Wilson, Phys. Rev. B 4, 3174, 3184 (1971).
[2]Reviews of the expansions in $\epsilon = 4 - d$ and $1/n$ are given by: K.G. Wilson and J. Kogut, Phys. Rep. C 12, 75 (1974); M.E. Fisher, Rev. Mod. Phys. 46, 597 (1974); S.K. Ma, Rev. Mod. Phys. 45, 589 (1973); E. Brezin, J.C. Le Guillou, and J. Zinn-Justin, *Phase Transitions and Critical Phenomena, Vol.* 6, edited by C. Domb and M.S. Green (Academic, New York, to be published).
[3]Approximate calculations in two and three dimensions have been performed by: Th. Niemeijer and J.M.J. van Leeuwen, Physica 71, 17 (1974); B. Nienhuis and M. Nauenberg, Phys. Rev. B 11, 4152 (1975); L.P. Kadanoff, Phys. Rev. Lett. 34, 1005 (1975); G.R. Golner and E.K. Riedel. Phys. Rev. Lett. 34, 856 (1975).
[4]J.R. Klauder, "A Characteristic Glimpse of the Renormalization Group" (Bell Laboratories Preprint); K. Subbarao, J. Math. Phys. 17, 255 (1976).
[5]J.S. Høye, Phys. Rev. B 12, 5219 (1975).
[6]C. Domb, in *Phase Transitions and Critical Phenomena, Vol.* 3, edited by C. Domb and M.S. Green (Academic, New York, 1974).
[7]L.P. Kadanoff, Physics 2, 263 (1966).
[8]D.R. Nelson and M.E. Fisher, Ann. Phys. (N.Y.) 91, 226 (1975); S. Krinsky and D. Furman, Phys. Rev. B 11, 2602 (1975).
[9]D. Ruelle, *Statistical Mechanics, Rigorous Results* (Benjamin, New York, 1969).
[10]R.B. Griffiths, J. Math. Phys. 8, 478, 484 (1967); D.G. Kelly and S. Sherman, J. Math. Phys. 9, 466 (1968); J. Ginibre, Phys. Rev. Lett. 23, 828 (1969).
[11]C.M. Newman, "Gaussian Correlation Inequalities for Ferromagnets," Z. Wahrscheinlichkeitstheorie (to be published).
[12]G. Stell, Phys. Rev. Lett. 24, 1343 (1970); M.E. Fisher, in *Collective Properties of Physical Systems, Proceedings of the Twenty-fourth Nobel Symposium*, edited by B. Lundqvist and S. Lundqvist (Academic, New York, 1973).
[13]E.R. Lorch, Bull. Am. Math. Soc. 45, 945 (1939); see also, F. Riesz and B. Sz-Nagy, *Functional Analysis* (Ungar, New York, 1955), Chap. 10.
[14]G.A. Baker, Jr., Phys. Rev. B 5, 2622 (1972).
[15]F.J. Dyson, Commun. Math. Phys. 12, 91, 212 (1969); 21, 269 (1971).
[16]G. Jona-Lasinio, Nuovo Cimento B 26, 99 (1975).
[17]G. Gallavott, and H. Knops, "The Hierarchical Model and the Renormalization Group" (University of Nijmegen preprint).
[18]R.B. Griffiths, C.A. Hurst, and S. Sherman, J. Math. Phys. 11, 740 (1970).
[19]J. Glimm and A. Jaffe, Ann. Inst. H. Poincaré A 22, 97 (1975).
[20]K. Kawasaki and J. Gunton, Phys. Rev. Lett. 35, 131 (1975).
[21]M. Duneau, B. Souillard, and D. Iagolnitzen, J. Math. Phys. 16, 1662 (1975); Commun. Math. Phys. 31, 191 (1973); 35, 307 (1974).
[22]A. Cayley, Quart. J. Math. 23, 376 (1889).
[23]E.F. Beckenbach and R. Bellman, *Inequalities* (Springer-Verlag, Berlin, 1961), p. 16.
[24]See, for example, J.D. Tamarkin and J.A. Shohat, *The Problem of Moments* (Amer. Math. Soc., New York, 1943); G.A. Baker, Jr., *Essentials of Padé Approximants* (Academic, New York, 1975).
[25]B.M. McCoy and T.T. Wu, Phys. Rev. 162, 436 (1967); 174, 546 (1968); *The Two Dimensional Ising Model* (Harvard U.P., Cambridge, 1973), Chap. 13.
[26]J.L. Lebowitz and A. Martin-Löf, Commun. Math. Phys. 25, 276 (1972).
[27]M.J. Buckingham and J.D. Gunton, Phys. Rev. 178, 848 (1969); see also, M.E. Fisher, Phys. Rev. 180, 594 (1969).
[28]G.A. Baker, Jr., Phys. Rev. Lett. 34, 268 (1975).
[29]G.A. Baker, Jr., J. Math. Phys. 16, 1324 (1975).
[30]J. Glimm and A. Jaffe, Phys. Rev. D 10, 536 (1974).
[31]M.E. Fisher, Rep. Prog. Phys. 30, 615 (1967).

Identification of the velocity operator for an irreducible unitary representation of the Poincaré group

Thomas F. Jordan

*Department of Physics, University of Minnesota, Duluth, Duluth, Minnesota 55812**
and Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627†
(Received 12 October 1976)

For a particle described by an irreducible unitary representation of the Poincaré group, for either positive mass or zero mass and discrete helicity, it is shown that the velocity operator can be identified by its transformations under the Poincaré group together with the assumption that it is a Hermitian operator whose different components commute with each other.

1. INTRODUCTION

Consider an irreducible unitary representation of the Poincaré group, for either positive mass or zero mass and discrete helicity, and nonnegative energy.[1] Let $\mathbf{P}$ denote the generators for space translations. Then the Hamiltonian, the generator for changes in time, is

$$H = (\mathbf{P}^2 + m^2)^{1/2} \tag{1.1}$$

with m a nonnegative number. (We use units such that $c = 1$ and later, when it enters, $\hbar = 1$ also.) Suppose we are to use the representation to describe a particle. Let us pretend that we do not know the physical meaning of any operators except that of the unitary operators for transformations of wavefunctions corresponding to the Poincaré group of transformations of space–time coordinates. For example, pretend that we know $\mathbf{P}$ and H only as generators and do not know that they represent the momentum and energy of the particle.

Our goal here will be just to identify the operator $\mathbf{V}$ that represents the particle's velocity as being

$$\mathbf{V} = \mathbf{P}/H. \tag{1.2}$$

The traditional way to do this is to identify the Newton–Wigner position operator[2] and take its time derivative by calculating its commutator with H.[3] Along the way one encounters difficult and subtle questions about the uniqueness of the position operator,[2,4] about its lack of Lorentz covariance,[2,5] and indeed about the meaning of positions in a relativistic quantum theory.[6] For zero mass and nonzero helicity there is no Newton–Wigner position operator.[2,7] Ahmad and Wigner[8] have suggested that it might be better to identify the velocity directly. They outline a method of doing so which is clear in its conceptual foundation but, as they were the first to point out, is rather involved with limits and approximations in its mathematical detail.

Here we wish to show, using simple commutator algebra, that the velocity operator is identified by its transformations under the Poincaré group and the assumption that it is a Hermitian operator whose different components commute with each other.

2. LORENTZ TRANSFORMATIONS OF VELOCITY

Our first assumption is that the velocity is translation invariant, so $\mathbf{V}$ commutes with $\mathbf{P}$. It follows from formula (1.1) for H that $\mathbf{V}$ commutes with H, so the velocity is constant in time. We are necessarily describing a freely moving particle.

If the velocity is measurable, the different components must be simultaneously measurable. We assume that the different components of the velocity operator commute with each other. Then the classical equations for Lorentz transformations of velocity can be quantized without any problem.

The velocity $\mathbf{V}'$ for an observer moving with velocity β in the z direction is

$$V_x' = (1-\beta^2)^{1/2} V_x (1-\beta V_z)^{-1},$$
$$V_y' = (1-\beta^2)^{1/2} V_y (1-\beta V_z)^{-1},$$
$$V_z' = (V_z - \beta)(1-\beta V_z)^{-1}.$$

To first order in β this is

$$V_x' = V_x + \beta V_x V_z,$$
$$V_y' = V_y + \beta V_y V_z,$$
$$V_z' = V_z + \beta V_z V_z - \beta.$$

Let $\mathbf{K}$ denote the generator for Lorentz transformations in the representation of the Poincaré group. We let

$$\exp(iK_z \tanh^{-1}\beta)\,\mathbf{V}\exp(-iK_z\tanh^{-1}\beta) = \mathbf{V}'.$$

Expanding the exponentials to first order in β, we see that

$$(1/i)[V_x, K_z] = V_x V_z,$$
$$(1/i)[V_y, K_z] = V_y V_z,$$
$$(1/i)[V_z, K_z] = V_z V_z - \mathbf{1}.$$

In this way we conclude that, in general,

$$(1/i)[V_j, K_k] = V_j V_k - \delta_{jk}, \tag{2.1}$$

for $j, k = 1, 2, 3$.

For an irreducible unitary representation of the Poincaré group for either positive mass[9] or zero mass and discrete helicity,[10] the generator for Lorentz transformations is of the form

$$\mathbf{K} = \tfrac{1}{2}(H\mathbf{Q} + \mathbf{Q}H) + \mathbf{N}, \tag{2.2}$$

where $\mathbf{N}$ is an operator that commutes with $\mathbf{P}$. We use momentum-space wavefunctions, on which $\mathbf{P}$ is multiplication and $\mathbf{Q}$ is $i\nabla$, with the ordinary inner product defined with the noninvariant $\int d^3p$.

 Copyright © 1977 American Institute of Physics

Suppose that, as we will show later, the velocity operator is of the form

$$\mathbf{V}=F(\mathbf{P}^2)\mathbf{P} \tag{2.3}$$

with some real function F of $\mathbf{P}^2$. Substituting this and the formula (2.2) for $\mathbf{K}$ into the commutation relations (2.1), we see immediately that these commutation relations are satisfied only for

$$F=H^{-1}.$$

Thus we obtain the formula (1.2) for $\mathbf{V}$.

3. REPRESENTATIONS FOR POSITIVE MASS

Now we will show that for a particle described by an irreducible unitary representation of the Poincare group with $m>0$ the velocity operator is of the form (2.3). The generators for the representation of the Poincaré group are $\mathbf{P}$ for space translations, H, given by Eq. (1.1), for changes in time,

$$\mathbf{J}=\mathbf{Q}\times\mathbf{P}+\mathbf{S} \tag{3.1}$$

for rotations, and

$$\mathbf{K}=\tfrac{1}{2}(H\mathbf{Q}+\mathbf{Q}H)+(H+m)^{-1}\mathbf{P}\times\mathbf{S} \tag{3.2}$$

for Lorentz transformations.[9] We use $(2s+1)$-component momentum-space wavefunctions on which $\mathbf{P}$ is multiplication, $\mathbf{Q}$ is $i\nabla$, and $\mathbf{S}$ is the usual $(2s+1)\times(2s+1)$ spin matrix. The inner product is

$$(\psi,\phi)=\sum_{\mu=-s}^{s}\int d^3p\,\psi_\mu(\mathbf{p})^*\phi_\mu(\mathbf{p}).$$

We assume, as we already mentioned, that the velocity $\mathbf{V}$ is invariant under space translations, which implies that it is also constant in time. We assume that $\mathbf{V}$ transforms as a vector for rotations, and as described in the previous section for Lorentz transformations. We also assume, as we already mentioned, that the different components of $\mathbf{V}$ commute with each other.

Let

$$U_0=(1-\mathbf{V}^2)^{-1/2},\quad \mathbf{U}=\mathbf{V}(1-\mathbf{V}^2)^{-1/2}.$$

From the Lorentz transformations of $\mathbf{V}$, described by the commutation relations (2.1), it follows that U_0, $\mathbf{U}$ transform as a four-vector, in particular that

$$(1/i)[U_0,K_k]=-U_k \tag{3.3}$$

for $k=1,2,3$. From the formulas (3.1) and (3.2) for $\mathbf{J}$ and $\mathbf{K}$ we can see that

$$H\mathbf{J}+\mathbf{P}\times\mathbf{K}-(H+m)^{-1}(\mathbf{P}\cdot\mathbf{J})\mathbf{P}=m\mathbf{S}. \tag{3.4}$$

Since U_0 must commute with H, $\mathbf{P}$, and $\mathbf{J}$, and $\mathbf{U}$ must commute with $\mathbf{P}$, it follows that

$$(1/i)[U_0,\mathbf{S}]=(1/m)\mathbf{U}\times\mathbf{P}, \tag{3.5}$$

$$(1/i)[U_0,\mathbf{P}\times\mathbf{S}]=(1/m)(\mathbf{P}^2)\mathbf{U}-(1/m)(\mathbf{P}\cdot\mathbf{U})\mathbf{P}. \tag{3.6}$$

The operators $\mathbf{P}$ and $\mathbf{P}\cdot\mathbf{S}=\mathbf{P}\cdot\mathbf{J}$ are a complete set of commuting operators. Since U_0 commutes with these operators, it must be a function of them. Since U_0 is a scalar for rotations, it must be a function of $\mathbf{P}^2$ and $\mathbf{P}\cdot\mathbf{S}$. Also $\mathbf{P}\cdot\mathbf{U}$ must be a function of $\mathbf{P}^2$ and $\mathbf{P}\cdot\mathbf{S}$; since $\mathbf{U}$ must be a vector for rotations, $\mathbf{P}\cdot\mathbf{U}$ must be a scalar, and $\mathbf{P}\cdot\mathbf{U}$ commutes with $\mathbf{P}$. Since the matrices $\mathbf{S}$ are $(2s+1)$-dimensional, a function of $\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|$ is a polynomial containing powers no higher than $(\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|)^{2s}$. Using the commutation relations

$$(1/i)[\mathbf{P}\cdot\mathbf{S},\mathbf{P}\times\mathbf{S}]=(\mathbf{P}^2)\mathbf{S}-(\mathbf{P}\cdot\mathbf{S})\mathbf{P},$$

$$(1/i)[\mathbf{P}\cdot\mathbf{S},\mathbf{S}]=-\mathbf{P}\times\mathbf{S},$$

we can see from Eq. (3.6) that

$$\mathbf{U}=A\mathbf{P}+B\mathbf{S}+C\mathbf{P}\times\mathbf{S}, \tag{3.7}$$

with A,B,C functions of $\mathbf{P}^2$ and $\mathbf{P}\cdot\mathbf{S}$. In particular, C as a function of $\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|$ is a polynomial containing powers no higher than $(\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|)^{2s-2}$ because such terms occur only after the commutator with $\mathbf{P}\cdot\mathbf{S}$ takes one power of $\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|$ and commutation moving the resulting $\mathbf{S}$ to the right takes another power of $\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|$. For $s=0$ or $\frac{1}{2}$ there is no C.

Using the form (3.7) for $\mathbf{U}$ and the commutation relations (3.5) and (3.6), we find that

$$\begin{aligned} m(1/i)[U_0,\mathbf{U}] &= mB(1/i)[U_0,\mathbf{S}]+mC(1/i)[U_0,\ \mathbf{P}\times\mathbf{S}] \\ &= B\{B\mathbf{S}\times\mathbf{P}+C(\mathbf{P}\times\mathbf{S})\times\mathbf{P}\}+C\{\mathbf{P}^2B\mathbf{S}-B(\mathbf{P}\cdot\mathbf{S})\mathbf{P} \\ &\qquad +\mathbf{P}^2C\mathbf{P}\times\mathbf{S}\} \\ &= -2BC(\mathbf{P}\cdot\mathbf{S})\mathbf{P}+2BC\mathbf{P}^2\mathbf{S}-(B^2-\mathbf{P}^2C^2)\mathbf{P}\times\mathbf{S}. \end{aligned} \tag{3.8}$$

This must be zero, because our assumption that the different components of $\mathbf{V}$ commute with each other implies that U_0 commutes with $\mathbf{U}$.

Taking the commutator of Eq. (3.8) with $\mathbf{P}\cdot\mathbf{S}$, we get

$$0=(B^2-\mathbf{P}^2C^2)\{\mathbf{P}^2\mathbf{S}-(\mathbf{P}\cdot\mathbf{S})\mathbf{P}\}+2BC\mathbf{P}^2\mathbf{P}\times\mathbf{S}.$$

Adding this to $\pm i|\mathbf{P}|$ times Eq. (3.8), we get

$$0=(B^2-\mathbf{P}^2C^2\pm i2BC|\mathbf{P}|)\{\mathbf{P}^2\cdot\mathbf{S}-(\mathbf{P}\cdot\mathbf{S})\mathbf{P}\mp i|\mathbf{P}|\mathbf{P}\times\mathbf{S}\}$$

Letting the z axis be in the direction of $\mathbf{P}$ for the moment, we see that we have

$$0=(B\pm i|\mathbf{P}|C)^2\{S_x\pm iS_y\}.$$

This implies that $B+i|\mathbf{P}|C$ could be nonzero only on eigenvectors of $\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|$ for the eigenvalue $-s$, and $B-i|\mathbf{P}|C$ could be nonzero only on eigenvectors for the eigenvalue s of $\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|$. Both B and C must be zero on the eigenvectors of $\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|$ for the other $2s-1$ eigenvalues between $-s+1$ and $s-1$. But C as a function of $\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|$ is a polynomial containing powers no higher than $(\mathbf{P}\cdot\mathbf{S}/|\mathbf{P}|)^{2s-2}$. Therefore, C is zero. (Remember there was no C at the start for $s=0$ or $\frac{1}{2}$.)

Now, from Eq. (3.8), we see that B is zero. Then, from Eq. (3.5) and what is left of (3.7), we see that U_0 commutes with $\mathbf{S}$, which means U_0 is a function of only $\mathbf{P}^2$. Finally, from Eq. (3.3) and the formula (3.2) for $\mathbf{K}$, we see that $\mathbf{U}$ is a function of $\mathbf{P}$ only and $\mathbf{V}=\mathbf{U}/U_0$ is a function of $\mathbf{P}$ only. Thus, with our assumption that $\mathbf{V}$ is a vector for rotations, we obtain the form (2.3) for $\mathbf{V}$.

4. REPRESENTATIONS FOR ZERO MASS

Now we establish the form (2.3) of $\mathbf{V}$ for $m=0$. We

assume that **V** is invariant under space translations, a vector for rotations, and a Hermitian operator whose different components commute with each other.

We let the particle states span two irreducible unitary representations of the Poincaré group for $m=0$ and two opposite helicites $\pm|\lambda|$. (This is what is usually done for photons. It allows the parity operator to be defined on the particle states. Although we do not use parity here, we do want the broadest definition of the particle states.) We use Pauli matrices σ to represent operators on the two helicity states, for example,

$$\mathbf{P}\cdot\mathbf{J}/|\mathbf{P}| = |\lambda|\sigma_3$$

for the helicity itself.

With the convention that the value of **P** left invariant by the little group is the unit vector in the z direction, the rotation generator is[10]

$$\mathbf{J}=\mathbf{L}+\mathbf{M}\sigma_3 \tag{4.1}$$

with

$$\mathbf{L}=\mathbf{Q}\times\mathbf{P}, \quad M_1=|\lambda|P_1/(|\mathbf{P}|+P_3),$$
$$M_2=|\lambda|P_2/(|\mathbf{P}|+P_3), \quad M_3=|\lambda|. \tag{4.2}$$

Here $\mathbf{Q}=i\nabla$ and **P** is multiplication on two-component momentum-space wavefunctions with inner product

$$(\psi,\phi)=\sum_{\lambda=\pm|\lambda|}\int d^3p\,\psi_\lambda(\mathbf{p})^*\,\phi_\lambda(\mathbf{p}).$$

The translation-invariant Hermitian velocity operator must be of the form

$$\mathbf{V}=\mathbf{A}+\mathbf{B}\sigma_1+\mathbf{C}\sigma_2+\mathbf{D}\sigma_3$$

with **A**, **B**, **C**, **D**, functions of **P**. The assumption that **V** rotates as a vector, expressed in the commutation relations

$$[V_j,J_k]=i\sum_{l=1}^{3}\epsilon_{jkl}V_l$$

for $j,k=1,2,3$, yields

$$[A_j,L_k]=i\sum_{l=1}^{3}\epsilon_{jkl}A_l \tag{4.3}$$

from the terms containing no Pauli matrices,

$$[B_j,L_k]+i2C_jM_k=i\sum_{l=1}^{3}\epsilon_{jkl}B_l \tag{4.4}$$

from the terms containing σ_1,

$$[C_j,L_k]-i2B_jM_k=i\sum_{l=1}^{3}\epsilon_{jkl}C_l \tag{4.5}$$

from the terms containing σ_2, and

$$[D_j,L_k]=i\sum_{l=1}^{3}\epsilon_{jkl}D_l \tag{4.6}$$

from the terms containing σ_3. Equations (4.3) and (4.6) simply tell us that **A** and **D** are vector functions of **P** for rotations, so

$$\mathbf{A}=a(\mathbf{P}^2)\mathbf{P}, \tag{4.7}$$

$$\mathbf{D}=d(\mathbf{P}^2)\mathbf{P}, \tag{4.8}$$

with some functions a and d of $\mathbf{P}^2$. We multiply Eqs. (4.4) and (4.5) by P_k and sum over $k=1,2,3$, to use the fact the $\mathbf{L}\cdot\mathbf{P}$ is zero, and obtain

$$2\mathbf{C}(\mathbf{M}\cdot\mathbf{P})=\mathbf{P}\times\mathbf{B}, \quad -2\mathbf{B}(\mathbf{M}\cdot\mathbf{P})=\mathbf{P}\times\mathbf{C}.$$

From Eqs. (4.2) we see that

$$\mathbf{M}\cdot\mathbf{P}=|\lambda|\,|\mathbf{P}|,$$

so we have

$$\mathbf{C}=\mathbf{P}\times\mathbf{B}/2|\lambda|\,|\mathbf{P}|, \tag{4.9}$$

$$\mathbf{B}=-\mathbf{P}\times\mathbf{C}/2|\lambda|\,|\mathbf{P}|. \tag{4.10}$$

From our assumption that the different components of **V** commute, we obtain

$$0=[V_j,V_k]$$
$$=(B_jC_k-C_jB_k)i2\sigma_3+(C_jD_k-D_jC_k)i2\sigma_1+(D_jB_k-B_jD_k)i2\sigma_2$$

for $j,k=1,2,3$, from which we see, in particular, that

$$\mathbf{B}\times\mathbf{C}=0.$$

From this and Eqs. (4.9) and (4.10) it follows that both **B** and **C** are zero.

Thus, for each of the two irreducible representations of the Poincaré group, we obtain the form (2.3) for the velocity operator,

$$\mathbf{V}=a(\mathbf{P}^2)\mathbf{P}+d(\mathbf{P}^2)\mathbf{P}$$

for the positive-helicity representation, and

$$\mathbf{V}=a(\mathbf{P}^2)\mathbf{P}-d(\mathbf{P}^2)\mathbf{P}$$

for the negative-helicity representation. This is all we need for the proof in Sec. 2 that **V** is in fact $\mathbf{P}/H$.

*Permanent.
†1976—77.

[1]E. P. Wigner, Ann. Math. 40, 149 (1939); V. Bargmann and E. P. Wigner, Proc. Nat. Acad. Sci. U.S. 34, 211 (1948).
[2]T. D. Newton and E. P. Wigner, Rev. Mod. Phys. 21, 400 (1949); A. S. Wightman, Rev. Mod. Phys. 34, 845 (1962).
[3]T. F. Jordan, Am. J. Phys. 43, 1089 (1975).
[4]A. Galindo, Nuovo Cimento 37, 413 (1965).
[5]T. O. Philips, Phys. Rev. 136, B893 (1964).
[6]G. C. Hegerfeldt, Phys. Rev. D 10, 3320 (1974).
[7]J. M. Jauch and C. Piron, Helv. Phys. Acta 40, 559 (1967); W. O. Amrein, Helv. Phys. Acta 42, 149 (1969).
[8]S. M. W. Ahmad and E. P. Wigner, Nuovo Cimento A 28, 1 (1975).
[9]L. L. Foldy, Phys. Rev. 102, 568 (1956); 122, 275 (1961).
[10]C. Fronsdal, Phys. Rev. 113, 1367 (1959).

Ashkin–Teller model as a vertex problem*

F. Y. Wu

Department of Physics, Northeastern University, Boston, Massachusetts 02115
(Received 22 October 1976)

It is shown that the Ashkin–Teller model on any planar lattice is equivalent to an eight-vertex model on a related lattice. The exact equivalence is given for finite lattices with a boundary. We show, in particular, that the AT model on the triangular or honeycomb lattice is related to an eight-vertex model on a Kagomé lattice. The occurrence of two phase transitions in the AT model in general is also discussed.

There has been considerable recent interest in the Ashkin–Teller (AT) model[1] and the Potts model[2] considered as models for phase transitions in lattice systems. Although neither of these models has been solved to this date, interesting transformation properties have been found which relate these models to other lattice statistical problems. For the Potts model on a square lattice, Temperley and Lieb[3] showed that the model is related to a six-vertex problem. This relationship has since been generalized to arbitrary planar lattice,[4] and has been used to obtain properties of the Potts model at the critical point.[5] For the AT model on a square lattice, it was shown by Wegner[6] that the model is reducible to an eight-vertex problem. It does not appear to be widely known, however, that the validity of this equivalence is more general, also extendible to arbitrary planar lattices. It is the purpose of this note to explicitly derive this relation. We shall establish the equivalence for finite lattices with a boundary. We also briefly discuss the occurrence of two phase transitions in the AT model in general when the lattice becomes infinite.

Consider a planar lattice L composed of N sites and E edges. For simplicity we shall assume L be multiply connected and without single-edge loops, but otherwise it is arbitrary. Let the sites of L be occupied by atoms which can be in one of four states, A, B, C, or D. Two atoms occupying the sites connected by an edge interact with an energy ϵ_0 if they are in the same state. The interaction is ϵ_1 between states AB, CD, ϵ_2 between AC, BD, and ϵ_3 between AD, BC. This defines the AT model. We shall suppose $\epsilon_0 \leq \epsilon_i$, $i=1, 2, 3$, so that the system is analogous to an Ising ferromagnet.

The Ising representation of the AT model introduced by Fan[7] proves to be useful. Introduce at the ith site of L two Ising spins (σ_i, τ_i) and identify the spin configuration $(+,+)$ as state A, $(+,-)$ as B, $(-,+)$ as C, and $(-,-)$ as D. Then, the energies ϵ_0, ϵ_1, ϵ_2, ϵ_3 can be represented by 2 and 4 spin interactions between neighboring pairs of spins (σ, τ) and (σ', τ'), and the partition function of the AT model takes the form

$$Z_{\mathrm{AT}}(\omega_0, \omega_1, \omega_2, \omega_3) = \sum_{\{\sigma,\tau\}} \prod_{\{E\}} \exp(K_0 + K_1\sigma\sigma' + K_2\tau\tau' + K_3\sigma\sigma'\tau\tau'), \tag{1}$$

where

$$\omega_i = \exp(-\epsilon_i/kT),$$
$$\exp(4K_0) = \omega_0\omega_1\omega_2\omega_3, \qquad \exp(4K_1) = \omega_0\omega_1/\omega_2\omega_3, \tag{2}$$
$$\exp(4K_2) = \omega_0\omega_2/\omega_1\omega_3, \qquad \exp(4K_3) = \omega_0\omega_3/\omega_1\omega_2.$$

The product in (1) is taken over the edge set $\{E\}$ of L, and Z_{AT} is symmetric in $(\omega_1, \omega_2, \omega_3)$ or (K_1, K_2, K_3).

Following Wegner,[6] we now effect a duality transformation[8,9] on one set of the spins, say $\{\tau\}$. For this purpose the sites of L_D, the dual of L, are located inside the faces of L, and the area outside the boundary of L is considered as one single face.[10] The number of of sites N_D of L_D is then given by the topological relation

$$N_D = E - N + 2. \tag{3}$$

Let μ_i be the Ising spin at the ith site of L_D. The exact duality relation for a finite Ising model now reads[8,9]

$$\sum_{\{\tau\}} \prod_{\{E\}} \exp(K_{ij}\tau_i\tau_j) = 2^{1-N_D} \sum_{\{\mu\}} \prod_{\{D\}} [\exp(K_{ij}) + \mu_i\mu_j \exp(-K_{ij})], \tag{4}$$

where the product on the rhs of (4) is taken over the edge set $\{D\}$ of L_D. We have also assumed the Ising parameters in (4) to be edge-dependent.

It is now a simple matter to substitute (4) into (1) to obtain

$$Z_{\mathrm{AT}} = 2^{1-N_D} \sum_{\{\sigma,\tau\}} \prod_{\{E\}} w(\sigma\sigma', \mu\mu'), \tag{5}$$

where the factors

$$w(\sigma\sigma', \mu\mu') = \exp(K_0 + K_1\sigma\sigma')\{\exp(K_2 + K_3\sigma\sigma') + \mu\mu' \exp[-(K_2 + K_3\sigma\sigma')]\} \tag{6}$$

are associated with the edges. It is convenient at this point to introduce the medial lattice[4,11] M obtained by connecting the midpoints of the adjacent edges of L. An example is shown in Fig. 1. The sites of M are located on the edges of L (or L_D) and the edges of M separate

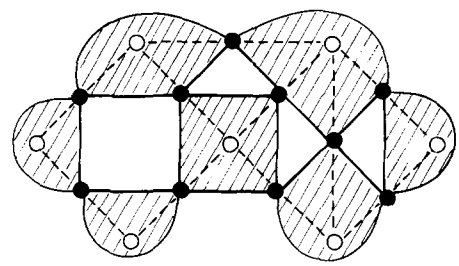


FIG. 1. A lattice L (open circles and broken lines) and its medial lattice M (solid circles and full lines). The 7 σ spins at the open circles are located inside the shaded faces of M and the 5 μ spins inside the plain faces. The number of sites on M is $E=10$.

Copyright © 1977 American Institute of Physics

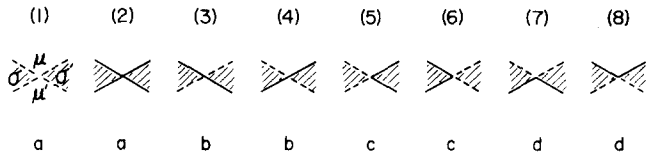


FIG. 2. The 8 vertex configurations on M and the associated vertex weights.

the σ spins from the μ spins. The faces of M containing the σ spins have been shaded in Fig. 1. Since we allow curved edges in M, it is not necessary to introduce the "exterior" sites of M as in Refs. 4 and 11, and the medial lattice has the exact coordination number 4 at all sites.

Following the standard procedure,[12] the spin configuration $\{\sigma, \mu\}$ on L and L_D can be mapped into vertex configurations on M. We do this by drawing bonds along the edges of M which separate spins of opposite signs. Then, as shown in Fig. 2, there are eight allowed configurations at a given vertex of M. A typical mapping of the configurations is shown in Fig. 3. The weights $w(\sigma\sigma', \mu\mu')$ are now considered as the vertex weights of M. Explicitly, we find

$$\begin{aligned} a &= w(+,+) = \omega_0 + \omega_1, \\ b &= w(-,-) = \omega_2 - \omega_3, \\ c &= w(-,+) = \omega_2 + \omega_3, \\ d &= w(+,-) = w_0 - \omega_1. \end{aligned} \tag{7}$$

Equation (5) now reads

$$Z_{\mathrm{AT}}(\omega_0, \omega_1, \omega_2, \omega_3) = 2^{2-N_D} Z_{8v}(a, b, c, d), \tag{8}$$

where Z_{8v} is the generating partition function of the eight vertex model on M. The extra factor 2 in (8) is due to the 2 to 1 mapping of the spin and vertex configurations. The equivalence (8) is our main result and is valid for any *finite* planar lattice.

Several remarks are in order. First, we point out that the equivalence (8) is readily extended to the AT model with edge-dependent energies. In this case one arrives at a vertex model with site-dependent weights provided that (7) holds locally at all sites of M. Next we remark on the consequences of the equivalence on regular lattices. If L is a square lattice, then the medial lattice M is also a square lattice. However, the geometry of M is such that its vertices are rotated 90° from site to site.[13] It follows that the vertex weights c and d are interchanged from site to site and the resulting vertex model is "staggered".[6] Conversely, the Baxter eight vertex model[14] with the weights $\bar{a}$, $\bar{b}$, $\bar{c}$, $\bar{d}$ uniformly at all sites is equivalent to an AT model whose interactions are different in the horizontal and vertical directions. Explicitly, we have

$$\begin{aligned} \omega_0^h &= \tfrac{1}{2}(\bar{a} + \bar{d}), & \omega_1^h &= \tfrac{1}{2}(\bar{a} - \bar{d}), \\ \omega_2^h &= \tfrac{1}{2}(\bar{b} + \bar{c}), & \omega_3^h &= \tfrac{1}{2}(\bar{b} - \bar{c}), \\ \omega_0^v &= \tfrac{1}{2}(\bar{a} + \bar{c}), & \omega_1^v &= \tfrac{1}{2}(\bar{a} - \bar{c}), \\ \omega_2^v &= \tfrac{1}{2}(\bar{b} + \bar{d}), & \omega_3^v &= \tfrac{1}{2}(\bar{b} - \bar{d}). \end{aligned} \tag{9}$$

If L is the triangular or honeycomb lattice, then M is the Kagomé lattice whose vertices are rotated 120° from site to site.[15] Now the 120° rotation is the only logical way of assigning directions and weights to the vertices of a Kagomé lattice. This suggests that the resulting eight vertex model on the Kagomé lattice is *regular* (as contrasted to *staggered*). Extending this reasoning to a square lattice, it is amusing to note that if the eight vertex model in (8) is considered *regular*, then it is the Baxter model that should be termed *staggered*!

Our result also leads to a duality relation for the AT model. Since going from L to L_D amounts to an interchange of $\{\sigma\}$ and $\{\mu\}$, or c and d in (7), we obtain the following duality relation for a *finite* AT model[9]:

$$2^{-N} Z_{\mathrm{AT}}(\omega_0, \omega_1, \omega_2, \omega_3) = 2^{-N_D} Z_{\mathrm{AT}}^{(D)}(\omega_0^*, \omega_1^*, \omega_2^*, \omega_3^*), \tag{10}$$

where $Z_{\mathrm{AT}}^{(D)}$ is the partition function of the AT model on L_D with weights

$$\begin{aligned} \omega_0^* &= \tfrac{1}{2}(\omega_0 + \omega_1 + \omega_2 + \omega_3), & \omega_1^* &= \tfrac{1}{2}(\omega_0 + \omega_1 - \omega_2 - \omega_3), \\ \omega_2^* &= \tfrac{1}{2}(\omega_0 + \omega_2 - \omega_1 - \omega_3), & \omega_3^* &= \tfrac{1}{2}(\omega_0 + \omega_3 - \omega_1 - \omega_2). \end{aligned} \tag{11}$$

Extension of the duality (10) to the AT model with edge-dependent weights is obvious.

The vertex weight (6) can also be written as

$$w(\sigma\sigma', \mu\mu') = \exp(K_0^* + K_1^* \sigma\sigma' + K_2^* \mu\mu' + K_4^* \sigma\sigma' \mu\mu'), \tag{12}$$

with

$$\begin{aligned} \exp(4K_0^*) &= abcd, & \exp(4K_1^*) &= ab/cd, \\ \exp(4K_2^*) &= ac/bd, & \exp(4K_4^*) &= ad/bc. \end{aligned} \tag{13}$$

This is a generalization of the usual Ising representation[12] of the Baxter model, and the AT model is again pictured as two Ising models coupled with 4-spin interactions.

Finally, we remark on the boundary conditions. The equivalence (8) is valid for any finite planar lattice with a boundary. It also holds for planar lattices with cylindrical boundary conditions, for such lattices can be embedded in a plane and the dual can be defined accordingly. It is not clear how to extend the present result to planar lattices with toroidal boundary conditions. While the dual of such lattices can be defined in

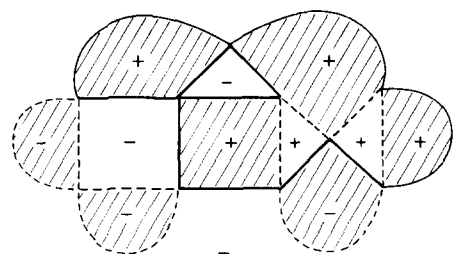


FIG. 3. The mapping of a typical spin configuration on L and L_D into a bond configuration on M.

an obvious way, the duality relation (4) for finite Ising models no longer holds.[16] Consequently, the equivalence (8) is not valid for such finite lattices.

For infinite lattices we expect the AT model to exhibit phase transitions. It has been pointed out[6] that from symmetry consideration alone two transitions are expected. Only when the two lower energies of ϵ_1, ϵ_2, and ϵ_3 become equal, do we expect the two transitions to coalesce into a single one. Furthermore, there are definitely two transitions in the parameter subspaces such as $K_3=0$ or $\epsilon_0+\epsilon_3=\epsilon_1+\epsilon_2$. These are the basic ingredients in the consideration of Wu and Lin[17] in their construction of the critical surface for the AT model on the square lattice. We then expect the critical surface of the general AT model behave similarly in the parameter space. The critical surfaces would intersect at some single-transition trajectories lying in the sectional planes

$$\omega_i=\omega_j>\omega_k,\ i,\ j,\ k \text{ distinct}. \tag{14}$$

The trajectories are straight lines for square lattice[17] and, very plausibly, parabolas for triangular and honeycomb lattices.[18,19] The three single-transition lines meet at the point $\omega_1=\omega_2=\omega_3=\omega^*$, where ω^* is the critical point of the 4-component Potts model.

*Supported in part by National Science Foundation Grant DMR 72-03213 A01.

[1]J. Ashkin and E. Teller, Phys. Rev. **64**, 178 (1943).
[2]R.B. Potts, Proc. Cambridge Philos. Soc. **48**, 106 (1952).
[3]H.N.V. Temperley and E.H. Lieb, Proc. R. Soc. London A **322**, 251 (1971).
[4]R.J. Baxter, S.B. Kelland, and F.Y. Wu, J. Phys. A **9**, 397 (1976).
[5]R.J. Baxter, J. Phys. C **6**, L445 (1973).
[6]F. Wegner, J. Phys. C **5**, L131 (1972).
[7]C. Fan, Phys. Lett. A **39**, 136 (1972).
[8]G.H. Wannier, Rev. Mod. Phys. **17**, 50 (1945).
[9]F.Y. Wu and Y.K. Wang, J. Math. Phys. **17**, 439 (1976).
[10]Cf., e.g., Figs. 4 and 5 of Ref. 8.
[11]F.Y. Wu, Adv. Math. (to be published).
[12]See, e.g., F.Y. Wu, Phys. Rev. B **4**, 2312 (1971).
[13]Cf. Fig. 7 of Ref. 4.
[14]R.J. Baxter, Ann. Phys. (N.Y.) **70**, 193 (1972).
[15]Cf. Fig. 8 of Ref. 4.
[16]A duality relation similar to (4) can indeed be written down for finite Ising models with toroidal boundary conditions. But this invokes an additional dual spin involving some three-spin interactions.
[17]F.Y. Wu and K.Y. Lin, J. Phys. C **7**, L181 (1974).
[18]I.G. Enting, J. Phys. A **8**, L5 (1975).
[19]F.Y. Wu, J. Phys. C **10**, L23 (1977).

Periods on manifolds, quantization, and gauge

R. M. Kiehn

Department of Physics, University of Houston, Houston, Texas 77004
(Received 27 April 1976)

It is suggested that the quantization of flux, charge, and angular momentum be interpreted as a set of independent natural concepts which physically exhibit certain topological properties of the fields on a space–time manifold. These quantum, or topological, properties may be described in terms of one-, two-, and three-dimensional periods, respectively. In terms of this viewpoint, topological constraints between the one-, two-, and three-dimensional periods can be put into correspondence with various gauge theories. If a dynamical system is to be nondissipative, in the sense that its one-, two-, and three-dimensional topological periods are reversible invariants of the motion, then it is proved herein that the dynamical field **V** must be a Hamiltonian vector field, the field currents must be proportional to **V**, and the Lagrangian difference between the elastic and inertial energy density must be twice the interaction energy density, respectively.

INTRODUCTION

It is the purpose of this article to display certain topological properties of physical theories and to demonstrate that the notions of quantized flux, charge, and spin can be interpreted in terms of independent topological ideas of fields built on manifolds. The fact that these quantities are in relation to the integers then follows from Brouwer's theorem about closed manifolds. Brouwer's theorem effectively states that values of the closed integrals on a closed oriented manifold are integer multiples of some smallest value (see Appendix A).

A manifold is a nice mathematical object in that the difficulties of singularities, in a sense, have been removed. The regions of difficulty, which lead to phrases like "fields which are zero almost everywhere," can be eliminated, but at the expense of changing the topology of the domain over which the fields are defined. Topology of a base space can be defined in terms of a system of constrained differential forms. (See Appendix B). Each form can be split into two parts, one of which is closed in the exterior derivative sense. This closed piece can further be decomposed into a part which is exact and integrable and another part which is not. The closed integral of an exact form is zero, but the closed integral of a closed, but not exact, form is not necessarily zero on a non-Euclidean manifold (see Appendix C). (In fluids, this effect implies the existence of circulation without vorticity.) The total closed portion of a form is defined to be the gauge of the form. This definition widens the usual concept of gauge to include more than just an exact contribution. The number of distinct closed integrals over the gauge parts of a form (the periods) depends upon the manifold topology. In this theory it appears that the quantization of flux concerns the periods of a fundamental 1-form, $\mathcal{A}$; the quantization of charge concerns the periods of a fundamental 2-form density, $\mathcal{H}$; the quantization of spin concerns the periods of a 3-form density, $\mathcal{S}$. The distinction between forms and form densities (as well as *pair* and *impair* forms) becomes evident with respect to the discrete symmetry transformations such as parity and time reversal.

Each quantization concept is independent, and this observation is a point of departure from the Dirac and Schwinger magnetic monopole theories, which, in a sense, are conditional theories, in that they claim that charge is quantized if flux, or angular momentum, is quantized. Any such relationship between the one- two-, and three-dimensional periods must be considered to be a topological constraint, and it is demonstrated below that these constraints form the basis of various gauge theories. The manifold approach indicates that flux and charge can be quantized even though $\mathrm{div}\mathbf{B} = 0$ everywhere on the manifold. (In a sense, these ideas lead to the geometrodynamic-like concept of a monopole without a monopole).

Of additional interest is the clarification of the fact that physical content is to be associated with potentials, counter to the usual historical claim that the potentials have no physical significance. Physical significance is to be associated not with an individual potential, but with the topological equivalence class to which it belongs. The quantum numbers of flux, charge, and spin relate to the one-, two-, and three-dimensional equivalence classes of periods, e.g., of integrals of closed forms over closed chains. In Euclidean space (or better, Euclidean topology) there exist no nontrivial periods and therefore, from the topological point of view, no quantization. The hypothesis of a Euclidean base space for classical field theories therefore precludes any notion of topological quantization. Classical theories contain within them the germs of quantization which become nontrivially evident when the domains of definition are not based on Euclidean topology.

In that which follows, the topology of fields built on a four-dimensional space—time will be studied in terms of a set of fundamental differential forms. For each physical system it is assumed that there exists such a set of fundamental forms, and the ensuing analysis then is applicable to all physical systems which are usually described in terms of a field theory. The presentation of the abstract theory will be given first in the more familiar language of electrodynamics; this presentation then will be followed by a hydrodynamic application which demonstrates the relationship between topological constraints and the Navier—Stokes equations.

 Copyright © 1977 American Institute of Physics

THE FUNDAMENTAL FORMS

For a rapid expose of the theory, consider the formulation of electromagnetism as given by Toupin,[1] in which it is assumed that a four-dimensional space–time supports a 2-form $\mathcal{J}$ of electric intensity and an $(N-1=3)$-form density, $\mathcal{J}$, of electric current. The arguments that $\mathcal{J}$ is conserved imply that $\mathcal{J}$ is closed. The idea of charge neutrality implies that $\mathcal{J}$ is exact, for then integrals of $\mathcal{J}$ over compact, oriented three-dimensional manifolds must vanish. It follows that, from de Rham's theorem, there exists an $(N-2)$-form potential density, $\mathcal{H}$, such that $d\mathcal{H}=\mathcal{J}$, which yields the second pair of Maxwell's equation involving **D**, **H**, **J**, and ρ. However, the potential $\mathcal{H}$ consists of a part which is not closed and a gauge part which is closed. That is,

$$\mathcal{H}=\mathcal{H}_0+\hat{\mathcal{H}}, \tag{1}$$

where $d\mathcal{H}_0=\mathcal{J}$, $d\hat{\mathcal{H}}=0$. The gauge of $\mathcal{H}$ is a closed form density and, if integrated over a closed 2-chain, yields a period of $\mathcal{H}$. The choice of gauge has physical significance, for the closed integrals of $\mathcal{H}$, that is, the periods of $\mathcal{H}$, are to be identified with the electric charge. Classically, charge is the topological, global, quantity of electric flux. The concept that the periods of $\mathcal{H}$ define charge is merely the abstraction of Gauss' law,

$$Q=\oiint\oiint\hat{\mathcal{H}}=\oiint\oiint\mathbf{D}\cdot d\mathbf{S}+\oiint\oiint\mathbf{H}\,dt\wedge dl. \tag{2}$$

Now the remarkable result is that all periods (integrals on closed manifolds) are integer multiples of some smallest value, by Brouwer's degree of a map theorem.[2] Hence, the idea that charge Q is quantized by the integers is a natural result of the notion that charge is a topological property of a two-dimensional gauge period. No magnetic monopole assumption is needed to provide a reason for the quantization of charge once the topological viewpoint is admitted. The nonintegrable parts of the gauge have physical content, counter to the usual claim that potentials have no physical meaning.[3]

A similar argument goes for $\mathcal{J}$, the electric intensity 2-form; the potential $\mathcal{A}$, such that $\mathcal{J}=d\mathcal{A}$, can be split into a nonclosed part and a gauge part which is closed, but which need not be exact:

$$\mathcal{A}=\mathcal{A}_0+\hat{\mathcal{A}},\quad d\hat{\mathcal{A}}=0. \tag{3}$$

As $\mathcal{J}$ is exact, $d\mathcal{J}=0$ by the Poincaré lemma, which yields the first Maxwell pair of equations involving **B** and **E** *without* any "monopole" source. The one-dimensional periods of $\mathcal{A}$ define the topological property physically called the flux quantum,

$$\Phi=\oint\mathcal{A}. \tag{4}$$

Again the Brouwer degree of a map theorem for closed manifolds asserts that these periods are quantized in terms of some smallest value. Again the choice of gauge has physical significance, as do the potentials themselves, for it is the topological properties of the gauge that yield the flux quantum. Note that $\mathcal{A}$ is a pair form, and not a form density.

The quantization of the flux (4) is related to the Bohr–Sommerfeld idea which quantized the periodic action-angle variables. Note that here the 1-form, $\mathcal{A}$, is not a single action component, but is a Pfaffian form summed over all space–time variables. Depending on the topology, many such distinct Pfaffian forms exist, each yielding a different topological period.

This author was made aware of the possibilities that charge and flux quanta were related to two-dimensional and one-dimensional period, by E. J. Post.[4] The Toupin specification of electromagnetic theory in terms of an exact 2-form $\mathcal{J}$, and an exact 3-form density, $\mathcal{J}$, was a culmination of ideas put together by Cartan, Weyl, Bateman, Murnaghan, Van Dantzig and Post, but Toupin's treatment does not consider the periods of the potentials. The relationship of quantization of the periods to an interpretation of Brouwer's theorem appears herein for the first time.

Now on a space–time of four dimensions, there also should exist a fundamental exact 4-form or 4-form density, which should carry the properties of the three-dimensional periods in space–time. The gauge parts of the associated 3-form should be quantized. Toupin[5] introduced a concept of a 4-form density of "action," but did not exploit the topological features of the idea. The present author found an exact representation for such a 4-form density in terms of the previously described fields, and interpreted the relationship between the divergence of the three potential and the 4-form density as an intrinsic transport theorem.[6] The exact 4-form density is closely related to the Lagrange density of classical field theory, save for a factor of two in the field energy densities. The explicit formula for the action is

$$\mathcal{L}=\mathcal{J}\wedge\mathcal{H}-\mathcal{A}\wedge\mathcal{J}\equiv(\mathbf{B}\cdot\mathbf{H}-\mathbf{D}\cdot\mathbf{E}-\mathbf{A}\cdot\mathbf{J}+\rho\phi)\Omega, \tag{5}$$

where Ω is the volume element $dx\wedge dy\wedge dz\wedge dt$. This 4-form density is exact with a potential $\mathcal{S}$ such the $\mathcal{L}=d\mathcal{S}=d(\mathcal{A}\wedge\mathcal{H})$, and leads to the three-dimensional period of action,

$$S=\oint\!\!\int\mathcal{S}=\oint\!\!\int\!\!\int(\mathcal{A}\wedge\mathcal{H})\equiv\oint\!\!\int\!\!\int i(\mathbf{A}\times\mathbf{H}+\mathbf{D}\phi,\ \mathbf{A}\cdot\mathbf{D})\Omega. \tag{6}$$

It is the periods of the gauge part of $\mathcal{S}$ that are of interest herein, for by Brouwer's theorem they are quantized. Note that formally the units of S are joule-sec, while the units Q are coulomb and the units of Φ are joule-sec/coulomb, for an electromagnetic representation.

The electromagnetic representation of the topological theory thereby demonstrates that the periods of a 1-form $\mathcal{A}$ can be put into correspondence with flux quanta, the periods of a 2-form density $\mathcal{H}$ can be put into correspondence with charge quanta, and the periods of the 3-form density $\mathcal{S}$ can be put into correspondence with spin or action quanta.

It is to be emphasized that by use of forms and form densities the previous equations, and those which follow, have an intrinsic significance. For any coordinate sequence abstractly labeled as (1, 2, 3, 4) and with corresponding 2-form [$(N-2)$-form density] coefficients labeled as **E**, **B** (or **D**, **H**), the intrinsic equations are invariant in form. For example, in any coordinate sys-

tem, if **B** is the sequence of coefficients associated with the (2,3), (3,1), and (1,2) coordinate pairs in the 2-form $\mathcal{J}$, and if **H** is the sequence of coefficients associated with the (1,4), (2,4), and (3,4) coordinate pairs of the $(N-2)$-form density $\mathcal{H}$, then the magnetic energy density is given exactly by the engineering format, $\frac{1}{2}\mathbf{B}\cdot\mathbf{H}$, even though the frame is not Cartesian. The results to be obtained, therefore, are independent of the choice of a reference coordinate frame when the engineering format is used.

INVARIANCE OF PERIODS

General theory

A vector field representing a transformation of the space may or may not preserve the system topology. If the vector field is a homeomorphism, then all topological properties are invariants of the transformation. In the context of this paper the pertinent topological properties are determined by the periods of one-, two-, and three-dimensions, and it is of some interest to determine the constraints on dynamical transformations which are necessary if these periods are to be individually invariant. Physically, the question may be stated as: For what transformations do the quanta of flux, charge, and spin remain invariant?

Cartan has demonstrated that an extremal vector field which reversibly leaves the one-dimensional periods invariant is represented by Hamilton's equations of motion. Cartan's concept has been formulated[7] in terms of a Lie derivative by the statement that,

$$\pounds_{\gamma\mathbf{V}}\Phi=\pounds_{\gamma\mathbf{V}}\oint\mathcal{A}=0,\quad \text{all } \gamma; \tag{7}$$

that is, it is necessary and sufficient that a vector field which leaves invariant the one-dimensional closed integral, $\oint\mathcal{A}$, independent of parametrization γ, be a Hamiltonian vector field. This work has been extended to nonhomeomorphic vector fields which indicate that the lack of topological invariance of the 1-form $\mathcal{A}$ is related to dissipation, for if $\mathcal{A}$ is mechanically represented in terms of the system momentum, then a nonzero value for (7) implies a nonzero value for the cyclic work, $\oint f_\mu\, dq^\mu$. The result (7) implies that Hamiltonian systems are necessarily adiabatic. In an electromagnetic sense, the result (7) implies that such systems conserve flux quanta.

It is also of some interest to study the invariance of the two-dimensional periods. For a vector field $\gamma\mathbf{V}$, invariance of the charge quanta implies that

$$\pounds_{\gamma\mathbf{V}}Q=\pounds_{\gamma\mathbf{V}}\oiint\mathcal{H}=\oiint\gamma i(\mathbf{V})\,d\mathcal{H}+\oiint d(\gamma i(\mathbf{V})\mathcal{H}) \tag{8}$$

$$=\oiint\gamma i(\mathbf{V})\mathcal{J}=0,\quad \text{all } \gamma.$$

Expanding the last integral in electromagnetic format yields

$$\oiint\gamma(\mathbf{J}-\rho\mathbf{V})\cdot d\mathbf{S}+\oiint\gamma(\mathbf{J}\times\mathbf{V})\,dt\wedge dl=0. \tag{9}$$

If this result is to be true for all periods for all γ, then the integrand must vanish (by deRham's theorem). Hence it is necessary for invariance of the two-dimension periods that the current density be proportional to the velocity field:

$$\mathbf{J}=\rho\mathbf{V}. \tag{10}$$

This requirement is generally *assumed* in most magneto-hydrodynamic treatments, *a priori*, which then implies that a flow which starts without two-dimensional periods, remains without two-dimensional periods. Conversely, when $\mathbf{J}\neq\rho\mathbf{V}$, then two-dimensional periods can be created or destroyed. This concept is related to another form of dissipation, intrinsically different from the dissipation associated with the one-dimensional periods (in the electromagnetic problem it takes work to separate charge). In a hydrodynamic context it is not apparent that all solutions to the Navier–Stokes equations yield flow vector fields for which the two-dimensional periods remain invariant. The seat of hydrodynamic instabilities will be related to the creation of one-, two-, and three-dimensional periods.

Similarly, the invariance of the three-dimensional periods may be examined by means of the action of the Lie derivative on S, and exhibited in electromagnetic format as

$$\pounds_{\gamma\mathbf{V}}S=\pounds_{\gamma\mathbf{V}}\oiint\mathcal{A}\wedge\mathcal{H}=\oiint\gamma i(\mathbf{V})\,d(\mathcal{A}\wedge\mathcal{H})$$

$$=\oiint[(\mathbf{B}\cdot\mathbf{H}-\mathbf{D}\cdot\mathbf{E})-(\mathbf{A}\cdot\mathbf{J}-\rho\phi)]\gamma i(\mathbf{V})\Omega. \tag{11}$$

If all three-dimensional periods are to be flow invariants independent of γ, then the integrand vanishes, putting a constraint on the magnetic and electric energy density difference. For a system without interactions, then, the field Lagrangian must vanish for invariance of the spin quanta. Electromagnetically such a result is approximated by traveling waves for which **E** and **B** are in phase. Note that for standing waves the invariance result is only valid in a time average sense. (This notion is reminiscent of Wheeler's geometrodynamic froth.) The invariance of the flux quanta (one-dimensional) periods and charge quanta (two-dimensional) periods does not (necessarily) guarantee the invariance of the angular momentum quanta (three-dimensional) periods.

Summing up the topological invariance properties in terms of three fundamental theorems, it is to be noted that:

1. Flux quanta reversible invariance requires Hamilton's equations of motion.

2. Charge quanta reversible invariance requires $\mathbf{J}/\rho=\mathbf{V}$.

3. Angular momentum reversible quanta invariance requires $(\mathbf{B}\cdot\mathbf{H}-\mathbf{D}\cdot\mathbf{E})-(\mathbf{A}\cdot\mathbf{J}-\rho\phi)=0$.

Parametrization and reversibility

The above discussion describes topological invariance theorems with respect to vector fields $\{\mathbf{V},1\}$ parametrized by γ, where γ is an arbitrary function on the domain. In a space of N dimensions, vector fields have N components, any one of which may be selected as a parametrization function, leaving, by division, a local flow field $\gamma\{\mathbf{V},1\}$. For γ completely arbitrary, it is to be noticed that the inversion $\mathbf{V}\to-\mathbf{V}(\gamma=1\to\gamma=-1)$

changes the direction of the flow field. Invariance w.r.t. all γ implies that the object is not only invariant with respect to propagation down the flow lines, but also the invariance concept is independent of the propagation direction. Not all systems admit such a reversibility invariance.

For example, suppose γ is chosen to satisfy the continuity requirement that the volume element be invariant, then

$$\pounds_{\gamma \mathbf{V}} \Omega = 0, \tag{12}$$

or

$$[d\gamma/dt + \gamma \operatorname{div}\mathbf{V}]\Omega = 0. \tag{13}$$

Then, if γ is continuous, solutions to (13) indicate that γ is one sign, either positive or negative: Even though div$\mathbf{V}$ may vary between positive and negative values, γ never goes through zero. Note that (13) implies a dual interpretation for γ. In addition to its interpretation as a parametrization function for the vector field $\mathbf{V}$, it may be interpreted also as a measure function for Ω.

For a particular choice of γ, the invariance of one-, two-, and three-dimensional periods implies

$$\pounds_{\gamma \mathbf{V}} \oint \mathcal{A} = 0, \quad \Rightarrow i(\mathbf{V}) d\mathcal{A} = dP/\gamma, \tag{14}$$

$$\pounds_{\gamma \mathbf{V}} \oiint \mathcal{H} = 0, \quad \Rightarrow i(\mathbf{V}) d\mathcal{H} = d\theta/\gamma, \tag{15}$$

$$\pounds_{\gamma \mathbf{V}} \oiiint \mathcal{S} = 0, \quad \Rightarrow i(\mathbf{V}) d\mathcal{S} = d\alpha/\gamma, \quad \text{fixed } \gamma. \tag{16}$$

The forms P, θ, and α are respectively zero-, one-, and two-dimensional fields which are completely arbitrary at this level of analysis. By examing the requirement for invariance of the one-dimensional periods, the function P can be identified as the *pressure* addition to Hamiltonian field. For $\mathcal{A} = p_\mu dq^\mu - H(p,q,t)dt$ on a state space of $2n+1$ dimensions, the modified Hamiltonian equations of motion become, for the vector field $\gamma\{\mathbf{v}, \mathbf{f}, \mathbf{1}\}$,

$$\mathbf{v} = +\partial H/\partial p_\mu + (1/\gamma)(\partial P/\partial p_\mu), \tag{17}$$

$$\mathbf{f} = -\partial H/\partial q^\mu - (1/\gamma)(\partial P/\partial q^\mu). \tag{18}$$

Such a field leaves the one-dimensional periods, $\oint \mathcal{A}$, invariant for a given choice of parametrization, γ. Similar arguments may be constructed for the other pressurelike terms, θ and α.

The set of continuously parametrized vector fields which leave the periods invariant fall into two classes, those with positive parametrizations and those with negative parametrizations. The intersections of these two sets are those vector fields that satisfy the Hamilton equations of motions and yield absolute relative invariants which are independent of propagation direction. The compliment of the intersection forms two disjoint sets which produce relative integral invariants that are sensitive to the direction of propagation. It appears that a positive choice of γ (mechanically interpreted in its measure sense as a *mass* density) selects one of these disjoint sets such that in dissipative (irreversible) systems at least one of the periods is not preserved as an invariant. The concept of an entropy increase may be put into correspondence with this lack of period invariance in the forward (predictive) direction. It is interesting to observe that a statistical concept like pressure is related to the lack of reversibility of a topological period. The relationship between specific choices of the parametrization function and the concept of orientation is described in Appendix D.

The topological ideas presented above do not yield information about the magnitude of the quanta (why is $Q = 1.6\times10^{-19}$ coulombs, why is $h = 6.67\times10^{-34}$ joule-sec?), but they do indicate that the magnitudes of the Φ, Q, and S quanta can be interrelated, especially for those cases where separability permits,

$$\oiiint \mathcal{S} = \oiint \mathcal{A} \wedge \mathcal{H} = \oint \mathcal{A} \oiint \mathcal{H}. \tag{19}$$

Then the flux quantum, charge quantum, and angular momentum quantum are not independent; the separability concept in effect imposes a topological constraint on the one-, two-, and three-dimensional gauges, in the sense of a gauge theory. Indeed, the Lagrange action potential $\mathcal{S}$, built from the product $\mathcal{S} \equiv \mathcal{A} \wedge \mathcal{H}$, is yet another gauge theory[6] relating periods of $\mathcal{A}$, $\mathcal{H}$, and $\mathcal{S}$. Several classical gauge theories which impose certain constraints on the cohomology of space—time are described in the next section. These are gauge theories of the second kind and are associated with restricted nonhomogeneous contact transformations. Gauge theories of the first kind based upon inhomogeneous contact transformations will be ignored in this article, except for the fact that they may be generalized to include the effects of dissipation. In certain cases gauge theories of the first kind are equivalent to gauge theories of the second kind. For example, it may be true that $\mathcal{A}\exp(i\theta) \equiv \mathcal{A} - d\psi$. However, such an equivalence requires that $\mathcal{A}$ be integrable; i.e., $\mathcal{A} \wedge d\mathcal{A} = 0$, which is certainly not the case for the most general physical system. The concept of integrability will be shown below to be related to the notion of when the Lagrange and Hamiltonian formalisms are equivalent.

GAUGE THEORIES

Classical developments

The recognition that the vector spaces of the fundamental forms $\mathcal{J}$ or $\mathcal{H}$, and $\mathcal{A}$, $\mathcal{J}$, or $\mathcal{S}$ are of the same dimensions leads to the notion that a map between the spaces of similar dimensionality effectively makes a statement about a choice of one-, two-, and three-dimensional gauges, and therefore influences a choice of quantization. It is the purpose of this section to demonstrate the topological basis for several historical gauge theories.

One of the earliest gauge theories was due to Maxwell who utilized the idea of a constitutive map,

$$X: \mathcal{J} \to \mathcal{H}, \tag{20}$$

to describe the electromagnetic properties of matter. In view of the fact that $\mathcal{J}$ is exact, but $\mathcal{H}$ need not be exact, the Maxwell constitutive map may be considered to be a gauge theory in that it specifies something about the two-dimensional periods of $\mathcal{H}$, or the quantized charge. Historically this observation was not made.

Somewhat later Lorentz utilized the idea that charge-

current meant particles in motion, and so developed a theory that maps the current $\mathcal{J}$ into the 1-form $\mathcal{A}$,

$$L:\mathcal{J}\to\mathcal{A}. \tag{21}$$

The Lorentz map effectively makes statements about the gauge or one-dimensional periods of $\mathcal{A}$. To update the theory, consider a metrically based constitutive map $X/Y_0=\sqrt{g}(g^{\mu\alpha}g^{\nu\beta}-g^{\mu\beta}g^{\nu\alpha})$, an idea which is an extension of that introduced by some of the early proponents of the use of tensor analysis in physics. The factor Y_0 represents the functional admittance of space—time. Its use permits the constitutive map, X, to be equated with the Hodge dualization operator, $*$, such that $\mathcal{H}=Y_0*\mathcal{J}$. This idea gives Lorentz's classic wave-equation-with-source for the currents and potentials, when the generalized differential form identity,

$$\Box^2\mathcal{A}=(d\delta+\delta d)\mathcal{A}=*\mathcal{J}/Y_0+d\delta\mathcal{A}-*\mathcal{H}\ dY_0/Y_0^2, \tag{22}$$

is constrained by the Lorentz gauge condition, now explicitly stated as a topological requirement,

$$\delta\mathcal{A}=0, \tag{23}$$

and the additional requirement that the admittance, Y_0, be constant over the domain. Note that the Lorentz theory makes a statement about the one-dimensional periods and the flux quanta. On a compact orientable manifold (23) requires by the Hodge—deRham decomposition theorem that $\mathcal{A}$ contain no exact part. In hydrodynamics, the Lorentz gauge is related to the isochoric condition, commonly called (incorrectly) incompressibility.

Closed, but nonexact gauges

Most of the early gauge theories were not concerned with quantization, and ignored the latent possibilities associated with the periods of the gauge. Such a statement is not applicable to the theory of London[8] which effectively analyzed the theory of super conductivity and the flux quantum in terms of one-dimensional periods and introduced the gauge condition,

$$f:\mathcal{J}\to\mathcal{A},$$

where explicitly

$$f:\mathcal{A}=-*(\mathcal{J}/\rho). \tag{24}$$

For ρ a constant, this gauge implies that

$$\mathcal{J}=d\mathcal{A}=-d(*\mathcal{J}/\rho), \tag{25}$$

or for ρ equal to the London parameter, λ^{-1},

$$\mathbf{B}=-\operatorname{curl}(\lambda\mathbf{J}) \tag{26}$$

and

$$\mathbf{E}=+\partial(\lambda\mathbf{J})/\partial t. \tag{27}$$

A related and more modern development states that $\mathcal{A}$ and $*\mathcal{J}/\rho$ are cohomologous; i.e., let

$$f:\mathcal{A}=-*\mathcal{J}/\rho+d\theta. \tag{28}$$

As $\mathcal{J}$ is exact, it satisfies the equation of continuity,

$$d\mathcal{J}=0, \tag{29}$$

with its usual local representation

$$\partial\rho/\partial t+\operatorname{div}\mathbf{J}=0. \tag{30}$$

By defining $\Psi=\sqrt{\rho}\exp(i\theta/\hbar)$, the gauge condition and the equation of continuity imply a complex root,[9]

$$i\hbar\partial\Psi/\partial t=(1/2m)(-i\hbar\nabla-q\mathbf{A})^2\Psi+q\phi\Psi, \tag{31}$$

which is Schrödinger's equation for a charged particle moving in an electromagnetic field given by the potential $(\mathbf{A},\phi)$. Remarkably, the phase function θ of quantum theory appears to be related to the concept of the cohomology function between the 1-form $\mathcal{A}$ and the Hodge dual, $*\mathcal{J}/\rho$.

Specific parametrizations and gauge theories

A remarkable feature of the invariance of two-dimensional periods is noted if a specific choice of parametrization of $\mathbf{V}$ is made to insure continuity. That is, choose a *mass* density γ such that $\gamma\mathbf{V}$ has zero divergence on the space. Then, with respect to this parametrized curve, the two-dimensional periods will be invariant if

$$\gamma i(\mathbf{V})\mathcal{J}=d(\mathcal{Z}), \quad \text{where } \mathcal{Z} \text{ is some } (N-3)\text{—form density.} \tag{32}$$

In Cartesian coordinates

$$\gamma(\mathbf{J}-\rho\mathbf{V})=\operatorname{curl}\mathbf{Z} \tag{33}$$

and

$$\gamma(\mathbf{J}\times\mathbf{V})=\operatorname{grad}\phi+\partial\mathbf{Z}/\partial t. \tag{34}$$

By letting $\mathbf{Z}$, $\phi\to 0$, the London gauge is retrieved, and by letting $\operatorname{curl}\mathbf{Z}=\operatorname{grad}\theta$, the Feynman gauge is obtained. It may be concluded that these two gauge theories (and hence the Schrödinger quantum theory) are gauge theories that leave invariant the two-dimensional periods of charge quanta, but not necessarily in a reversible manner.

If in four dimensions the gauge field $\mathcal{Z}$ is taken to be proportional to the 1-form $-\lambda(\mathcal{A})$, then

$$\mathbf{J}=\rho\mathbf{V}-\lambda/\gamma(\partial\mathbf{A}/\partial t+\operatorname{grad}\phi), \tag{35}$$

which is both of the Ohmic format

$$\mathbf{J}=\rho\mathbf{V}+\sigma\mathbf{E}, \tag{36}$$

and related to the requirement that $\{\mathbf{V},1\}$ be a characteristic vector field.

Hence it is proved that Ohm's law for the current preserves two-dimension periods, but not reversibly. Ohmic dissipation does not involve topological variations of two-dimensional periods (but it is a dissipative mechanism that does involve topological variation of one- and three-dimensional periods). It is to be noted that only in four dimensions it is possible to use $\mathcal{Z}=-\lambda\mathcal{A}$, for only in four dimensions does $\mathcal{Z}$, an $(N-3)$-form density, have the vector space qualities of a 1-form, $\mathcal{A}$.

It is of physical interest to ask when two-dimensional periods are not topological invariants. In electromagnetism it is observed that friction (somehow) produces two-dimensional periods, and bombardment of matter with γ rays of more than 1.02 MeV produces two-dimensional periods. In hydrodynamics, the notion of two-dimensional periods has yet to be experimentally identified, but it is suspected that they are related to (non-

steady) dissipative mechanisms. The dissipative effects of radiation are associated with change of three-dimensional periods (spin). The dissipative effects of nonadiabatic processes are related to the change of one-dimensional periods (flux). These dissipative mechanisms intrinsically are distinct from those changes of two-dimensional periods—changes which must represent other dissipative mechanisms (See Appendix E). The relationship between forms, topology, and thermodynamics will appear elsewhere. Herein the pertinent theme to be remembered is that dissipation implies noninvariant periods, which in turn implies changing topology.

A HYDRODYNAMIC APPLICATION

The initial development of this theory of periods was in the language of electromagnetism for which it is well known that the desired 1-form is action per unit charge. It would seem natural for hydrodynamic systems to develop the theory in terms of a fluid action per unit mass. Hence from the 1-form,

$$\mathcal{A} = u\,dx + v\,dy + w\,dz - h\,dt, \tag{37}$$

it is possible to develop over space–time the fundamental forms $\mathcal{E}$, H, $\mathcal{J}$, $\mathcal{A}\wedge H$, etc., and study the topological properties of such systems. Indeed, for a spatially Euclidean metric it is possible to demonstrate[10] how the Navier–Stokes equations take the form of a constraint relation:

$$\pounds_{\mathbf{v}}\mathcal{A} = -dP/\rho + \nu\delta\, d\mathcal{A} + d\theta. \tag{38}$$

By action of the exterior derivative on (38), the basic Helmholtz equation for vorticity becomes

$$\pounds_{\mathbf{v}}\mathcal{J} = +(1/\rho^2)\,d\rho\wedge dP + \nu\, d\delta\mathcal{J}, \tag{39}$$

and demonstrates the effect of viscosity and equation of state on the generation of vorticity.

Every 1-form may be adjusted by a gauge transformation (of the second kind) to its cohomological equivalent, $\mathcal{A}$-$d\theta$. Furthermore, the dual of $(\mathcal{A}$-$d\theta)$ is an $(N-1)$-form which locally satisfies the conditions of integrability, hence admits an integrating factor β such that the "current,"

$$\mathcal{J}_m \triangleq \beta\{*(\mathcal{A} - d\theta)\}, \tag{40}$$

is closed. The following questions arise: Is there a global integrating factor? Can the integrability of $\mathcal{A}$ be accomplished by the gauge transformation, $d\theta$, alone {implying that β is a constant, or $d\beta$ is orthogonal to $*(\mathcal{A} - d\theta)$}? If $\mathcal{J}_m$ is closed, is it exact? Is $\mathcal{J}_m$ a topological constraint such that $\mathcal{J}_m = \mathcal{J} \equiv d^* d\mathcal{A}$?

As a first example, consider the two-dimensional fluid ($w=0$) on space–time endowed with a metric $[1,1,1,-c^2/n^2]$. For simplicity, define $Y_0 n/c = \epsilon = (n^2/\mu c^2)$. Assume that the specific Hamiltonian h is given by the formula

$$h = [c^2 - \nu\partial(\ln\zeta)/\partial t], \tag{41}$$

where the $\{x,y,z\}$ components of the vorticity are given by the usual notation, $\mathrm{curl}\mathbf{V} = \{\xi,\eta,\zeta\}$. Now make the gauge transformation, based on the phase function

$$\theta = \nu \ln\zeta, \tag{42}$$

and use the z component of vorticity, ζ, as the integrating factor β to yield

$$\mathcal{J}_m = \zeta\{*[(u - \partial\theta/\partial x)\,dx + (v - \partial\theta/\partial y)\,dy - c^2\,dt]\}. \tag{43}$$

By assuming that n and c are constants, the closure of $\mathcal{J}_m$, $d\mathcal{J}_m = 0$, yields

$$\partial\zeta/\partial t + u\partial\zeta/\partial x + v\partial\zeta/\partial y + \zeta(\partial u/\partial x + \partial v/\partial y) = \nu(\partial^2\zeta/\partial x^2 + \partial^2\zeta/\partial y^2), \tag{44}$$

which is precisely the Navier–Stokes–Helmholtz equation for the vorticity of a two-dimensional, compressible, viscous fluid, and demonstrates how the topologically based gauge theory is applicable to hydrodynamics.

It is to be noted that even for the three-dimensional case, if the phase function θ is chosen to be such that

$$(n/c)^2(h + \partial\theta/\partial t) = 1, \tag{45}$$

then the closure of $\mathcal{J}_m$ is always an equation of continuity, such that the volume measure $\Omega = \tilde{\beta}\,dx\wedge dy\wedge dz\wedge dt$ is an invariant of the flow field given by

$$\bar{\mathbf{V}} = \{u - \partial\theta/\partial x,\ v - \partial\theta/\partial y,\ w - \partial\theta/\partial z,\ 1\}. \tag{46}$$

The density $\tilde{\beta}$ is defined as $\beta\sqrt{g} = \beta(ic/n)$. If the gauge is chosen such that $\bar{\mathbf{V}}$ is isochoric, then

$$\nabla^2\theta = \mathrm{div}\mathbf{V}, \tag{47}$$

but such a choice does not necessarily lead to the concept of incompressibility, for the closure of $\mathcal{J}_m$ and the constraint given by (45) yield the result that

$$d\beta/dt = \partial\{1 - (h + \partial\theta/\partial t)(n^2/c^2)\}(\beta c/n)/\partial t. \tag{48}$$

The density "β" is incompressible only if $d\beta/dt$ vanishes, which is not necessarily true, even though the fluid is isochoric.

Again focus attention on a two-dimensional fluid, and ask the question: Is $\mathcal{J}_m$ integrable? The answer, subject to (45), is always yes in two dimensions, which admits the concept of a stream function ψ. Consider the 2-form density

$$-\mathcal{G}/(ic/n) = (1/2)(\{x\,dy\wedge dz - y\,dx\wedge dz\} - \psi\,dt\wedge dz). \tag{49}$$

The exterior derivative of $\mathcal{G}$ yields the 3-form density (for constant c/n)

$$-d\mathcal{G} = (ic/n)(dx\wedge dy\wedge dz + \partial\psi/\partial x\,dx\wedge dx\wedge dt + \partial\psi/\partial y\,dy\wedge dz\wedge dt). \tag{50}$$

Direct comparison to $*(\mathcal{A} - d\theta)$ yields the standard representations,

$$u = +\partial\theta/\partial x + \partial\psi/\partial y, \tag{51a}$$

$$v = +\partial\theta/\partial y - \partial\psi/\partial x, \tag{51b}$$

for the velocity components in terms of the "potential" function θ and the "stream" function ψ. Note that

$$\mathrm{div}\mathbf{V} = \nabla^2\theta, \tag{52a}$$

and

$$\zeta = (\mathrm{curl}V)_z = -\nabla^2\psi. \tag{52b}$$

The remarkable feature of two dimensions is that the

viscosity of the fluid and its associated dissipative features may be transformed away by means of a dilitation [compare (42) and (52b)]. This effect has been observed in other (mechanical) dissipative systems.[11] (Consider the simple example as given by the damped harmonic oscillator, followed by a transformation to a space whose coordinates are exponentially shrinking with time. In the shrinking coordinate system, the equation of motion is that of an undamped oscillator.)

The two-dimensional stream function permits the evaluation of $\mathcal{J}$ as

$$\mathcal{J} = d\overline{\mathcal{A}} = d(\mathcal{A} - d\theta) = -\nabla^2\psi(dx\wedge dy) + (\partial^2\psi/\partial y\,\partial t)\,dt\wedge dx + (-\partial^2\psi/\partial x\,\partial t)\,dt\wedge dy, \tag{53}$$

which allows the z component of vorticity, $\zeta = -\nabla^2\psi$, to be compared with the z component of the electromagnetic **B**-field. The dual of $\mathcal{J}$ is the 2-form density H:

$$H = Y_0^*\mathcal{J} = i\{(1/\mu)(-\nabla^2\psi)\,dz\wedge dt - (\epsilon)(\partial^2\psi/\partial y\,\partial t)\,dy\wedge dz - (\epsilon)(\partial^2\psi/\partial x\,\partial t)\,dz\wedge dx\}, \tag{54}$$

from which the "electromagnetic" current density 3-form $\mathcal{J}$ can be derived (for constant velocity, ϵ, and μ) in terms of a Hamiltonian format built on a function, χ, equal to the result of the application of the wave operator on ψ.

That is, define

$$\chi = \nabla^2\psi - (n/c)^2\,\partial^2\psi/\partial t^2, \tag{55}$$

then

$$\mathcal{J} = -(i/\mu)\{(\partial\chi/\partial y)\,dy\wedge dz\wedge dt - (-\partial\chi/\partial x)\,dx\wedge dz\wedge dt\}. \tag{56}$$

Note that the free charge density ρ vanishes identically.

If χ is proportional to ψ such that (55) becomes the Klein—Gordon equation, then (56) indicates that the electromagnetic current is proportional to the "rotational" components of the velocity. Therefore, it is to be observed that the topological constraint that $\mathcal{J}_m = \mathcal{J}$ is related to the constraint that the stream function, ψ, satisfies the Klein-Gordon equation.

For all cases, the form $\mathcal{J}\wedge H$ equals

$$\mathcal{J}\wedge H = i[(1/\mu)(\nabla^2\psi)^2 - \epsilon\{(\partial^2\psi/\partial x\partial t)^2 + (\partial^2\psi/\partial y\,\partial t)^2\}]\,dx\wedge dy\wedge dz\wedge dt. \tag{57}$$

Except for very rapid variations of the stream function, the energy density of the fluid tends to be magneticlike. That is, the fluid field energy density is dominated by the vorticity and not by the fluid accelerations. Therefore, there exist Lorentz transformations which can eliminate the accelerations (**E** fields) entirely. A critical point in the fluid flow is reached when the energy density becomes electriclike.

The topological periods of one, two and three dimensions are determined by the forms $\mathcal{A}$, H, and $\mathcal{A}\wedge H$. The periods of $\mathcal{A}$ are determined entirely by the stream function, ψ, as are the periods of H. However, the periods of $\mathcal{A}\wedge H$ depend upon the choice of the gauge function θ as well.

It is readily observed that if the fluid is steady there do not exist any timelike ($dt = 0$) two-dimensional periods of H or three-dimensional periods of $\mathcal{A}\wedge H$. The one-dimensional periods are related to nonzero values of the circulation integral which can exist for steady (streamline) flows. Timelike two- and three-dimensional periods require that the flow be nonsteady.

Note, however, that for a fluid governed by the Navier—Stokes equations [set $\theta = \mathcal{L} + \nu\delta\mathcal{A}$ in (38)], the choice of parametrization $\gamma = \rho$ reduces the intrinsic equations of motion to the statement,

$$\mathcal{L}_{\rho\mathbf{V}}\mathcal{A} = d(\rho\mathcal{L} - P) + \nu(d\delta + \delta d)\mathcal{A}. \tag{58}$$

For the inviscid fluid this (common) choice of parametrization implies that the one-dimensional periods are invariant w.r.t. the flow $\rho\mathbf{V}$, for all **V**. In the atmosphere where the dominant mechanism for vorticity creation is due to a nonbarytropic equation of state (and where viscosity is often neglected) this result implies that the one-dimensional periods are invariant along the flow lines of the momentum flux $\rho\mathbf{V}$ (which, of course, are different from the flow lines of **V**). Only viscosity effects can change the vorticity, or the one-dimensional periods, of the fluid along the flow lines of $\rho\mathbf{V}$, the momentum flux.

SUMMARY

This article demonstrates that topological concepts of periods on manifolds may be put into correspondence with the quantized concepts of flux, charge, and spin, each in an independent manner and without direct utilization of a quantum hypothesis. Physically important information is contained in the potentials in that the specification of a gauge selects an equivalence class in the topological sense. Physical theories of matter can be put into correspondence to various gauge theories relating or defining the one-, two-, and three-dimensional periods supported by space—time. Processes without dissipation leave the topological periods invariant. Schrödinger's quantum theory can be interpreted as a gauge theory that leaves the two-dimensional periods invariant. Hamiltonian theories leave the one-dimensional periods invariant. On the other hand, there must exist three (independent) modes of dissipation depending upon whether or not the process individually modifies the one-, two-, or three-dimensional periods of the system.

ACKNOWLEDGMENT

Funds for the partial support of this study have been allocated by the NASA—Ames Research Center, Moffett Field, California under interchange NCA-2-OR295-602.

APPENDIX A: THE BROUWER DEGREE OF A MAP THEOREM

Several proofs of the Brouwer theorem may be found in texts[2] on topology. The purpose of this appendix is to demonstrate its use in space—time situations. Consider an initial space M of coordinates x^A and a final space N of coordinates y^μ, with a transitional map ϕ from M to N that need not be a homemorphism. Suppose $\mathcal{A} = A_\mu\,dy^\mu$ is a 1-form on N which may be retrodicted to M by the pullback, $\phi^*\mathcal{A}$. Now suppose a closed curve

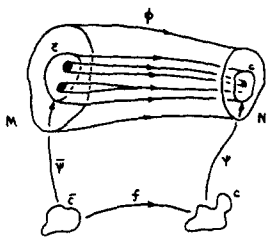


FIG. 1. Brouwer's degree of a map theorem applied to hydrodynamic dissipative systems.

$\bar{C}$ (one-dimensional manifold) may be immersed into M by $\bar{\psi}$ and another closed curve C may be immersed into N by ψ. Then a map f between $\bar{C}$ and C will have a degree such that one-dimensional integrals are related by Brouwer's theorem,

$$\int_{\bar{C}} \bar{\psi}^*(\phi^* A) = \deg f \int_C \psi^* A, \tag{A1}$$

and $\deg f$ is an integer. If the image of $\bar{C}$ in M encircles a number of holes in M different from the number of holes encircles by C in N, then that information will appear in the degree of the map, f. The two curves C and $\bar{C}$ are not homotopic in a space of the same dimension, but may be homotopic in spaces of higher dimension.

Now it is not necessary to think of the topological holes as literal holes in the space. A physical example of the holes may appear as regions of closed and open sets in the field built over space—time. Consider a fluid stream in which there exist zones which are curl free and zones which are not. It is possible to visualize the zones which are rotational as "holes" in the field. As the fluid flows from an initial to a final configuration, the number of rotational zones can change. The first Betti number of field space then must not be an invariant. The flow cannot (by Cartan's theorem) be described by Hamilton's equations of motion. The changing topology will be exhibited by a nondiabatic dissipation. Figure 1 portrays the situation which is described by Brouwer's theorem. It is most remarkable that dissipative, radiative transitions from initial to final states take place in discrete increments in this topological analysis. These ideas are in agreement with a quantum theory point of view, but did not invoke a quantum hypothesis.

APPENDIX B: KURATOWSKI CLOSURE AND DIFFERENTIAL FORMS

The theory of exterior differential forms should not be considered as just another formalism of fancy. It should be considered as a theory that contains both geometrical information (as does classical tensor analysis) and topological information. Consider a system Σ of forms and their exterior derivatives specified on some space. Then note that the operator

$$K = I \oplus d, \tag{B1}$$

where I is the identity map and d is the exterior derivative, when operating on subsets of the system of forms, obeys the following rules:

1. $K\phi = \phi$,
2. $KA \supset A$,
3. $K^n A = KA$,
4. $K(A \cup B) = KA \cup KB$. (B2)

These rules permit K to be identified as a Kuratowski closure operation.[12] Hence the specification of the system Σ defines a topology.

The most amazing property of differential forms is that with respect to maps which need not be invertible, and which may not have invertible Jacobians, the differential form is always retrodictable (but *not* necessarily predictable). The physicist should adopt the viewpoint that the most primitive field problem is: Given the final state, what was the initial state from which it came? This retrodictive question is more fundamental[13] than the usual predictive version: Given the initial state, what is the final state? With respect to nonhomeomorphic maps the retrodictive question is often deterministic, while the predictive question is not.

APPENDIX C: AN EXAMPLE OF A NONEXACT GAUGE FIELD

Consider the vector field,

$$\mathbf{A} = (iy - jx)/(x^2 + y^2), \tag{C1}$$

on the space $E^3 - \{0\}$, which has a hole. Now $\mathrm{curl}\mathbf{A} = 0$, everywhere; therefore, the field $\mathbf{A}$ is a gauge field. The closed integral of $\mathbf{A} \cdot d\mathbf{l}$ about the hole may be evaluated by means of the coordinate transformation $x = r\cos\theta$, $y = r\sin\theta$. The circulation integral about a loop encircling the origin is given by the expression

$$\oint A = \oint \mathbf{A} \cdot d\mathbf{l} = \oint \{y\,dx - x\,dy\}/(x^2 + y^2). \tag{C2}$$

Substitution of dx and dy in terms of r and θ yields

$$\oint A = \int_0^{2\pi} d\theta = 2\pi. \tag{C3}$$

As the circulation is not zero, $\mathbf{A}$ cannot be globally represented by the gradient of a single potential function and represents a gauge field. Under the transformation the closed loop in (x,y) space becomes an open curve in (r,θ) space.

APPENDIX D: PARAMETRIZATION, ORIENTATION, AND OTHER TOPOLOGICAL IDEAS

It appears that the continuity of the parametrization function γ is related to the topological concept of orientability. When γ is continuous (and by convention chosen to be positive), then it is possible to mechanically identify γ with the mass density ρ and a choice of orientation. On the other hand, the identification of a γ which satisfies (12) with the charge density ρ is a notion associated with a nonorientable manifold that supports both positive and negative measures. An elementary geometric example of such phenomena is given by two separated spherical surfaces, one with an outward normal field and the other with an inward normal field. First, because the sets are disconnected they cannot be represented by a single immersion, for immersions have a unique range. The questions arises, can two spherical surfaces be represented by a submersion?, and the

answer to the question is yes, if the spheres were oriented. The implicit equation $[(x-a)^2+y^2+z^2-(a/2)^2] \times[(x+a)^2+y^2+z^2-(a/2)^2]=0$ yields a closed set of two components representing two spheres of radii $a/2$ separated by $2a$. The implicit representation, however, guarantees that the surfaces are closed and *oriented*. The normal (gradient) field is outbound on *both* spheres. If *one* of the factors is multiplied by minus one, then the implicit representation gives an inbound normal field on both spheres. The nonorientable two component spheres with one normal inbound and one normal outbound *cannot* be given an implicit representation. Nature states this idea by demonstrating charge pairs with the field lines "existing" from the positive charge regions and "entering" the negative charge regions. Mass on the other hand is oriented, with field lines only existing from regions of (positive) mass.

A few remarks about the topological question of separability of the integration chains is of related physical significance. Some closed chains separate, others do not. For example the closed curves on a torus which yield cycles do not separate the space, but the shrinkable closed curves do. Sometimes a cut can be made that will separate and sometimes not. A cut along the closed "diametrical" curve on a Möbius band does not separate the space, but actually produces from a spin-$\frac{1}{2}$ representation a spin-2 representation,[14] which then may be folded into a double layered spin-$\frac{1}{2}$ representation again (fermions to gravitons to fermions??). On the other hand a cut along the nonshrinkable closed curve of a cylinder separates the space. In the physics of electromagnetic materials such phenomena often is observed. A conductor placed in an electric field is polarized. If the conductor is then cut in two, the charge is separated. On the other hand, a cut through a dielectric does not separate the charge.

Amperian magnetic needles when cut do not separate the flux quanta, which imply that such cuts do not separate and do not produce flux monopoles. The concept of a magnetic monopole has to be put in correspondence with the topological theory that there exists a one-dimensional cut which can separate the flux quanta, and yet $\mathrm{div}\mathbf{B}=0$ on the manifold. The existence of these cuts is an open question, but it is to be observed that this possible topological property of the theory does not require a source term to be appended to the 1st Maxwell set of equations ($d\mathcal{F}=0$; not $d\mathcal{F}=\mathcal{J}_m$). Flux quantization is permissable, with free flux quanta (monopoles) permissible, if separable cuts exist.

APPENDIX E: GAUGE INVARIANCE AND HAMILTONIAN THEORIES

The relationship of gauge and nonintegrability was introduced to the physics community by Weyl.[15] A most informative picture can be obtained by propagating a vector about some closed loop where infinitesimally the vector is kept "parallel" to itself. When the vector returns to its starting point, Weyl pointed out that there is no obvious reason why the length of the vector should be the same as it was when it started. The vector length happens to be length, or gauge, invariant with respect to parallel transport on Euclidean spaces, but it need not be invariant on other spaces.

A more general viewpoint notes that upon return to the starting point the final vector may be of different length, and also it may have a line of action not parallel to the initial vector. The decrement vector betwen this initial and final state may have a component in the "plane" of the curves and the initial vector, and perhaps a component out of the initial "plane." The decrement angle out of the plane is related to the torsion of the space, and the decrement angle in the plane is related to the curvature of the space. If the closed loop is defined by segments of vector fields, the propagated vector may or may not be algebraically closed with respect to these vectors. If the propagated vector, upon being transported around the loop, is not closed with respect to the vector generators of the loop, but in a sense points out of the plane of the loop, then the space has torsion. Immersed spaces do not have torsion; submersed spaces can have torsion.

Now a great emphasis is placed on the gauge invariance of physical theories, some of it in an almost mystical sense. However, according to the ideas expressed in this text the notion of gauge invariance relates to a concept of topological invariance, and this concept must be extended also to two and three-dimensional gauges.

It is to be noted that if the one-dimensional gauges (one-dimensional periods) are to be invariant with respect to a dynamical transformation generated by the field $\mathbf{V}$, for any parametrization γ, then

$$\pounds_{\gamma\mathbf{V}}\oint_{\gamma}\mathcal{A}=0. \tag{E1}$$

However, such a requirement leads to the necessary and sufficient conditions (Cartan's theorem) that the equations of motion are described by Hamilton's partial differential equations.[12]

A necessary and sufficient condition for a theory to be one-dimensionally gauge invariant (for any γ) is then that the theory be a Hamiltonian theory. It would appear that the predilection for gauge invariance possibly can be related to the demand for a quantum theory, which requires a Hamiltonian basis for quantization in the Dirac sense.

It should be recognized that a gauge invariant theory *cannot* explain the details of nonadiabatic phenomena, for dissipation implies that gauge is not invariant. As was mentioned in the text, the invariance of the one-dimensional gauges does not guarantee the invariance of the two- and three-dimensional periods, unless further constraints are put on the system. Hence, there may be dissipative effects in a physical system if the two-dimensional and three-dimensional periods are not invariant. A Hamiltonian theory may still be dissipative in the sense that the two- and three-dimensional periods may not be invariant, even though the one-dimensional periods are invariant. For those topological systems where the p- and $(N-p)$-dimensional periods are dualistically related (flux dual to angular momentum) then a Hamiltonian analysis implies both conservation of flux and angular momentum, but still does not guarantee conservation of two-dimensional periods (charge pairs can be created in an independent dissipative manner).

For hydrodynamic systems where the unit source is

mass (not charge) the integer relations of the one-dimensional periods have been experimentally observed, and the quantum of vortices in superfluids is h/m (not h/e). Physically, what the two-dimensional periods are for systems with mass as the source (gravity) is not known. Neither is there experimental evidence available at this time for three-dimensional periods in a hydrodynamic system. The relationship between creation of nonzero periods and instability concepts leads to the conjecture that these effects are to be observed in the turbulent state.

APPENDIX F: DUALITY AND PERIODS

It is known that topological periods are related to Betti numbers and that for certain spaces the Betti numbers obey the Poincaré–Alexander duality principle; that is $B_p = B_{N-p}$, on compact oriented spaces.[16] The idea of duality can be used to explain the apparent paradox between the Bohr–Sommerfeld quantization rule,

$$\oint p_\mu \, dq^\mu = n\hbar, \tag{F1}$$

which implies that mechanical action is related to a one-dimensional period, and the electromagnetic concept that action is related to a three-dimensional period [see Eq. (6)]. To resolve the paradox, consider those systems where the factorization of Eq. (19) is admissible; then the three-dimensional period of action becomes proportional to the one-dimensional period of flux mod the two-dimensional period of charge. Classically this result is expressed by the statement that the electromagnetic momentum is Q times the vector potential. Topologically, the concept is a duality constraint on the periods of dimension $p = 1$ and $N - p = 3$ for fixed values of Q, and states that the flux periods are proportional to the action periods. Under this topological constraint the three-dimensional period, or integral, of action may be reduced to a one-dimensional period, or integral, in the Bohr–Sommerfeld case.

In mechanics, a similar paradox arises when the Lagrangian approach is compared to the Hamiltonian-action theory. The question asked is: When is the Lagrange integrand, $L\,d\tau$, equivalent to the Hamiltonian action $\mathcal{A} = p_\mu \, dq^\mu - H\,dt$? If the equivalence is to be globally valid, then it follows that $\mathcal{A} \wedge d\mathcal{A} = 0$; i.e., the Hamiltonian action must be integrable. Such special cases imply that the momentum vector either has zero curl, or that the curl of **p** be perpendicular to **p**. Many physical systems admit such descriptions, but others do not. A fluid with vorticity is a classic counter example to the exact case, and a fluid with a twisted secondary flow yields an example for the nonintegrable case.

The resolution to the paradox of LaGrange–Hamiltonian action equivalence is not only related to the duality principle, but also sheds some light on the ergodic problem of when phase averages are equivalent to time averages. In the duality sense, the Lagrange function is interpreted not as the coefficient of a 1-form, but as the decomposed coefficient of a $2N$ density in $2N+1$ space:

$$\mathcal{L} = L(p,q,\tau)\, dq^1 \wedge \cdots \wedge dp_N. \tag{F2}$$

All $2N$ forms are decomposable, but they do not necessarily represent $2N$ volumes. The above format suggests that the Lagrange density approach appears to be associated with phase averages. For each $2N$ form in $2N+1$ space there exists a metric dual which is a 1-form, $\mathcal{A}$. That is,

$$*\mathcal{L} = \mathcal{A}, \tag{F3}$$

by means of the Hodge dualization operator *. The question may now be asked: Does the dual of $\mathcal{L}$ admit a single parameter (time) representation? If true, then there exists an ergodic equivalence between $\int \cdots \int \mathcal{L}$ as a phase average and $\int \mathcal{A}$ as a time average. The equivalence of the two methods requires again that the Hamiltonian action, $\mathcal{A}$, be integrable:

$$\mathcal{A} \wedge d\mathcal{A} = 0 = (*\mathcal{L}) \wedge d(*\mathcal{L}). \tag{F4}$$

The Hamiltonian–Lagrange equivalence subsumes the ergodic principle.

The reverse question is somewhat different: Suppose there exists an integrable 1-form, "$\mathcal{L}$" $d\tau = \mathcal{A}$. Is the metrically related dual $*\mathcal{A} = \sqrt{g}\, g^{0\nu} \mathcal{L}\, dq^1 \wedge \cdots \hat{dq}^\nu \wedge \cdots \wedge d\tau$ divergence free? (In order words is the "phase volume" closed?) The answer is yes if $\sqrt{g}\mathcal{L}$ is an integrating factor for the time like component of the metric $g^{0\nu}$.

A somewhat more general view considers the duality constraints between $\mathcal{S}$ and $\mathcal{A}$ in terms of a cohomology theory, and compare the closed one-dimensional integrals and the closed $2N$-dimensional integrals on a state space of $2N+1$ variables p, q, t. In a preliminary way, assume that the single-dimensional integral implies the existence of a 1-form $\mathcal{A} = p_\mu \, dq^\mu - H\,dt$ on state space, and a one-parameter map ϕ: $\tau \to p, q, t$, such that the pullback of the integral of the 1-form exists as a function on τ; i.e.,

$$\phi^* \int \mathcal{A} = \int L\, d\tau. \tag{F5}$$

In the second situation, consider a map ψ: $\bar{q}, \bar{p} \to p, q, t$ such that the $2N$-dimensional form $\mathcal{S} = \rho\, dq^1 \wedge \cdots dq^N \wedge dp_2 \wedge \cdots \wedge dp_N - J^1 dt \wedge d\hat{q}^1 \wedge \cdots \wedge dp_N + \cdots$ has a pullback image,

$$\psi^* \int \cdots \int \mathcal{S} = \int \cdots \int S\, d\bar{q}^1 \wedge \cdots \wedge d\bar{p}_N. \tag{F6}$$

Time averages are equal to phase averages when it is possible to equate (dually) the forms of action L to action density S. As the ergodic hypothesis asserts that the open integrals are equivalent, it follows that a similar question may be posed about the closed integrals; but then the question becomes when are the one-dimensional periods equivalent to the $(N-1)$-dimensional periods. That is, assert $d\hat{\mathcal{A}} = 0$ and $d\hat{\mathcal{L}} = 0$ so that the question becomes

$$\oint \hat{\mathcal{A}} \overset{?}{\equiv} \oint\cdots\oint \hat{\mathcal{S}}; \tag{F7}$$

but this question is similar to the notion that the flux quanta are equivalently related to the action quanta.

The basic idea of importance here is the notion that the periods of $\mathcal{A}$ and the periods of $\mathcal{S}$ are dually related in ergodic systems. If the one-dimensional periods of $\mathcal{A}$ are quantized flux and if the periods of $\mathcal{S}$ are quantized angular momentum, then they are individually in

the ratio of the integers by Brouwer's theorem and directly related to one another (that is, dually equivalent) if $B_n = B_{n-p}$, e.g., if the space is compact and orientable. As not all spaces have $B_p = B_{n-p}$, the flux and action quanta are not always dually related, and not all systems are ergodic.

[1]C. Truesdell and R.A. Toupin, *Handbuch der Physik* (Springer-Verlag, Berlin, 1960), III/1, p. 660.
[2]This author was alerted to the Brouwer theorem by H. Flanders, *Differential Forms* (Academic, New York, 1963), p. 72; the theorem is given in many texts, for example, see J.G. Hocking and G.S. Young, *Topology* (Addison-Wesley, Reading, Mass., 1961).
[3]F. Rohrlich, *Classical Theory of the Electron* (Addison-Wesley, Reading, Mass., 1965), p. 65.
[4]E.J. Post, "Current Status of Geometrodynamics", McGill University Progress Report to the "Foundations and Philosophy of Science Unit" (1972).
[5]R.A. Toupin, "Centro Internazionale Matematico Estivo" (C.I.M.E.) course held at Bressanone (March—June 1965).
[6]R.M. Kiehn and J.F. Pierce, Phys. Fluids 12, 1971 (1969).
[7]R.M. Kiehn, J. Math. Phys. 15, 9 (1974).
[8]F. London, *Superfluids* (Dover, New York, 1950), Vol. I.
[9]R.P. Feynman, R.B. Leighton, and M. Sands, *Lectures on Physics Vol III* (Addison-Wesley, Reading, Mass., 1965), p. (21—8); R.E. Collins, Found. Phys. (to be published, 1977).
[10]R.M. Kiehn, Int. J. Eng. Sci. 13, 941 (1975).
[11]R.M. Kiehn, J. Math. Phys. 16, 1032 (1975).
[12]K. Kuratowski, *Topology* (Warsaw, 1948), Vol. I.
[13]R.M. Kiehn, Int. J. Eng. Sci. 14, 749 (1976).
[14]R.M. Kiehn, Lett. Nuovo Cimento 12, 300 (1975).
[15]H. Weyl, *Space, Time, Matter* (Dover, New York, 1950).
[16]P. Alexandroff, *Elementary Concepts of Topology* (Dover, New York, 1961), p. 66.

Superfields as an extension of the spin representation of the orthogonal group

J. O. Winnberg

Institute of Theoretical Physics, Fack, S-402 20 Göteborg 5, Sweden
(Received 13 October 1976)

We show that a superfield can be interpreted as a spinor belonging to the spin representation of a Clifford algebra. A subset of this algebra is connected by a linear automorphism to the orthogonal group. With this interpretation it should be possible to give a physical meaning to Grassmann variables.

1. INTRODUCTION

Grassman algebra, although invented in the mid-nineteenth century,[1] has only recently been used in theoretical physics. To put bosonic and fermionic operators on an equal footing in field theory, Berezin[2] introduces Grassmann parameters and develops a calculus for these. More recently, the same type of technique has been used in supersymmetric theories,[3,4] where there exists a symmetry between particles of different spins, spaced one-half unit apart. In these theories, one introduces the concept of superspace, which is essentially a direct sum of ordinary space and a Grassmann algebra.

Supersymmetric theories were originally deduced from a study of the Neveu–Schwarz–Ramond (NSR) dual model.[5] The superspace techniques have later been used in this model, to find a Lagrangian for a spinning string,[4] corresponding to the NSR model, and to reformulate the operator formalism.[6] In fact, one finds that, introducing superoperators into the Veneziano dual model,[5] one obtains the NSR model. In a series of papers[7] it has been shown that the Grassmann parameters can be used to put an internal symmetry into dual models in a nontrivial way. Finally, the superspace technique has recently been used to give a classical correspondence to spin-half particles.[8]

It should be pointed out that in the cases cited, the Grassmann parameters can be done without. They serve as a mathematical tool that greatly simplifies the formal manipulations by keeping track of various symmetries. Mostly, however, no deeper significance is given to the Grassmann parameters themselves. Thus, in the study of the mathematical foundations of supersymmetry, one has concentrated on the graded Lie algebra (GLA) structure of the theory.[9] The anticommuting parameters then keep track of the grading and, when introduced, turns the GLA into a "pseudo Lie algebra" with only commutators. The great computational simplification obtained and such examples as the unification of the NSR and Veneziano dual models make us feel, however, that these variables could indeed have a deeper significance. The question then arises about their interpretation. A step towards an answer is taken in this note by pointing out the following fact:

A Grassmann algebra, as used in supersymmetry, is equivalent, in a sense which will be clear below, to the spin representation of a Clifford algebra. This algebra in turn is closely related to the orthogonal group.

The questions about the nature and origin of the vector space on which this orthogonal group acts are completely open. If it is a tangent space or the space of a local internal symmetry, the vectors will be functions of space–time, and the Clifford algebra local. In other cases we will have a global Clifford algebra. In the case of dual string theory, one candidate is the tangent space of the string world sheet. For a supersymmetric field theory there are several possibilities, which apply to strings as well. In standard supersymmetry, the Grassmann parameters are Majorana spinors, which means that the vector space is a subspace of the space of Lorentz spinors. Other possibilities that we mention are space–time itself and the space of an internal symmetry. The plan of this paper is as follows: In Sec. 2 we define a Grassmann algebra and indicate briefly its use in supersymmetric theories. In Sec. 3, we define Clifford algebra and show its relation to the orthogonal group. In Sec. 4 we study the spin representation and show how the Grassmann algebra is contained in it. In the conclusion we sum up our results and discuss them briefly. It should be mentioned at this point that the application of the present ideas to the development of physical theory is beyond the scope of the present paper and will have to await further investigations.

2. GRASSMANN ALGEBRA

Let V be a d-dimensional vector space and define $\Lambda^n V$ to be the space of totally antisymmetric tensor products of n vectors. Obviously $\Lambda^d V$ is the highest possible such space and has only one element. The Grassmann or exterior algebra $G(V)$ of V is now defined as follows:

Let $\theta_i \in V$. Define the product $\circ$ by

$$\theta_i \circ \theta_j \in \Lambda^2 V, \quad \theta_i \circ \theta_j = -\theta_j \circ \theta_i, \tag{2.1}$$

$$\theta_i \circ \theta_j \circ \theta_k \in \Lambda^3 V, \quad \theta_i \circ (\theta_j \circ \theta_k) = (\theta_i \circ \theta_j) \circ \theta_k, \tag{2.2}$$

i.e., the product is anticommutative and associative. Then by recursion

$$\theta_i \in V, \quad \xi_n \in \Lambda^n V, \quad \theta_i \circ \xi_n \in \Lambda^{n+1} V \tag{2.3}$$

and

$$G(V) = \mathbf{1} + \bigoplus_{n=1}^{d} \Lambda^n V. \tag{2.4}$$

[Below we can give a mathematically more exact definition of $G(V)$ as the Clifford algebra of the null form.] A superfield ϕ is now a field which is also a general element of $G(V)$, if V is defined over a suitable field. Let $\{\theta_i\}$ be a basis for V. Then

Copyright © 1977 American Institute of Physics

$$\phi(x)=A(x)+\theta_i\psi_i(x)+\theta_i\circ\theta_jB_{ij}(x)+\cdots+\left[\prod_{i=1}^{d}\theta_i\right]F(x). \tag{2.5}$$

A well-known example is the superfield for the NSR string, where $d=2$ and the two thetas are collected in a two-component Majorana spinor θ. The superfield then consists of a scalar ϕ, a spinor ψ, and a scalar field F,

$$x(\xi,\theta)=\phi(\xi)+\theta\psi(\xi)+\bar{\theta}\theta F(x). \tag{2.6}$$

On the superfield we can perform the following "Grassmann operations": (from now on $\{\theta_i\}$ is a basis for V)

(i) Multiplication by any element of $G(V)$.

(ii) Taking the "derivative" $\partial/\partial\theta$ defined by

$$\frac{\partial}{\partial\theta_i}\theta_j=\delta_{ij}-\theta_j\frac{\partial}{\partial\theta_i} \tag{2.7}$$

so that

$$\frac{\partial}{\partial\theta_1}\theta_1=1,\quad \frac{\partial}{\partial\theta_1}\theta_2\theta_1=-\theta_2. \tag{2.8}$$

(iii) Performing "integrations" defined by

$$\int d\theta=0,\quad \int d\theta_i\theta_j=\delta_{ij} \tag{2.9}$$

with the rule that the variable to be integrated must be anticommuted to the left in the integrand before integrating, thus

$$\int d\theta_1\theta_2\theta_1=-\int d\theta_1\theta_1\theta_2=-\theta_2. \tag{2.10}$$

It is easily seen that the operators $\partial/\partial\theta$ and $\int d\theta$ are identical. The two different forms are used for esthetic reasons, the derivative occurring in conjunction with space derivatives, the integration will space integrals. Algebraically, $\partial/\partial\theta_i$ is the antiderivation of the Grassmann algebra corresponding to the linear form $g_i(\theta_j)=\delta_{ij}$. More precisely, the statement is the following[10]:

Let g be a linear form on V. Then there exists a uniquely determined antiderivation δ_g of $G(V)$ such that $\delta_g(\theta)=g(\theta)$ for all $\theta\in V$. The operator δ_g is homogeneous of degree -1, i.e., it maps Λ^nV into $\Lambda^{n-1}V$ and $\mathbb{I}$ upon $\{0\}$. Furthermore, $\delta_g^2=0$, $\delta_{ag}=a\delta_g$, $\delta_{g+g'}=\delta_g+\delta_{g'}$, and $\delta_g\delta_{g'}+\delta_{g'}\delta_g=0$.

In the Grassmann algebra obtained below, the linear form stems from the scalar product of V such that

$$g_i(\theta_j)=(\theta_i,\theta_j).$$

3. CLIFFORD ALGEBRA

Clifford algebra was invented a hundred years ago,[11] and a few particular examples are well-known in physics since the advent of quantum theory. For example, the Dirac algebra is the Clifford algebra of Minkowski space and the Pauli algebra is that of 3-space. An extensive study of the application of Clifford algebra to physics has been made by Hestenes.[12] This section is based on that reference and, mostly, on the monograph by Chevalley.[10] In the latter, the proofs of theorems stated here as well as the exact form of those theorems not explicitly stated can be found. Since most readers will be quite unfamiliar with the formalism used, we give some examples of its use in Minkowski space.

A Clifford algebra is similar to a Grassmann algebra in that the product is anticommutative and associative, the difference being that the anticommutator of some vectors is nonzero. Furthermore, the Clifford algebra depends by definition on the quadratic form (essentially the norm) of the vector space, and is through this closely connected to the orthogonal group. We can take over the definition of $G(V)$ in Sec. 2 if we change (2.1) to

$$\alpha_i\alpha_j+\alpha_j\alpha_i=2g_{ij},\quad \alpha_i\in V, \tag{3.1}$$

g_{ij} being the metric of V and denoting the Clifford product by juxtaposition. The exact definition goes as follows:

Let V be a vector space over a field K and let $Q(\alpha)$ be a quadratic form on V. Let T be the full tensor algebra over V, $T=\oplus_{m=0}^{\infty}T_m$, and let Q be the two-sided ideal of T generated by the elements $\alpha\otimes\alpha-Q(\alpha)$ for all $\alpha\in V$. Then the Clifford algebra of V with respect to Q is defined as the factor algebra $C_Q(V)=T/I_Q$.

In the following we will assume that the field K does not correspond to any pathological case and that $Q(\alpha)$ is nondegenerate. To gain an understanding of the above abstract definition, let us study the Clifford algebra of Minkowski space M. Denote the vectors of a basis of M by γ_0, γ_1, γ_2, and γ_3. Then an element of $T(M)$ is $\gamma_\alpha\otimes\gamma_\beta+\gamma_\beta\otimes\gamma_\alpha$. In $C_Q(M)$, with $Q(\gamma_\mu)=\gamma_\mu^2$, we can factor out, i.e., subtract, any element of I_Q. We get

$$\begin{aligned}\gamma_\alpha\gamma_\beta+\gamma_\beta\gamma_\alpha\sim{}&\gamma_\alpha\otimes\gamma_\beta+\gamma_\beta\otimes\gamma_\alpha-[(\gamma_\alpha+\gamma_\beta)\otimes(\gamma_\alpha+\gamma_\beta)\\&-Q(\gamma_\alpha+\gamma_\beta)]-[-(\gamma_\alpha\otimes\gamma_\alpha-Q(\gamma_\alpha))]\\&-[-(\gamma_\beta\otimes\gamma_\beta-Q(\gamma_\beta))]\\={}&Q(\gamma_\alpha+\gamma_\beta)-Q(\gamma_\alpha)-Q(\gamma_\beta)\\\equiv{}&2(\gamma_\alpha,\gamma_\beta)=2g_{\alpha\beta}.\end{aligned} \tag{3.2}$$

Thus $C_Q(M)$ is the ordinary Dirac algebra

$$\{\gamma_\alpha,\gamma_\beta\}=2g_{\alpha\beta}.$$

We see that we can now define $G(V)$ as the Clifford algebra of the null form $Q(\alpha)=0$ for all $\alpha\in V$,

$$G(V)=C_0(V).$$

Next we study the relation between $C_Q(V)$ and the orthogonal group. Certain elements of $C_Q(V)$ belong to the Clifford group Γ, defined below. There is a mapping from Γ to the orthogonal group $O(V,Q)$, the latter being defined as the group of transformations of V that leaves $Q(\alpha)$ invariant. In the next section we will study the spin representation of $C_Q(V)$. When restricted to Γ this gives the spin representation of $O(V,Q)$. The Clifford group Γ is the group of invertible elements s of $C_Q(V)$ such that $s\alpha s^{-1}\in V$ for all $\alpha\in V$. Denote by $\chi(s)$ the linear automorphism

$$\chi(s):\alpha\to s\alpha s^{-1}. \tag{3.3}$$

χ is a linear representation of Γ on V called the vector representation. The relation to $O(V,Q)$ is given by the following theorem.

Theorem: Let d be the dimension of V. If d is even, then $\chi(\Gamma)=O(V,Q)$. If d is odd, then $\chi(\Gamma)=SO(V,Q)$, the group of operations of determinant one in $Q(V,Q)$. If τ is any nonsingular vector of V, i.e., if $Q(\tau)\neq 0$, then $\tau\in\Gamma$ and $\chi(\tau)$ is the mapping $\alpha\to R_\tau\alpha$ where R_τ is the reflection through the line defined by τ.

Again let us give an example in M. A nonsingular vector is given by $\tau=a\gamma_1$ with $\tau^{-1}=-a^{-1}\gamma_1$. Then

$$\begin{aligned}\chi(\tau)\cdot\alpha &= -\alpha\gamma_1\alpha^\mu\gamma_\mu a^{-1}\gamma_1\\ &= -\gamma_1\alpha^\mu\{\gamma_\mu,\gamma_1\}+\gamma_1^2\alpha^\mu\gamma_\mu\\ &= 2\alpha^1\gamma_1-\alpha^\mu\gamma_\mu.\end{aligned}$$

Thus all components of α except the one along τ changes sign. It is easy to check that two transformations τ and τ' with $\alpha/2$ the angle between the vectors, gives a rotation through the angle α in the plane spanned by τ and τ'.

4. THE SPIN REPRESENTATION OF $C_Q(V)$

From now on we shall assume that the dimension d of V is even. We begin with the definition of the spin representation:

All simple representations of the simple algebra $C_Q(V)$ are equivalent. We select one of them, denoted by ρ, and call the space S of this representation the space of spinors of $C_Q(V)$. The representation ρ is called the spin representation.

Define a totally singular subspace P of V as a subspace on which $Q(\alpha)$ is the null form, $Q(\beta)=0$ for all $\beta\in P$. We now assume that the quadratic form $Q(\alpha)$ is of maximal index, i.e., that the dimension of a maximal totally singular subspace of V is $d/2$. Such is always the case if K is an algebraically closed field such as the complex numbers. In other cases, the assumption is a restriction on $Q(\alpha)$ so that, for example, in the case where K is the real numbers, $O(V,Q)$ is $O_{d/2,d/2}$. When Q is of maximal index, V can be split (in infinitely many ways) into the sum of two totally singular subspaces N and P, $V=N+P$. Denote be $C(N)$ the Clifford algebra of N [identical to $G(N)$], by $C(P)$ that of P, and by f the product of the elements of some basis of P, i.e., f is the element of $\Lambda^{d/2}P$. Obviously $C_Q(V)f$ is a left ideal of $C_Q(V)$. It can be shown that it is also a minimal left ideal and that $C_Q(V)f=C(N)f$. There is a representation ρ of $C_Q(V)$ on $C(N)$ such that $(\rho(\xi)\cdot\theta)f=\xi\theta f$ for $\xi\in C_Q(V)$, $\theta\in C(N)$. Since $C(N)f$ is a minimal left ideal, ρ is simple. Thus we may take $C(N)$ to be the space S of spinors, ρ being the spin representation. This definition of a spinor is due to Riesz[13] and is much later than the Dirac theory.

Let us give an example in M. Different choices of N and P correspond to different representations of the γ matrices. We choose

$$\begin{aligned}&N:\theta_1=\tfrac12\{\alpha_0+\alpha_3\},\quad \theta_2=\tfrac12\{\alpha_1+i\alpha_2\},\\ &P:\bar\theta_1=\tfrac12\{\alpha_0-\alpha_3\},\quad \bar\theta_2=\tfrac12\{-\alpha_1+i\alpha_2\},\end{aligned}\tag{4.1}$$

with α_μ a basis for M such that $\alpha_\mu^2=g_{\mu\mu}$. A spinor ψ is a general element of $C(N)$,

$$\psi=\psi_1+\theta_1\theta_2\psi_2+\theta_1\psi_3+\theta_2\psi_4.\tag{4.2}$$

A rotation through the angle ϑ in the $\alpha_1\alpha_2$ plane is given by the reflection $R_{\alpha_1}=\theta_2-\bar\theta_2$ followed by the reflection

$$\begin{aligned}&R_{\alpha_1\cos\vartheta/2+\alpha_2\sin\vartheta/2}\\ &\quad=(\theta_2-\bar\theta_2)\cos\vartheta/2-i\{\theta_2+\bar\theta_2\}\sin\vartheta/2,\end{aligned}$$

and thus

$$\begin{aligned}D_{\alpha_1\alpha_2}(\vartheta)&=[(\theta_2-\bar\theta_2)\cos\vartheta/2-i(\theta_2+\bar\theta_2)\sin\vartheta/2](\theta_2-\bar\theta_2)\\ &=-\{\theta_2,\bar\theta_2\}\cos\vartheta/2+i[\theta_2,\bar\theta_2]\sin\vartheta/2.\end{aligned}\tag{4.3}$$

We will apply this to a spinor below.

We are now in a position to show that the spin representation is equivalent to a Grassmann algebra formalism. First, let $\theta\in N$. Then

$$(\rho(\theta)\cdot\psi)f=\theta\psi f=(\theta\psi)f\tag{4.4}$$

since $C_Q(V)$ is associative and thus

$$\rho(\theta)\cdot\psi=\theta\psi,\tag{4.5}$$

where the rhs is a product in the Grassmann algebra $C(N)$. Next, let $\bar\theta\in P$ and choose a particular basis $\{\theta_i\}$ for N and $\{\bar\theta_i\}$ for P such that $(\bar\theta_i,\theta_j)=\frac12\delta_{ij}$. Then $\{\bar\theta_i,\theta_j\}=\delta_{ij}$ and with $\bar\theta=\vartheta^i\bar\theta_i$ we get

$$\begin{aligned}(\rho(\bar\theta)\cdot\psi)f&=\bar\theta\psi f\\ &=\vartheta^i\bar\theta_i\{\psi_0+\psi_i\theta_i+\psi_{ij}\theta_i\theta_j+\cdots\}f\\ &=\psi_0\vartheta^i\bar\theta_i f+\vartheta^i\psi_i[\{\bar\theta_i,\theta_i\}f-\theta_i\bar\theta_i f]+\vartheta^i\psi_{ij}[\{\bar\theta_i,\theta_i\}\theta_j f\\ &\quad-\theta_i\{\bar\theta_i,\theta_j\}f+\theta_i\theta_j\bar\theta_i f]+\cdots.\end{aligned}\tag{4.6}$$

But $C_Q(V)$ is associative and $\bar\theta_i f=0$ since $C(P)$ is a Grassmann algebra. By comparing (4.6) with (2.7) we see that

$$\rho(\bar\theta)\psi=\frac{\partial}{\partial\theta}\psi\tag{4.7}$$

or in general, with $g_{\bar\theta}(\theta)$ the linear form $g_{\bar\theta}(\theta)=2(\theta,\bar\theta)$,

$$\rho(\bar\theta)\psi=\delta_{g_{\bar\theta}}\psi.\tag{4.8}$$

Applying now $D_{\alpha_1\alpha_2}(\vartheta)$ to a spinor in M, we have

$$\begin{aligned}\rho(\bar\theta_2)\psi&=\frac{\partial}{\partial\theta_2}\{\psi_1+\theta_1\theta_2\psi_2+\theta_1\psi_3+\theta_2\psi_4\}\\ &=-\theta_1\psi_2+\psi_4\end{aligned}\tag{4.9}$$

and

$$\begin{aligned}&D_{\alpha_1\alpha_2}(\vartheta)\psi\\ &=\left\{-\cos\frac{\vartheta}{2}+i\left(\theta_2\frac{\partial}{\partial\theta_2}-\frac{\partial}{\partial\theta_2}\theta_2\right)\sin\frac{\vartheta}{2}\right\}\psi\\ &=-\left\{\cos\frac{\vartheta}{2}-i\left(2\theta_2\frac{\partial}{\partial\theta_2}-1\right)\right\}\psi\\ &=\psi_1\left\{-\cos\frac{\vartheta}{2}-i\sin\frac{\vartheta}{2}\right\}+\theta_1\theta_2\psi_2\left\{-\cos\frac{\vartheta}{2}+i\sin\frac{\vartheta}{2}\right\}\\ &\quad+\theta_1\psi_3\left\{-\cos\frac{\vartheta}{2}-i\sin\frac{\vartheta}{2}\right\}+\theta_2\psi_4\left\{-\cos\frac{\vartheta}{2}+i\sin\frac{\vartheta}{2}\right\},\end{aligned}\tag{4.10}$$

or, in matrix notation

$$D_{\alpha_1\alpha_2}(\vartheta)\psi=-\begin{pmatrix}e^{i\vartheta/2}&0&0&0\\0&e^{-i\vartheta/2}&0&0\\0&0&e^{i\vartheta/2}&0\\0&0&0&e^{-i\vartheta/2}\end{pmatrix}\begin{pmatrix}\psi_1\\\psi_2\\\psi_3\\\psi_4\end{pmatrix}.\tag{4.11}$$

As was pointed out in Sec. 2 there is no need for the operator $\int d\theta$, since it is identical to $\partial/\partial\theta$. It appears natural to use it, however, in taking the spinor norm. There is a theorem stating that there exist only two bilinear invariants on the space of spinors. These can be defined as follows:

Let $\beta(\psi, \psi')$ be the first invariant. To calculate it, take the main antiautomorphism α of ψ, i.e., change the order of all products of thetas in ψ,

$$\alpha(\psi) = \psi_0 + \psi_i\theta_i + \psi_{ij}\theta_j\theta_i + \psi_{ijk}\theta_k\theta_j\theta_i + \cdots . \tag{4.12}$$

Then multiply $\alpha(\psi)$ and ψ' and find the coefficient multiplying the product of all θ_i of N, i.e., find the term belonging to $\Lambda^{d/2}N$. Using the Grassmann integral, this can conveniently be written

$$\beta(\psi, \psi') = \int \left(\prod_1^{d/2} d\theta_i \right) \alpha(\psi)\psi' . \tag{4.13}$$

The other spinor invariant $\tilde{\beta}(\psi, \psi')$ is obtained if one also takes the main involution (corresponding to Hermitian conjugation and thus denoted in the same way) of ψ, i.e., one lets an element $\xi \in \Lambda^n N$ go into $(-1)^n\xi$. Thus

$$\psi^\dagger = \psi_0 - \psi_i\theta_i + \psi_{ij}\theta_i\theta_j - \psi_{ijk}\theta_i\theta_j\theta_k + \cdots . \tag{4.14}$$

Then

$$\tilde{\beta}(\psi, \psi') = \int \left(\prod_1^{d/2} d\theta_i \right) \alpha(\psi^\dagger)\psi' . \tag{4.15}$$

It is interesting to note that the singular vectors are exactly those elements of V that does not belong to Γ. Thus $\rho(\theta)$ is *not* a transformation of the orthogonal group, whereas transformations of the form $\rho(a\theta + b\bar{\theta})$ are.

If we want to use a superfield in quantum field theory, we would like this field to have definite commutation relations. Since the spinor components are what we have left to arrange for this, we must have that ψ_i, ψ_{ijk} and all odd-number-of-indices components are fermionic, whereas the others are bosonic or vice versa. However, the product of Clifford spinors is not defined within the Clifford algebra representation but belongs to the tensor product of this representation with itself or some factor algebra thereof.

5. CONCLUSION

In conclusion we may sum up our results in the statements:

(i) A superfield is a spinor of a Clifford algebra $C_Q(V)$.

(ii) θ and $\partial/\partial\theta$ are the operations of left multiplication by θ and $\bar{\theta}$, respectively, in the spin representation of this algebra.

(iii) $\int d\theta$ occurs naturally in taking the spinor norm.

(iv) θ and ∂/θ are precisely those elements of V which do not belong to the Clifford group and thus do not belong to the orthogonal group in the spin representation.

In superspace, it is being a direct sum of Minkowski space and a Grassmann algebra $G(N)$, one introduces the transformation

$$\begin{aligned} \theta' &= \theta + \delta\theta = \theta + \xi, \\ \chi^{\mu'} &= \chi^\mu + \delta\chi^\mu = \chi^\mu + (\xi, A^\mu\theta), \end{aligned} \tag{5.1}$$

where $(\cdot, \cdot)$ is a scalar product of N and A^μ is an operator on N which is also a Lorentz vector. In standard supersymmetry, V is the space of Majorana spinors and A^μ is then taken to be the Dirac matrices. In other cases there is the possibility that A^μ could provide a nontrivial connection between an inner symmetry and the Poincaré group. Since a superfield is a spinor in the coordinate-independent formulation (cf. $\hat{A} = A^i\hat{X}_i$), it is obvious that it transforms as a scalar w.r.t. coordinate transformations of V, i.e., when $\theta_i \to \theta_i' = \theta_i + \delta\theta_i$. The transformation properties under supersymmetry transformations will then depend wholly on to which representation of the Poincaré group the field belongs. The question of a physical interpretation of a superfield when considered as a Clifford spinor has not been touched upon yet. First one has to decide which is the space V on which $C_Q(V)$ is defined. Second, the theory has to be specified, for example, by superspace transformations under which it is to be invariant, which is to say that N is to be the representation space of some group to which the operators A^μ of (5.1) belong. We hope to come back to these questions elsewhere.

Note added in proof: In a recent paper[14] by myself *et al.*, it has been found that a pseudoclassical particle interacting with a gauge field after quantization is represented by a Clifford spinor. We refer the reader to that paper for further details.

ACKNOWLEDGMENT

We wish to thank Lars Brink for helpful discussions and for critically reading the manuscript.

[1] H. Grassmann, *Die Ausdehnungslehre* (1862).
[2] F. A. Berezin, *The Method of Second Quantization* (Academic, New York, 1966).
[3] B. Zumino, in *Proceedings of the XVII International Conference on High Energy Physics* (Science Research Council, London, 1974).
[4] B. Zumino, in *Proceedings of the Conference on Gauge Theory and Modern Field Theory, North Eastern University*, edited by R. Arnowitt and R. Nath (MIT, 1975); L. Brink, P. Di Vecchia, and P. Howe, Gothenburg preprint 76-22 (1976); P. A. Collins and R. W. Tucker, University of Lancaster preprint (1976); S. Deser and B. Zumino, CERN preprint TH 2276 (1976).
[5] M. Jacob., ed., *Dual Theory* (North-Holland, Amsterdam, 1974).
[6] L. Brink and J. O. Winnberg, Nucl. Phys. B 103, 445 (1976).
[7] M. Ademollo, L. Brink, A. D'Adda, R. D'Auria, E. Napolitano, S. Scinto, E. Del Giudice, P. Di Vecchia, S. Ferrara, F. Gliozzi, R. Musto, and R. Petterino, Phys. Lett. B 62, 105 (1976); Nucl. Phys. B **114**, 297 (1976); same authors and J. Schwarz, Nucl. Phys. B **111**, 77 (1976).
[8] L. Brink, S. Deser, B. Zumino, P. Di Vecchia, and P. Howe, preprint TH-2208 CERN (1976).
[9] L. Corwin, Y. Ne'eman, and S. Sternberg, Rev. Mod. Phys. **47**, 573 (1975).
[10] C. Chevalley, *The Algebraic Theory of Spinors* (Columbia U. P., New York, 1954).
[11] W. K. Clifford, Amer. J. Math. **1**, 350 (1878).
[12] D. Hestenes, *Spacetime Algebra* (Gordon and Breach, New York, 1966).
[13] M. Riesz, *Dixième Congrès des Mathématiciens Scandinaves, Copenhague* (Copenhagen, 1946), p. 123. 1946), p. 123.
[14] A. P. Balachandran, P. Salomonsson, B. S. Skagerstam, and J. O. Winnberg, Gothenburg preprint 76-35 (1976).

A C^*-algebra formulation of the quantization of the electromagnetic field

A. L. Carey, J. M. Gaffney,* and C. A. Hurst

Department of Mathematical Physics, University of Adelaide, South Australia 5001
(Received 4 November 1976)

A presentation of the Fermi, Gupta–Bleuler, and radiation gauge methods for quantizing the free electromagnetic field is given in the Weyl algebra formalism for quantum field theory first introduced by Segal. The abstract Weyl algebra of the vector potential is defined using the formalism of Manuceau. Then the Fermi and Gupta–Bleuler methods are given as schemes for constructing representations of the algebra. The algebra of the physical photons is shown to be a factor algebra of a certain subalgebra of the original algebra of the vector potential. In this formalism, the application of the supplementary condition in the Fermi method, and the supplementary condition and indefinite metric in the Gupta–Bleuler method, can be interpreted as the means by which a representation of this factor algebra is obtained. The Weyl algebra of the physical photons is the Weyl algebra associated with the radiation gauge method. It is also shown that in the Fock representation of the Weyl algebra given by the Fermi method, automorphisms of the algebra corresponding to Lorentz transformations cannot always be implemented by unitary transformations. This leads us to construct a new representation of the Weyl algebra which provides a covariant representation for the vector potential.

1. INTRODUCTION

It is well known that the methods which have been developed for quantizing the electromagnetic field cannot be fitted into the framework of conventional field theory. There is, however, no general understanding why this should be so. Usually, the accounts of the various methods can only describe the departure from conventional field theory as an unpleasant feature of the particular formulation. Some time ago, Strocchi,[1] in studying the quantization of the electromagnetic field in the Wightman formulation of field theory, obtained results which suggest that "unpleasant" features are inevitable for a theory of the electromagnetic field.

It is against this background that the results obtained in this paper are to be understood. Conventional field theory cannot be used to establish the consistency of a procedure for quantizing the electromagnetic field and hence no method should be dismissed simply because it leads to a field theory with unusual features. The Fermi method[2] fell into disfavor largely because the physical states were found to be unnormalizable.[3] The primary aim of this paper is to show that it is possible to give a consistent mathematical structure for the Fermi method, following ideas which were adumbrated in Ref. 4.

The key idea is that the Fermi method should not be interpreted as a Hilbert space theory with an unsatisfiable supplementary condition, but rather as a C^*-algebra theory with the Fermi condition to be interpreted as a prescription for singling out a "gauge invariant" subalgebra and constructing a Poincaré invariant quotient algebra of this subalgebra. This quotient algebra becomes the algebra of the physical photons.

For reasons that will become evident, Segal's Weyl algebra formalism[5] (and in particular, the variant due to Manuceau[6]) which is a C^*-algebra formulation of field theory, provides a convenient basis for a rigorous formulation of the Fermi method. It is interesting to see how this formalism can be modified to accommodate the radiation gauge[7] and Gupta—Bleuler[8] methods. However, because these methods are well established, their relationship with the Weyl algebra formalism will only be sketched.

We shall begin in Sec. 2 with a brief discussion of the Weyl algebra formalism. To show how this formalism applies in a familiar situation we shall discuss in Sec. 3 the canonical commutation relations (CCR) for four degrees of freedom,

$$[q_\mu, p_\nu] = -ig_{\mu\nu}. \tag{1.1}$$

This example is particularly relevant to the present discussion because in the Weyl algebra formulation various features are present in the four-dimensional case which are also present in a Weyl algebra formulation of the Gupta—Bleuler and Fermi methods. These features are not found in a Weyl algebra formulation of the scalar field.

In Sec. 4, the Weyl algebra of the vector potential will be defined. The Fermi method will be discussed in Sec. 5 as a particular representation of this algebra. We will show that in this representation automorphisms of the algebra corresponding to Poincaré transformations are not necessarily implementable by unitary transformations. Finally, we show in Sec. 6 that the Fermi method nevertheless has a covariant representation in which the Poincaré transformations act unitarily.

2. THE WEYL ALGEBRA FORMALISM

The basic formalism of the Weyl algebra approach to field theory was developed by Segal.[5] We will primarily use a variant of this approach, due to Manuceau,[6] and discussed in detail by Emch.[9] We shall first present an account of the formalism which emphasizes its relationship to the more familiar formulations. Discussions of the Weyl algebra approach in which the formalism is developed in a more natural fashion are given in Ref. 10.

We begin by considering the CCR in one dimension.

 Copyright © 1977 American Institute of Physics

They can be formulated either in the Heisenberg form

$$[q,p]=i \tag{2.1}$$

or in the Weyl form

$$U(a)V(b)=V(b)U(a)\exp(iab), \tag{2.2}$$

where

$$U(a)=\exp(iap), \quad V(b)=\exp(ibq). \tag{2.3}$$

These forms of the CCR are not mathematically equivalent and the relationship between them given by (2.3) has only formal significance. The technical advantage of working with $U(\cdot)$ and $V(\cdot)$ rather than p and q is well recognized and hence the Weyl form of the CCR is usually mathematically simpler to handle in a rigorous approach.

If we let

$$W(a,b)=U(a)V(b)\exp(-iab/2), \tag{2.4}$$

then a multiplication law for these objects is defined by

$$W(a_1,b_1)W(a_2,b_2)=\exp[-i(a_2b_1-a_1b_2)/2] \times W(a_1+a_2,b_1+b_2). \tag{2.5}$$

An algebra is formed by considering the set of all finite linear combinations of such $W(\cdot,\cdot)$ with complex coefficients. This algebra is then turned into an involutive algebra by defining the $*$ operation by

$$W(a,b)^*=W(-a,-b). \tag{2.6}$$

Consider the Schrödinger representation of the CCR. This is obviously a representation of the algebra we have just defined. In the Schrödinger representation, since $U(\cdot)$ and $V(\cdot)$ are represented by operators that are also strongly continuous as unitary representations of the real line R, Stone's theorem[11] ensures the existence of self adjoint operators that from (2.3) are a representation of q and p. Normally, we want to be able to set up representations of q and p and hence it is reasonable to require that the representations of $U(\cdot)$ and $V(\cdot)$ also be strongly continuous representations of R so that Stone's theorem will apply.

The continuity and unitarity conditions on the representation of the Weyl relation that enable us to use Stone's theorem also serve to establish the essential uniqueness of the Schrödinger representation. More precisely, it is the content of the von Neumann uniqueness theorem[12] that any irreducible representation of the CCR satisfying these continuity and unitarity conditions is unitarily equivalent to the Schrödinger representation.

The elements $W(\cdot,\cdot)$ are labeled by two real numbers which can be considered to be components of a vector in R_2. Also, if we let

$$B(a_1,b_1;a_2,b_2)=a_2b_1-a_1b_2, \tag{2.7}$$

then B is a nondegenerate bilinear, antisymmetric (i.e., symplectic), real form on R_2. If z denotes a vector in R_2, then the multiplication law defined by (2.5) can be written in a coordinate free fashion, i.e.,

$$W(z)W(z')=W(z+z')\exp[-iB(z,z')/2]. \tag{2.8}$$

This multiplication law can also be considered to be the multiplication law for the CCR for any number of degrees of freedom by appropriately interpreting z and B. If the number of degrees of freedom, n, is finite then z is a vector in R_{2n}. The explicit form of B can be calculated quite easily. For a field with an infinite number of degrees of freedom, it is possible to interpret z as a vector in a suitable space of wavefunctions and to identify B with the commutation relation.

We have already noted that the imposition of certain conditions on representations of the Weyl relation ensures that operators satisfying the Heisenberg form of the CCR will exist. In the Weyl algebra approach to the CCR, representations which satisfy these conditions are said to be Weyl systems.

Definition 2.1: Suppose that M is a real, locally convex topological vector space and $B(z,z')$ is symplectic form on M which is continuous in the given topology. Then we have a Weyl system over (M,B) if to each pair of elements z and z' in M there corresponds unitary operators $W(z)$ and $W(z')$ acting on a Hilbert space and satisfying

$$W(z)W(z')=W(z+z')\exp[-iB(z,z')/2].$$

Furthermore, we require $W(tz)$ to be a strongly continuous function of $t\in R$, for any fixed z.

The existence of Weyl systems over (M,B) is guaranteed by the following additional structure.[10] Let J be a real linear operator on M satisfying $J^2=-1$, $B(Jz_1,Jz_2)=B(z_1,z_2)$ for all $z_1,z_2\in M$ and $B(z,Jz)>0$ if and only if $z\neq 0$. Then M becomes a complex pre-Hilbert space if we define a complex inner product by

$$\langle z_1,z_2\rangle = B(z_1,Jz_2)+iB(z_1,z_2),$$

and interpret J as multiplication by i. Clearly B is the imaginary part of the inner product.

To define an abstract C*-algebra for the CCR we follow Manuceau[6] (cf. Emch[9]).

Definition 2.2: Let $\Delta(M)$ be the normed $*$-algebra defined as follows:

(i) The elements of $\Delta(M)$ are complex valued functions on M which vanish everywhere except for a finite subset of M.

(ii) Equip $\Delta(M)$ with the obvious linear structure and the multiplication

$$(f_1f_2)(z)=\sum_{z_1\subset M} f_1(z_1)f_2(z-z_1)\exp[-(i/2)B(z_1,z)].$$

(iii) The involution is given by $f^*(z)=\overline{f(-z)}$ and the norm by

$$\|f\|_1=\sum_{z\in M}|f(z)|.$$

Let $\Delta_1(M)$ denote the completion of $\Delta(M)$ in the above norm. A basis of $\Delta(M)$ is given by the functions δ_z defined by $\delta_z(z')=1$ if $z'=z$ and zero otherwise. Let P be the set of all nondegenerate continuous $*$-representations π of $\Delta_1(M)$ by bounded operators on a Hilbert space such that $\pi(\delta_{\lambda z})$ is weakly continuous in $\lambda\in R$ for all z in M.

Definition 2.3: The C*-algebra of the CCR over (M,B) is defined to be the completion of $\Delta_1(M)$ in the

norm

$$\|f\| = \sup_{\pi \in P} \|\pi(f)\|, \quad f \in \Delta(M).$$

We denote this algebra by $\Delta_c(M)$.

The Weyl systems over (M, B) can be put into one to one correspondence with the elements of P by setting $W_\pi(z) = \pi(\delta_z)$ for each $\pi \in P$. In this way we recover the operators $W(z)$ of Eq. (2.8). Moreover, since $\Delta_c(M)$ is a C^*-algebra, cyclic representations of $\Delta_c(M)$ can be associated with each positive, linear normalized functional (i.e., state) on the algebra by the Gel'fand and Naimark,[14] Segal[14] (GNS) construction. The representation obtained via the GNS construction, however, is not necessarily a Weyl system, since $W(tz)$ need not be a continuous function of t. The following result (Theorem III.1.7 in Emch[9]) classified those states which give rise to Weyl systems.

Theorem 2.4: Suppose that ρ is a complex-valued function on M such that

(i) $\rho(0) = 1$,

(ii) $\rho(\lambda z + z')$ is continuous in $\lambda \in R$ for all $z, z' \in M$,

(iii) for arbitrary z_i in M and complex numbers α_i,

$$\sum_{i,j \in F} \rho(z_i - z_j)\alpha_i \bar{\alpha}_j \exp\left(-\frac{i}{2}B(z_i, z_j)\right) \geq 0,$$

where F is a finite index set. Then there corresponds a unique (up to unitary equivalence) cyclic representation π_ρ of $\Delta_c(M)$ with cyclic vector v such that $\pi_\rho \in P$ and

$$\rho(z) = \langle v, W_\rho(z)v\rangle \quad (= \langle v, \pi_\rho(\delta_z)v\rangle).$$

Every such ρ defines (in Segal's terminology[5]) a regular state E_ρ on $\Delta_c(M)$ by $E_\rho(A) = \langle v, \pi_\rho(A)v\rangle$ where $A \in \Delta_c(M)$. Conversely, every cyclic Weyl system over (M, B) may be obtained from a generating functional ρ, in this way.

3. CANONICAL COMMUTATION RELATIONS FOR FOUR DEGREES OF FREEDOM

Consider the Hilbert space, $L^2(R_4, d^4x)$, consisting of equivalence classes of Lebesque square integrable functions over R_4. If $\mathbf{x}$ denotes a vector in R_4 with components (x_0, x_1, x_2, x_3), then the Schrödinger representation of the CCR,

$$[q_\mu, p_\nu] = -ig_{\mu\nu}, \quad g_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1), \tag{3.1}$$

will be given by

$$\begin{aligned}
&(q_i f)(\mathbf{x}) = x_i f(\mathbf{x}),\\
&(p_i f)(\mathbf{x}) = -i\frac{\partial f(\mathbf{x})}{\partial x_i}, \quad i = 1, 2, 3,\\
&(q_0 f)(\mathbf{x}) = x_0 f(\mathbf{x}),\\
&(p_0 f)(\mathbf{x}) = i\frac{\partial f(\mathbf{x})}{\partial x_0},
\end{aligned} \tag{3.2}$$

where there will be certain domain conditions on the functions f which will not concern us here (cf. Ref. 9).

We now consider the unitary operators

$$\begin{aligned}
&U(\mathbf{a}) = \exp(-i\Sigma g_{\mu\nu}a_\mu p_\nu),\\
&V(b) = \exp(-i\Sigma g_{\mu\nu}b_\mu q_\nu),
\end{aligned} \tag{3.3}$$

where $a_\mu, b_\mu \in R$, $\mu = 0, 1, 2, 3$. Their action on functions in L^2 will be

$$\begin{aligned}
&U(a)f(\mathbf{x}) = f(\mathbf{x} + \mathbf{a}),\\
&V(\mathbf{b})f(\mathbf{x}) = \exp(-i\Sigma g_{\mu\nu}b_\mu x_\nu)f(\mathbf{x}).
\end{aligned} \tag{3.4}$$

The Weyl operator is then defined by

$$W(\mathbf{a}, \mathbf{b}) = U(\mathbf{a})V(\mathbf{b})\exp(\tfrac{1}{2}i\Sigma_{\mu\nu}g_{\mu\nu}a_\mu b_\nu) \tag{3.5}$$

and from this we find the Weyl relation:

$$\begin{aligned}
&W(\mathbf{a}, \mathbf{b})W(\mathbf{a}', \mathbf{b}')\\
&\quad = W(\mathbf{a} + \mathbf{a}', \mathbf{b} + \mathbf{b}')\exp(-i\Sigma g_{\mu\nu}(a_\mu b'_\nu - a'_\mu b_\nu)/2).
\end{aligned} \tag{3.6}$$

Identity M of definition 2.1 with R_8 and define

$$B(\mathbf{c}, \mathbf{c}') = \Sigma g_{\mu\nu}(a_\mu b'_\nu - a'_\mu b_\nu) \tag{3.7}$$

where $\mathbf{c} = (\mathbf{a}, \mathbf{b}) \in R_8$. Then B is a nondegenerate real symplectic form on M and is invariant under the transformations $\mathbf{c} \to \mathbf{c}'$ defined by

$$\mathbf{c}' = (\Lambda\mathbf{a}, \Lambda\mathbf{b}) \tag{3.8}$$

where Λ is a Lorentz transformation. These symplectic transformations induce automorphisms of the algebra $\Delta_c(R_8)$ which will be given by Ref. 6,

$$\delta_{\mathbf{a},\mathbf{b}} \to \delta'_{\mathbf{a},\mathbf{b}} = \delta_{\Lambda\mathbf{a},\Lambda\mathbf{b}}. \tag{3.9}$$

In terms of the operators q and p these automorphisms are

$$q \to q' = \Lambda^{-1}q \quad \text{and} \quad p \to p' = \Lambda^{-1}p, \tag{3.10}$$

since

$$\Sigma g_{\mu\nu}a'_\mu p_\nu = \Sigma g_{\mu\nu}a_\mu p'_\nu$$

and

$$\Sigma g_{\mu\nu}b'_\mu q_\nu = \Sigma g_{\mu\nu}b_\mu q'_\nu.$$

So the Schrödinger representation is a Weyl system over (R_8, B) in the sense of Definition 2.1.

The CCR (3.1) are invariant under the usual matrix representation of the Lorentz group and hence the operators q_μ and p_ν are four vectors. We see that in the Weyl algebra formalism (3.6) the four vector automorphisms (3.10) are induced by symplectic transformations in the parameter space, but these symplectic transformations do not always leave the usual Euclidean scalar product in R_8 invariant.

Now the Schrödinger representation is an irreducible Weyl system for the Weyl algebra $\Delta_c(R_8)$. Hence, in accordance with Theorem 2.4, the state E on the algebra that generates the Schrödinger representation can be specified by a generating functional ρ where

$$\rho(\mathbf{a}, \mathbf{b}) = E(\delta_{\mathbf{a},\mathbf{b}}) = (\Omega_0, W(\mathbf{a}, \mathbf{b})\Omega_0) \tag{3.11}$$

and Ω_0 is the cyclic vector of Theorem 2.4.

We will take Ω_0 to be the vacuum state:

$$\Omega_0(x_0, x_1, x_2, x_3) = \pi^{-1}\exp(-\tfrac{1}{2}\Sigma_\mu x_\mu^2). \tag{3.12}$$

On performing the required integration we obtain

$$\rho(\mathbf{a}, \mathbf{b}) = \exp[-\tfrac{1}{4}\Sigma_\mu(a_\mu^2 + b_\mu^2)].$$

If we put

$$d\mu(\mathbf{a}, \mathbf{b}) = (\Omega_0, W(\mathbf{a}, \mathbf{b})\Omega_0)\, d\mathbf{a}\, d\mathbf{b},$$

then $d\mu(\mathbf{a},\mathbf{b})$ is the only normalized product measure which is invariant under the group of orthogonal transformations on R_8. (The uniqueness of this measure and the uniqueness of the Schrödinger representation are interrelated.) Consider the effect of four-vector automorphisms of the algebra on the generating functional. The automorphism of the algebra in which $x \to \Lambda^{-1}x$, is induced by the transformation $(\mathbf{a},\mathbf{b}) \to (\Lambda\mathbf{a}, \Lambda\mathbf{b})$ in the parameter space. Therefore, under these automorphisms the generating functional transforms according to

$$\begin{aligned}\rho \to \rho' : \rho[\mathbf{a},\mathbf{b}] &= \rho[\Lambda\mathbf{a}, \Lambda\mathbf{b}],\\ &= \exp(-\tfrac{1}{4}\Sigma_\mu[(\Lambda\mathbf{a})_\mu^2 + (\Lambda\mathbf{b})_\mu^2]),\\ &\neq \rho[\mathbf{a},\mathbf{b}].\end{aligned}$$

The new generating functional, ρ', generates a representation of the Weyl algebra which by the von Neumann uniqueness theorem must be unitarily equivalent to the representation generated by ρ. Since ρ' is the new vacuum expectation value of the Weyl operator the noninvariance of ρ under Lorentz transformations implies the noninvariance of the vacuum. Explicitly, in the Schrödinger representation

$$\begin{aligned}\Omega_0'(\mathbf{x}) &= \Omega_0(\Lambda^{-1}\mathbf{x})\\ &= \exp[-\tfrac{1}{2}\Sigma_\mu(\Lambda^{-1}\mathbf{x})_\mu^2]\\ &= U\Omega_0(\mathbf{x}).\end{aligned}$$

Another way of proving that the vacuum is necessarily noninvariant in a representation of the CCR with q and p self-adjoint operators, is to consider the relationship between the operators q_μ, p_μ and the annihilation and creation operators a_μ, $\bar{a}_\mu$ where $\bar{a}_\mu$ is the Hermitian conjugate of a_μ. In the Schrödinger representation with vacuum given by (3.12) we may define

$$a_\mu = \frac{1}{\sqrt{2}}\left(x_\mu + \frac{\partial}{\partial x_\mu}\right), \quad \bar{a}_\mu = \frac{1}{\sqrt{2}}\left(x_\mu - \frac{\partial}{\partial x_\mu}\right). \tag{3.13}$$

In terms of the q_μ and p_μ of Eq. (3.2) we have

$$\begin{aligned}&a_0 = \frac{1}{\sqrt{2}}(q_0 - ip_0), \quad a_i = \frac{1}{\sqrt{2}}(q_i + ip_i),\\ &\bar{a}_0 = \frac{1}{\sqrt{2}}(q_0 + ip_0), \quad \bar{a}_i = \frac{1}{\sqrt{2}}(q_i - ip_i), \quad i = 1,2,3\end{aligned} \tag{3.14}$$

in the Schrödinger representation. However, the operators a_μ, $\bar{a}_\mu$ as defined by (3.14) will, from (3.1), satisfy the commutation relations

$$[a_\mu, \bar{a}_\nu] = \delta_{\mu\nu} \tag{3.15}$$

and this will be true in any representation of the CCR (3.1). It is clear from (3.15) that a_μ cannot transform as a four-vector, nor can $\bar{a}_\mu$, and indeed the appropriate combinations must be $(a_0, \bar{a}_i)$ and $(\bar{a}_0, a_i)$ if we wish to form four-vectors. We conclude therefore that if q_μ and p_μ transform as four-vectors, then the annihilation and creation operators will be mixed up under Lorentz transformations and hence neither the vacuum defined by

$$a_\mu\Omega_0 = 0, \quad \mu = 0,1,2,3$$

nor the generating functional can possibly be Lorentz invariant.

In (3.14) we note the difference between the representation of p_0 and p_i in terms of annihilation and creation operators. In the Fermi method the CCR are satisfied by an analogous representation of field operators in terms of annihilation and creation operators. In this connection we recall Dirac's remark[16] concerning the Fermi method. He noted that the reality condition for the field was satisfied by reversing the roles of the annihilation and creation operators for the zero components of the field. Since Fock space is often thought of as being built up from states formed by creation operators acting on a vacuum state it might be more appropriate for the roles of the annihilation and creation operators to be maintained while the representation of the field operator is different so that the commutation relations can be satisfied.

It is a useful exercise to formulate the CCR given by (3.1) within the Weyl algebra framework because it is an example in which the parameter space does not have a complex structure invariant under the appropriate symmetry group. Also we see that in choosing a Weyl system for the algebra we are forced into having a noninvariant vacuum.

Now a representation of this algebra can be given which is analogous to the Gupta–Bleuler representation. In this representation, the operators q_0 and p_0 are, in terms of the annihilation and creation operators,

$$q_0 = \frac{1}{\sqrt{2}}(a_0 - \bar{a}_0), \quad p_0 = \frac{-i}{\sqrt{2}}(a_0 + \bar{a}_0). \tag{3.16}$$

On the Hilbert space $L^2(R_4, d^4x)$, q_0 and p_0 are given by the skew-adjoint operators

$$(q_0 f)(x) = ix_0 f(x), \quad (p_0 f)(x) = \frac{\partial}{\partial x_0} f(x). \tag{3.17}$$

The vacuum will now be invariant under automorphisms of the algebra, $q \to \Lambda q$ and $p \to \Lambda p$. Clearly however, neither $\exp(-ia_0 p_0)$ nor $\exp(-ib_0 q_0)$ (defined using the functional calculus for the skew-adjoint p_0 and q_0) can be unitary whence the Weyl operators defined in terms of these exponentials cannot be unitary. The "Gupta–Bleuler" representation is not a Weyl system. In discussions of the Gupta–Bleuler method an indefinite metric is used from the outset. The indefinite metric, however, does not arise from the prescription (in terms of generating functionals) for obtaining representations of the abstract CCR algebra. It comes in fact, as part of a prescription for obtaining a representation of a factor algebra.

The discussion of the CCR of (3.1) will now be taken a little further by showing how we can construct a certain invariant factor algebra from the original Weyl algebra. This discussion will be given in some detail as a completely analogous situation occurs for the Weyl algebra of the vector potential.

Suppose that instead of the full Lorentz group we only consider a Euclidean group which is a subgroup of symplectic transformations on R_8. The representations of E(2) contained in the finite dimensional representations of the Lorentz group are indecomposable.[17] We

shall choose the generators of E(2) on R_8 to be

$$E_1=\begin{bmatrix}E_1' & 0\\ 0 & E_1'\end{bmatrix},\quad E_2=\begin{bmatrix}E_2' & 0\\ 0 & E_2'\end{bmatrix},\quad M_3=\begin{bmatrix}M_3' & 0\\ 0 & M_3'\end{bmatrix},$$

$$E_1'=\begin{bmatrix}0&0&-1&1\\0&0&0&0\\1&0&0&0\\1&0&0&0\end{bmatrix},\quad E_2'=\begin{bmatrix}0&0&0&0\\0&0&-1&1\\0&1&0&0\\0&1&0&0\end{bmatrix},\quad M_3'=\begin{bmatrix}0&1&0&0\\-1&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}. \tag{3.18}$$

Let N, T, and S be subspaces of R_8 of vectors of the form $(a_1, a_2, a, a; b_1, b_2, b, b)$, $(0, 0, a, a, 0, 0, b, b)$, and $(a_1, a_2, 0, 0, b_1, b_2, 0, 0)$ respectively. Then N is the direct sum of S and T. On vectors in the subspace T, the generators of E(2) are zero. Hence the group E(2) acts trivially on T. The subspace N is also invariant under E(2) but S is not.

We shall define the algebra $\Delta_c(N)$ to be the C^*-subalgebra of $\Delta_c(R_8)$ which is the closure of the algebra $\Delta(N)$ consisting of functions $\Sigma\lambda_i\delta_{n_i}$ where $n_i \in N$. Similarly, we define $\Delta(S)$, $\Delta_c(S)$, $\Delta(T)$, and $\Delta_c(T)$. In the Schrödinger representation $\Delta_c(N)$ is faithfully represented by a C^*-algebra of bounded operators, say $W_c(N)$. Let $\mathcal{N}$ be the weak closure of $W_c(N)$ and similarly define algebras $W_c(T)$, $W_c(S)$ and hence the von Neumann algebras $\mathcal{S}$ and $\mathcal{T}$. Let us record some facts about these various algebras.

(i) The algebra $\mathcal{T}$ is Abelian, in fact, since $\Delta(T)$ is the center of $\Delta(N)$ is follows that $\mathcal{T}$ is the center of $\mathcal{N}$.

(ii) Consider the automorphisms of $\Delta_c(N)$ induced by the group E(2) acting in R_8,

$$\delta_{\mathbf{a},\mathbf{b}} \to \delta'_{\mathbf{a},\mathbf{b}} = \delta_{G\mathbf{a},G\mathbf{b}} \tag{3.19}$$

where $G \in$ E(2). The algebras $\Delta_c(N)$ and $\Delta_c(T)$ will be invariant under these automorphisms because of the invariance of N and T. However, since S is not invariant, neither is $\Delta_c(S)$.

(iii) Following Manuceau,[6] one can show that $\Delta_c(N)$ is isomorphic to the C^* tensor product $\Delta_c(S)\otimes\Delta_c(T)$.

[That there is just one C^* norm on the algebraic tensor product of $\Delta_c(S)$ and $\Delta_c(T)$ follows from the fact that $\Delta_c(T)$ is Abelian and from proposition 1.22.5 of Sakai.[18]] Note that the degeneracy of the symplectic form on T does not alter the validity of Manuceau's proof[6] in this case, as $W_c(T)$ provides a faithful representation of $\Delta_c(T)$.

We now construct an algebra isomorphic to $\Delta_c(S)$ but which is invariant under these E(2) automorphisms. Define a homomorphism π: $\Delta(N) \to \Delta(S)$ by $\pi(\Sigma\lambda_i\delta_{n_i}) = \Sigma\lambda_i\delta_{s_i}$ where $n_i = s_i + t_i$ is the unique decomposition of n_i into elements of S and T.

Theorem 3.1: The map π extends to a continuous homomorphism from $\Delta_c(N)$ into $\Delta_c(S)$ such that $\Delta_c(S)$ is isometrically isomorphic to $\Delta_c(N)/I$, where I is the kernel of π.

Proof: We note that π is a homomorphism from the algebra $\Delta(N)$ onto $\Delta(S)$ because of the degeneracy of B on T. Further, it is straightforward to check that π is continuous as a map from $\Delta_1(N)$ into $\Delta_c(S)$, where $\Delta_1(N)$ is the closure of $\Delta(N)$ in the norm $\|\ \|_1$ on $\Delta(R_8)$, introduced in Definition 2.2. Now let ρ denote the restriction of the Schrödinger representation of $\Delta_c(R_8)$ to $\Delta_c(N)$. Then $\rho\circ\pi$ is a continuous nondegenerate representation of $\Delta_1(N)$. Introduce a new C^*-algebra $\overline{\Delta(N)}$ defined by taking all nondegenerate continuous representations σ of $\Delta_1(N)$ and defining a C^* norm by

$$\|A\|_0 = \sup_\sigma \|\sigma(A)\|, \quad A \in \Delta_1(N).$$

Clearly, $\overline{\Delta(N)} \subseteq \Delta_c(N)$ as $\|A\|_0 \geq \|A\|$, $A \in \Delta_1(N)$. Now ρ is a faithful representation of $\Delta_c(N)$ and hence of $\overline{\Delta(N)}$. By the well-known norm preserving properties of faithful representations we therefore have

$$\|\rho(A)\| = \|A\| = \|A\|_0 \quad \text{for } A \in \Delta_1(N).$$

That is, $\overline{\Delta(N)} = \Delta_c(N)$. However, $\rho\circ\pi$ must satisfy

$$\|\rho\circ\pi(A)\| \leq \|A\|_0 = \|A\|, \quad A \in \Delta_1(N).$$

That is, $\rho\circ\pi$ extends to a continuous representation of $\Delta_c(N)$. As ρ is faithful and $\pi(\Delta_1(N)) = \Delta_1(S) \subseteq \Delta_1(N)$ we have

$$\|\rho(\pi(A))\| = \|\pi(A)\|, \quad A \in \Delta_1(N).$$

But this implies $\|\pi(A)\| \leq \|A\|$, so π is continuous. The rest follows by Corollary 1.8.3 of Dixmier.[19]

Denote the cosets of I in $\Delta_c(N)$ by $n + I = \hat{n}$ where $n \in \Delta_c(N)$. Distinct elements in $\Delta_c(S)$ belong to distinct cosets and from the isomorphism between $\Delta_c(S)$ and $\Delta_c(N)/I$ we see that the set of equivalence classes $s = s + I$, $s \in \Delta_c(S)$ exhausts $\Delta_c(N)$.

Let us again consider the automorphisms of $\Delta_c(R_8)$ induced by E(2). We note that these automorphisms act as the identity on $\Delta_c(T)$ (as they act trivially on T). Further, the ideal I is generated by the elements $\Sigma\lambda_i\delta_{t_i}$ of $\Delta(T)$ such that $\Sigma\lambda_i = 0$. Since these elements are invariant under the E(2) automorphisms we conclude that I is invariant. (The heuristic reasons for the invariance of I were described in Ref. 4.)

We can define automorphisms of $\Delta_c(N)/I$ from automorphisms of $\Delta_0(R_8)$ which leave both $\Delta_c(N)$ and I invariant in the obvious way. If $n \to n'$ under some automorphism, then

$$(n + I) \to (n + I)' = n' + I \tag{3.20}$$

will be an automorphism of $\Delta_c(N)/I$. Note that for cosets of the form $\delta_{s+t} + I$, $s \in S$ and $t \in T$, an E(2) induced automorphism maps s to $s' + t'$ and t to itself. Hence from (3.20)

$$(\delta_{s+t} + I)' = \delta_{s'}\delta_{t'+t} + I = \delta_{s'} + I.$$

Identifying $\delta_s + I$ with δ_s and $\delta_{s'} + I$ with $\delta_{s'}$, we see that forming the factor algebra $\Delta_c(N)/I$ implies a redefinition of the transformation law for elements of $\Delta_c(S)$.

We shall now show how it is possible to construct a representation of the algebra $\Delta_c(N)/I$ from the Schrödinger representation of $\Delta_c(R_8)$. Any realization of a Hilbert space as a space of functions over a locally compact space can be considered as a representation of an abstract Hilbert space in the form of a direct integral. Hence we can think of the Schrödinger representation (for one degree of freedom) as the direct integral, with respect to the Lebesque measure, of one-dimensional Hilbert spaces assigned to each point on the real

line. Furthermore, to each direct integral of Hilbert spaces there corresponds a commutative weakly closed algebra of bounded linear operators containing the identity operator.[20] The algebra that corresponds to the direct integral decomposition for the Schrödinger representation is the weakly closed algebra generated by $W(0, \mathbf{b})$. Alternatively, it is a representation of this algebra by essentially bounded measurable functions over the operators q_μ.

Suppose that we decompose a Hilbert space into a direct integral $H = \int d\lambda H_\lambda$. Then, if R is the algebra corresponding to the direct integral decomposition, any operator, T, on H that commutes with all elements of R is reducible to "diagonal" form, i.e., $T = \{T(\lambda)\}$.

These considerations can be applied to the subalgebras $\mathcal{N}$ and $\mathcal{T}$. Since $\mathcal{T}$ is a commutative weakly closed algebra, there will exist a direct integral decomposition of the Hilbert space with respect to $\mathcal{T}$. Also, since the elements of $\mathcal{N}$ commute with the elements of $\mathcal{T}$ we can reduce elements of $\mathcal{N}$ to "diagonal" form.

This direct integral decomposition can be constructed explicitly. If

$$W(\mathbf{a}, \mathbf{b}) = \exp[-i\Sigma g_{\mu\nu}(a_\mu p_\nu + b_\mu q_\nu)]$$

is in $\mathcal{T}$ then

$$\Sigma g_{\mu\nu}(a_\mu p_\nu + b_\mu q_\nu) = a(p_0 - p_3) + b(q_0 - q_3)$$
$$= ia\left(\frac{\partial}{\partial x_0} + \frac{\partial}{\partial x_3}\right) + b(x_0 - x_3).$$

Therefore, we seek a unitary mapping of these operators such that

$$i\left(\frac{\partial}{\partial x_0} + \frac{\partial}{\partial x_3}\right) \to \lambda_1 \quad \text{and} \quad (x_0 - x_3) \to \lambda_2. \tag{3.21}$$

This is achieved with the unitary mapping

$$f(x_0, x_1, x_2, x_3) \xrightarrow{U} g(x_1, x_2, \lambda_1, \lambda_2),$$

where

$$g(x_1, x_2, \lambda_1, \lambda_2)$$
$$= \frac{1}{\sqrt{\pi}} \int du f(\tfrac{1}{2}(u + \lambda_2), x_1, x_2, \tfrac{1}{2}(u - \lambda_2)) \exp(iu\lambda_1/2). \tag{3.22}$$

For $(\mathbf{a}, \mathbf{b})$ in N with components $(a_1, a_2, a, a, b_1, b_2, b, b)$,

$$W(\mathbf{a}, \mathbf{b}) g(x_1, x_2, \lambda_1, \lambda_2)$$
$$= g(x_1 + a_1, x_2 + a_2, \lambda_1, \lambda_2) \exp[-i(a\lambda_1 + b\lambda_2)]$$
$$\times \exp[i(b_1 x_1 + b_2 x_2) + \tfrac{1}{2}(a_1 b_1 + a_2 b_2)]. \tag{3.23}$$

Hence,

$$H = \int d\lambda_1' d\lambda_2' H_{\lambda_1'\lambda_2'}$$

represents $L^2(R_4, d^4x)$ as a direct integral of the spaces

$$H_{\lambda_1'\lambda_2'} = L^2(R_2, d^4x).$$

All the elements in $\mathcal{N}$ are in "diagonal" form, i.e., if $T \in \mathcal{N}$ then $T = \{T(\lambda_1', \lambda_2')\}$.

Let the mapping $\Pi_{\lambda_1'\lambda_2'}$ be defined by

$$\Pi_{\lambda_1'\lambda_2'} T \to T(\lambda_1', \lambda_2'),$$

where $T(\lambda_1', \lambda_2')$ is an operator acting on $H_{\lambda_1'\lambda_2'}$. Then the operators $T(\lambda_1', \lambda_2')$ for T in $W_c(N)$ form an irreducible representation of the algebra $\Delta_c(N)$. In this decomposition the fact that $\Delta_c(N)$ can be considered as the direct product of the algebras $\Delta_c(S)$ and $\Delta_c(T)$ is clearly evident. The reducible representation of $\Delta_c(N)$ can be considered to be constructed as a direct integral of the direct product of irreducible representations of $\Delta_c(S)$ and $\Delta_c(T)$.

The mapping $T \to T(0, 0)$ takes a representation of $\Delta_c(N)$ to a representation of the factor algebra $\Delta_c(N)/I$. This follows directly from the way in which $\Delta_c(N)/I$ was defined since the elements of $\Delta_c(N)/I$ are the equivalence classes of elements in $\Delta_c(N)$ that are mapped onto a given element of $\Delta_c(S)$. By inspection, the mapping $T \to T(0, 0)$ takes the representing elements of I to zero, the operators $T(0, 0)$ for $T \in W_c(N)$ form a representation of $\Delta_c(S)$, and hence we have a representation of $\Delta_c(N)/I$.

Although the operators $T(0, 0)$ act only on H_{00} we can identify them with the operators acting on H in the following way. The operators $T(0, 0)$ are an irreducible representation of the algebra $\Delta_c(S)$. We therefore identify the operator $T(0, 0)$ with an element of $W_c(N)$ which belongs to the subalgebra $W_c(S)$.

Explicitly, consider the operator $W(\mathbf{a}, \mathbf{b})$ of (3.23). $H_{00} = L^2(R_2, d^4x)$ and under the mapping $T \to T(0, 0)$,

$$W(\mathbf{a}, \mathbf{b}) \to W(\mathbf{a}, \mathbf{b})(0, 0),$$

where

$$W(\mathbf{a}, \mathbf{b})(0, 0) g(x_1, x_2)$$
$$= g(x_1 + a_1, x_2 + a_2) \exp[i(b_1 x_1 + b_2 x_2) + \tfrac{1}{2}(a_1 b_1 + a_2 b_2)]. \tag{3.24}$$

We therefore identify $W(\mathbf{a}, \mathbf{b})(0, 0)$ with $W(\mathbf{a}', \mathbf{b}')$ acting in H with $(\mathbf{a}', \mathbf{b}) = (a_1, a_2, 0, 0, b_1, b_2, 0, 0)$.

The operators are of the same form in both H and H_{00}. However, while it does not matter whether we think of the operators of the factor algebra $\Delta_c(N)/I$ as acting in either H or H_{00} we cannot identify vectors in H_{00} and H. The vectors in H_{00} are not normalizable with respect to the measure in H. The function $g(x_1, x_2)$ can be an element of H_{00} but it will certainly not be an element of H.

Essentially what we are trying to do is solve the eigenvalue problem

$$(p_0 - p_3) f = 0, \quad (q_0 - q_3) f = 0. \tag{3.25}$$

Because the operators $p_0 - p_3$ and $q_0 - q_3$ have continuous spectra this is not possible in H. Nevertheless, one can see formally that on the space of solutions of (3.25), $W_c(T)$ would act trivially while every operator T in $W_c(N)$ would be formally identical with its image $T(0, 0)$ under $\Pi_{0,0}$. To make this rigorous one observes that the set of solutions of (3.25) consists of distributions $g(x_1, x_2)\delta(\lambda_1)\delta(\lambda_2)$ where g is in $L^2(R_2, dx_1, dx_2)$. These vectors are not normalizable and hence do not lie in H. However we can identify them with vectors in H_{00} by simply dropping the δ functions. We then work with the operators $T(0, 0)$ (acting in the space H_{00}) rather than

the operators $T \in W_c(S)$ with which they are formally identical.

In the next section we shall show that the algebra of the physical photons is *a factor algebra of the algebra of the vector potential*. We find that the Fermi method is the generalization to an infinite number of degrees of freedom of the Schrödinger representation that we have been discussing in this section. The operators $p_0 - p_3$ and $p_0 - q_3$ are analogs of the supplementary condition operators in the Fermi method. The functions $g(x_1, x_2, \lambda_1, \lambda_2)$ are therefore the analogs of the Fermi physical states. Thus, the difficulty of the nonnormalizability of the Fermi physical states can be overcome by reinterpreting them as states in a Hilbert space of a direct integral decomposition. It is clear from the foregoing that it is only when attention is restricted to the subalgebra $\Delta_c(N)$ that this interpretation is possible. Physically, this means that we must work only with gauge invariant quantities. Furthermore, the arguments of this and the later sections show that if one wishes to treat the electromagnetic potential, and the related field tensor, as a Weyl system one is automatically led to quantization in accordance with the Fermi method. It is therefore the natural extension of both the heuristic and rigorous treatments of the scalar field, to the case of a zero mass vector field, and so can be regarded as a suitable alternative to the other methods usually employed.

4. THE WEYL ALGEBRA OF THE VECTOR POTENTIAL

We begin by writing the commutation relations for the vector potential in the Weyl form. In heuristic formulations, the CCR for the field operators are

$$[\mathcal{A}_\mu(x), \mathcal{A}_\nu(x')] = -ig_{\mu\nu}D(x - x'), \tag{4.1}$$

where $D(x - x')$ satisfies $\Box D = 0$, with the initial data

$$D(\mathbf{x}, 0) = 0, \quad \frac{\partial}{\partial t} D(\mathbf{x}, t)\Big|_{t=0} = -\delta(\mathbf{x}).$$

Let F, with components f_μ $(\mu = 0, 1, 2, 3)$ be a real test function in the Schwartz space $\sum_{i=0}^{3} \mathcal{S}^{(i)}(R_4)$. Then the CCR for the smeared fields

$$A(F) \equiv \int d^3x \mathcal{A}^\mu(x) f_\mu(x) \tag{4.2}$$

will be

$$[A(F), A(F')] = i\sigma(F, F'), \tag{4.3}$$

where

$$\sigma(F, F') = -\int d^4x\, d^4x' D(x - x') f^\mu(x) f'_\mu(x'). \tag{4.4}$$

Let $U(F) = \exp[iA(F)]$, then the CCR in Weyl form are

$$U(F)U(F') = U(F + F') \exp[-i\sigma(F, F')/2]. \tag{4.5}$$

The map $F \to U(F)$ just as for the scalar field, is not one-to-one. We therefore introduce the space of real solutions of the wave equation

$$\Box \phi_\mu = 0 \quad (\mu = 0, 1, 2, 3) \tag{4.6}$$

and the map $F \to \phi$ where ϕ has components

$$\phi_\mu = D * f_\mu. \tag{4.7}$$

We obtain a one-to-one map by defining

$$W(\phi) = U(F). \tag{4.8}$$

The bilinear form, σ, induces the symplectic form, B, on the space of real solutions of (4.6),

$$(F, F') \to B(\phi, \phi') = -\int_n d\eta^\mu(x) \left[\phi^\nu(x) \frac{\partial \phi'_\nu(x)}{\partial x^\mu} - \frac{\partial \phi^\nu(x)}{\partial x^\mu} \phi'_\nu(x)\right], \tag{4.9}$$

where n is a spacelike surface. This form is independent on n and is usually written in terms of the Cauchy data for the solutions. If the real solution $\phi_\mu(\mathbf{x}, t)$ at time t_0 has Cauchy data $\phi_\mu(\mathbf{x})$ and $\dot{\phi}_\mu(\mathbf{x})$, then (4.9) can be written in the familiar form

$$B(\phi, \phi') = -\int d^3\mathbf{x} [\phi^\mu(\mathbf{x}) \dot{\phi}'_\mu(\mathbf{x}) - \dot{\phi}^\mu(\mathbf{x}) \phi'_\mu(\mathbf{x})]. \tag{4.10}$$

We can now define the Weyl algebra of the vector potential for the electromagnetic field. We need only specify the appropriate M and B. B comes from (4.9) or (4.10) while to define M we firstly take M_0 to be the space of functions $D * F$ defined by (4.7) where the components f_μ of F range over the Schwartz space $\mathcal{S}(R_4)$, $\mu = 0, 1, 2, 3$. Then M is taken to be the completion of M_0 in the Fermi norm whose definition we defer until the next section [Eq. (5.3)].

The automorphisms of the algebra corresponding to Poincaré transformations are induced by the symplectic transformations

$$\phi \in M; \quad \phi \to \phi'; \quad \phi'_\mu(x) = \Lambda_\mu{}^\nu \phi_\nu(\Lambda^{-1}(x - a)), \tag{4.11}$$

which form a (nonunitary) representation of the Poincaré group on M. The symplectic form B is invariant under these transformations and hence the transformation $\delta_\phi \to \delta'_\phi$ induced by $\phi \to \phi'$ defines the relevant automorphism of the algebra $\Delta_c(M)$.

Our analysis so far applies equally well to both the Fermi and Gupta–Bleuler methods of quantization. The same abstract algebra $\Delta_c(M)$ is associated with each method and furthermore, corresponding Poincaré transformations in each theory are given by the same automorphism of $\Delta_c(M)$.

In descriptions of the Gupta–Bleuler and Fermi methods, the field operator is often interpreted in terms of annihilation and creation operators for physical and unphysical photons. Such an interpretation is however, misleading because strictly speaking operators for physical photons only emerge after supplementary conditions (an indefinite metric in the case of the Gupta–Bleuler method) have been applied.

The situation is quite analogous to that which we discussed in Sec. 3. The Weyl algebra of the physical photons is, as already mentioned, *a factor algebra of a subalgebra* of the original abstract Weyl algebra.

The subalgebra of the Weyl algebra associated with the so-called "physical photons" will be denoted by $\Delta_c(S)$. We shall define $\Delta_c(S)$ as the algebra generated by the elements δ_ϕ of $\Delta_c(M)$ where ϕ belongs to the subspace S, of M, consisting of wavefunctions which satisfy

$$\phi_0 = 0, \quad \nabla \cdot \boldsymbol{\phi} = 0. \tag{4.12}$$

Thus S is the set of wavefunctions with transverse components only.

We note that the two Poincaré invariant subspaces of M are N, defined as the closure of the subspace of M_0 consisting of elements that satisfy

$$\frac{\partial \phi^\mu}{\partial x^\mu} = 0 \tag{4.13}$$

and the subspace T, which is the closure of the subspace of M_0 consisting of elements of the form $\partial\Lambda/\partial x^\mu$, where Λ is a scalar solution of the wave equation.

In order to exhibit N as the direct sum $S \oplus T$ we need to project out the transverse components of any ϕ in N. This is most easily done by operating on the Fourier transform $\hat{\phi}$ of ϕ. For each vector (k_1, k_2, k_3, k) where $k = |\mathbf{k}|$, the Lorentz transformation

$$\Lambda(k) = \begin{pmatrix} 1 + k_1^2/k(k_3 - k) & k_1 k_2/k(k_3 - k) & k_1/k & 0 \\ k_1 k_2/k(k_3 - k) & 1 + k_2^2/k(k_3 - k) & k_2/k & 0 \\ k_1/k & k_2/k & k_3/k & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.14}$$

maps this vector onto $(0, 0, k, k)$. The transverse components of $\Lambda(k)\hat{\phi}(k)$ are obtained by applying the projection

$$P' = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}$$

and then applying the inverse transformation $\Lambda(k)^{-1}$. A simple calculation shows that they are

$$\begin{aligned} \hat{\phi}'_0(k) &= 0, \\ \hat{\boldsymbol{\phi}}'(k) &= \hat{\boldsymbol{\phi}}(k) - \mathbf{k}\mathbf{k} \cdot \hat{\boldsymbol{\phi}}(k)/k^2. \end{aligned} \tag{4.15}$$

In the coordinate representation one obtains the element ϕ' of S,

$$\boldsymbol{\phi}'(x) = \boldsymbol{\phi}(x) - \frac{1}{\nabla^2} \boldsymbol{\nabla}\boldsymbol{\nabla} \cdot \boldsymbol{\phi}(x) \quad \text{and} \quad \phi'_0(x) = 0. \tag{4.16}$$

Since $\hat{\phi}_\mu(k) - \hat{\phi}'_\mu(k) = k_\mu(\mathbf{k} \cdot \hat{\boldsymbol{\phi}}(k)/k^2)$, then ϕ and ϕ' differ by an element of T. Clearly $S \cap T = (0)$ and so N is the direct sum of S and T with the S component of each $\phi \in N$ given by (4.16). The fact that S is not Poincaré invariant is obvious from definitions (4.12) and (4.11). Hence the automorphisms of $\Delta_c(M)$ induced by Poincaré transformations in M, i.e., $\delta_\phi \to \delta_{\phi'}$, will not leave $\Delta_c(S)$ invariant. Thus we cannot use $\Delta_c(S)$ as the algebra of the physical photons and maintain the Lorentz covariance of the theory.

On the other hand, $\Delta_c(N)$ and $\Delta_c(T)$ are invariant subalgebras of $\Delta_c(M)$. As in the previous section we can use the decomposition $N = S \oplus T$ to write $\Delta_c(N)$ as the tensor product $\Delta_c(S) \otimes \Delta_c(T)$. Also, we can prove that $\Delta_c(T)$ is the center of $\Delta_c(N)$. $\Delta_c(T)$ will be called the algebra of the supplementary condition operators. To see this, consider the following formal calculation. Let

$$W(\phi) = \exp[iA(\phi)]$$

where

$$A(\phi) = \int d^4x \mathcal{A}^\mu(x) f_\mu(x) \quad \text{with } \phi_\mu = D * f_\mu.$$

For those ϕ in T, $\phi_\mu = \partial\Lambda/\partial x^\mu$,

$$A\left(\frac{\partial \Lambda}{\partial x^\mu}\right) = -\int d^4x \mathcal{A}^\mu(x) \frac{\partial f(x)}{\partial x^\mu} \quad \text{with } \Lambda = D * f.$$

On integrating this formal expression by parts we obtain

$$A\left(\frac{\partial \Lambda}{\partial x^\mu}\right) \equiv \int d^4x \frac{\partial \mathcal{A}^\mu}{\partial x^\mu}(x) f(x).$$

Now $\partial \mathcal{A}_\mu/\partial x^\mu$ is the supplementary condition operator in the heuristic formulation. Hence $\Delta_c(T)$ is the invariant subalgebra of $\Delta_c(M)$ of Lorentz gauge transformations.

To obtain an invariant factor algebra isomorphic to $\Delta_c(S)$ we use the procedure of the previous section. Define π as before to be the homomorphism of $\Delta(N)$ onto $\Delta(S)$ and use Theorem 3.1 to extend π to $\Delta_c(N)$ [replacing the work Schrödinger by Fock in the proof and $\Delta_c(R_3)$ by $\Delta_c(M)$]. Then we have $\Delta_c(S)$ isomorphic to $\Delta_c(N)/I$ where I is the kernel of π. All the remarks made in the previous section about $E(2)$ induced automorphisms now carry over to Poincaré induced automorphisms of $\Delta_c(M)$.

By labeling the cosets of I in $\Delta_c(N)$ as before by elements of $\Delta_c(S)$ we may identify the algebra $\Delta_c(N)/I$ with the Weyl algebra associated with the radiation gauge method. For, consider the element in $\Delta_c(N)/I$, $\delta_\phi + I$, where ϕ is in S. Then under Poincaré transformations we have $\phi \to \phi'$ and $\hat{\delta}_\phi \to \hat{\delta}_{\phi'}$. In the direct product decomposition of $\delta_{\phi'}$ the component in $\Delta_c(S)$ will be $\delta_{\phi''}$ where ϕ'' is the component of ϕ' in S [i.e., the transverse components of ϕ', cf. (4.15)]. Hence

$$\hat{\delta}_{\phi'} = \hat{\delta}_{\phi''}. \tag{4.17}$$

The mapping $\phi \to \phi''$ is the *transformation law of the potential in radiation gauge*.

The Weyl algebra Δ' say, associated with the radiation gauge method can be constructed without any reference to the algebra of the vector potential. It is the Weyl algebra over (M', B') with

$$M' = N/T = (S \oplus T)/T$$

and B' the restriction of B to S. The automorphisms of Δ' representing physical transformations will be induced by Poincaré transformations in M'. These transformations in M' are the usual transformations for the radiation gauge potential or alternatively can be defined from the Poincaré transformations in M since N and T are invariant subspaces of M. It is this algebra Δ' that Segal constructed[10] in applying the Weyl algebra formalism to the electromagnetic field. It should now be clear that $\Delta_c(N)/I$ and Δ' are isomorphic C^*-algebras. To see this we merely identify $\Sigma\lambda_i \delta_{s_i}$ in $\Delta_c(S)$ with $\Sigma\lambda_i \delta_{s_i + T}$ in Δ'. This extends to an isomorphism of $\Delta_c(S)$ with Δ' and hence to an isomorphism of $\Delta_c(N)/I$ and Δ'.

The Gupta–Bleuler and Fermi methods are different representations of this algebra. We can infer from other considerations that in the Gupta–Bleuler representation of the algebra $\Delta_c(M)$, the representation of

the factor algebra $\Delta_c(N)/I$ is constructed on a Hilbert space of states that are equivalence classes of states satisfying the supplementary condition. The indefinite metric is the mechanism for constructing such equivalence classes.

5. THE FERMI METHOD OF QUANTIZING THE ELECTROMAGNETIC FIELD

The Fermi method leads to a representation of $\Delta_c(M)$ that is a Weyl system and hence can be characterized by a generating functional. We can calculate it by a method which closely parallel that for the neutral scalar field. In this case, the generating functional arises from the scalar product in the Hilbert space of positive energy wavefunctions. So, we begin by associating real solutions with positive energy solutions in the usual fashion:

$$\phi_\mu(\mathbf{x},t)=\frac{1}{(2\pi)^{3/2}}\int\frac{d^3\mathbf{k}}{2\sqrt{k}}\,(\phi^+_\mu(\mathbf{k})\exp[-i(kt-\mathbf{k}\cdot\mathbf{x})] + \overline{\phi^+_\mu(\mathbf{k})}\exp[i(kt-\mathbf{k}\cdot\mathbf{x})]). \tag{5.1}$$

The Poincaré invariant form on M_0 is given by

$$(\phi,\phi')=-\int d^3\mathbf{k}\phi^{\mu+}(\mathbf{k})^*\phi'^+_\mu(\mathbf{k}). \tag{5.2}$$

$B(\phi,\phi')$ is certainly the imaginary part of $(\phi^+,\phi^{+\prime})$ but since this form is indefinite M_0 is not a pre-Hilbert space. Thus the usual methods for quantization of the electromagnetic field break down at this point.

In the Fermi method however one redefines the complex structure on M_0 in such a way as to give a nondegenerate complex inner product which is not Lorentz invariant. Specifically, one defines $(J\phi_\mu)^+(p) = -i\epsilon_{\mu 0}\phi^+_\mu(p)$, where $\epsilon_{\mu 0}=-1$ when $\mu=0$ and 1 otherwise. Clearly J does not commute with Lorentz transformations. Now define the Fermi inner product as

$$\langle\phi,\phi'\rangle_F=B(\phi,J\phi')-iB(\phi,\phi') =\int d^3\mathbf{k}\,(\boldsymbol{\phi}^+(\mathbf{k})^*\boldsymbol{\phi}'^+(\mathbf{k})+\phi^+_0(\mathbf{k})\phi'^+_0(\mathbf{k})^*). \tag{5.3}$$

Clearly $B(\phi,\phi')$ is the imaginary part of the scalar product (5.3). As in the previous section M denotes the completion of M_0 in this norm. The generating functional for the Fermi representation of $\Delta_c(M)$ is

$$\rho(\phi)=\exp(-\tfrac{1}{4}\langle\phi,\phi\rangle_F). \tag{5.4}$$

Poincaré transformations on M will be given by (4.11). Hence the scalar product in M will not be invariant under Poincaré transformations and therefore ρ cannot be invariant.

In the example of the CCR for four degrees of freedom, we encountered a noninvariant generating functional. However, this system is finite and the von Neumann uniqueness theorem enabled us to be certain that a unitary operator would exist to implement the automorphism of the Weyl algebra, which is induced by the motion of the generating functional, and also to map old vacuum to a new one. For infinite systems the von Neumann theorem does not hold and in general, relatively few automorphisms of the Weyl algebra of the infinite system are implemented in a given representation of the algebra.

The automorphisms of $\Delta_c(M)$ induced by Poincaré transformations fall into a class of automorphisms that have been studied by Shale.[21]

There is however one technical point. Shale's results depend on Segal's version[5] of the definition of the Weyl algebra for the CCR. That his results also apply to Manuceau's formalism can be most easily seen by the following. Let $\mathcal{J}$ be the family of all complex finite dimensional linear subspaces of M on which B is nondegenerate. Then $\Delta_c(M)$ is the inductive limit[13] of the C^*-algebras $\Delta_c(F)$ where F is equipped with the symplectic form obtained by restriction of B.[6] In Segal's formulation the Weyl algebra of the CCR's, $\mathcal{W}(M,B)$, is defined as the C^* inductive limit of the algebras $\mathcal{W}(F,B)$ where $\mathcal{W}(F,B)$ is the weak closure of $\Delta_c(F)$ in the weak topology of the Schrödinger representation. Clearly, therefore, there is a continuous injection of $\Delta_c(M)$ into $\mathcal{W}(M)$ and, using the characterization of Weyl systems by generating functionals (Theorem 2.4), we see that there is a 1-1 correspondence between the Weyl systems defined by each algebra. We will use Shale's notation[21] and write $W(z)$ for δ_z when it is to be considered as an element of $\mathcal{W}(M,B)$.

With these preliminary remarks out of the way let us quote Shale's criterion for the unitary implementability of automorphisms of $\mathcal{W}(M,B)$.

Theorem 5.1: Let (M,B) be given as in Definition 2.1. Let H be the complex Hilbert space obtained from a complex structure on M such that B is the imaginary part of the scalar product. Then in the representation of $\mathcal{W}(M,B)$ specified by the generating functional

$$\rho(z)=\exp(-\tfrac{1}{4}\|z\|^2),$$

where $\|z\|$ is the norm induced by the scalar product in H, the automorphism of the algebra induced by

$$W(z)\to W(Tz)\exp[if(z)]$$

is unitarily implementable if and only if

(1) $f(z)$ is a real-valued continuous real-linear functional on H.

(2) T is a symplectic transformation on H such that $T^*T-1=$ Hilbert-Schmidt operator.

Theorem 5.2: In the Fermi representation of the Weyl algebra of the vector potential, the automorphism of the algebra induced by Poincaré transformations is only implemented by unitary transformations when the transformations are spatial rotations and translations.

Proof: Let $X(a,\Lambda)$ be defined by

$$\phi'=X(a,\Lambda)\phi$$

with ϕ' and ϕ given by (4.11). Then with respect to the scalar product in M given by (5.3),

$$X^*(a,\Lambda)X(a,\Lambda)=\Lambda^T\Lambda\phi.$$

Suppose that $\Lambda=\Lambda(l_1)$ where $\Lambda(l_1)$ is a Lorentz transformation in the direction of the x_1 axis. If $\tanh u$ is the relative velocity of the two frames the vector $\phi''=(\Lambda^T(l_1)\Lambda(l_1)-1)\phi$ will have components

$$\phi^{+\prime\prime}_0(\mathbf{k})=2\sinh^2u\phi^+_0(\mathbf{k})+2\sinh u\cosh u\phi^+_1(\mathbf{k}),$$

$$\phi^{+\prime\prime}_1(\mathbf{k})=2\sinh u\cosh u\phi^+_0(\mathbf{k})+2\sinh^2u\phi^+_1(\mathbf{k}),$$

$$\phi^{+\prime\prime}_2(\mathbf{k})=0,$$

$$\phi^{+\prime\prime}_3(\mathbf{k})=0.$$

For an arbitrary basis of $M(e_1, e_2, \cdots)$, the series

$$\sum_{n=1}^{\infty} \| (\Lambda^T(l_1)\Lambda(l_1) - 1)e_n \|^2$$

will not be convergent, since we can easily find a basis for which $\| (\Lambda^T(l_1)\Lambda(l_1) - 1)e_n \| \not\to 0$ as $n \to \infty$. The operator $[X^*(a, \Lambda(l_1))X(a, \Lambda(l_1)) - 1]$ is therefore not Hilbert–Schmidt on M since an operator O is Hilbert–Schmidt if and only if the series

$$\sum_{n=1}^{\infty} \| Oe_n \|^2$$

converges for an arbitrary basis.[22]

Then by Shale's criterion the automorphism induced by $\Lambda(l_1)$ is not implementable by a unitary transformation.

If Λ is a spatial rotation, then $(R^TR - 1) = 0$ is obviously Hilbert–Schmidt and hence induces an unitarily implementable automorphism.

We can parametrize every Lorentz transformation in the form

$$\Lambda = \Lambda(R_1)\Lambda(l_1)\Lambda(R_2),$$

where $\Lambda(R_1)$ and $\Lambda(R_2)$ are spatial rotations. Hence every Lorentz transformation that is not purely spatial induces automorphisms that cannot be implemented in the Fermi representation.

This calculation shows that the Fermi representation is significantly different from other free field representations. In all approaches to quantum electrodynamics that start with a vector potential, there is some price that has to be paid. For Gupta–Bleuler it is the indefinite metric, for radiation gauge it is the lack of manifest covariance, and here it is the noninvariant vacuum.

The Weyl algebra of the physical photons is the factor algebra $\Delta_c(N)/I$. We can obtain a representation of this algebra from the Fermi representation by procedures that are completely analogous to those used in Sec. 3. It is possible to set up the Schrödinger representation of the field operators and find a unitary transformation to "diagonalize" the elements of $\mathcal{N}$. We can then think of the representation space for the Schrödinger representation as a direct integral of Hilbert spaces labeled by eigenvalues of the supplementary condition operators.

To resolve the problem of the nonnormalizability of the Fermi physical states we interpret them as states in that Hilbert space which are labeled by the eigenvalue zero in the direct integral decomposition, without reference to the large space in which they are "embedded."

A detailed account of the direct integral decomposition of the representation space for the Schrödinger representation is given in Ref. 23.

The discussion that we have given here has only referred to the free field. However, calculations have been performed by Belinfante[3] in perturbation theory using the Fermi representation. We shall not reproduce them here as the only comment that needs to be added to his account is the mathematical justification of his handling of the infinite normalization factor of the wavefunctions. He quite rightly observes that it is consistent to put this factor equal to one. The mathematical justification for this procedure is that this normalization is appropriate for states belonging to the Hilbert space in the direct integral decomposition of the original Hilbert space with respect to the supplementary condition operators. The detailed working out of the usual Feynman rules for perturbation theory is the subject of a separate investigation which, it is hoped, will be presented in a later publication.

6. COVARIANT REPRESENTATION THEORY OF THE FERMI POTENTIAL

Our object in this section is to exhibit a Hilbert space K which carries a representation of the CCR algebra $\mathcal{W}(M, B)$ and a unitary representation of the Poincaré group, and which may be regarded as a covariant formulation of the Fermi method.

One way of thinking about our procedure is to take the algebra $\mathcal{W}(M, B)$ and form the "semidirect product" $\mathcal{W}(M, B) \circledS P$ where P is the Poincaré group regarded as a subgroup of the automorphism group of $\mathcal{W}(M, B)$. For each Poincaré transformation (a, Λ) there is a corresponding automorphism $\theta(a, \Lambda)$ of $\mathcal{W}(M, B)$ and we may therefore define composition in $\mathcal{W}(M, B) \circledS P$ by

$$(A; a, \Lambda)(A'; a', \Lambda') = (A\theta(a, \Lambda)(A'); (a, \Lambda)(a', \Lambda')),$$

where $A, A' \in \mathcal{W}(M, B)$. By constructing a representation of this enlarged object, we obtain a Poincaré invariant theory.

To do this we first consider the representations of $\mathcal{W}(M, B)$ defined by the transformed generating functionals,

$$\rho_{a,\Lambda}(z) = \exp(-\tfrac{1}{4} \| (a, \Lambda)^{-1}z \|_F^2), \quad z \in M.$$

These representations, $\pi_{a,\Lambda}$, act in Hilbert spaces $K_{a,\Lambda}$ with vacua $\Omega_{a,\Lambda}$. We can define maps $Y_{a,\Lambda}: K_{a',\Lambda'} \to K_{(a,\Lambda)(a',\Lambda')}$ by

$$Y_{a,\Lambda}(\pi_{a',\Lambda'}(A)\Omega_{a',\Lambda'}) = \pi_{(a,\Lambda)(a',\Lambda')}(\theta(a, \Lambda)A)\Omega_{(a,\Lambda)(a',\Lambda')}.$$

Then it follows by Theorem 6.1(b) of Shale[21] that $Y_{a,\Lambda}$ is unitary and that

$$Y_{a,\Lambda}W_{a',\Lambda'}(z)Y^{-1}_{a,\Lambda} = W_{(a,\Lambda)(a',\Lambda')}((a, \Lambda) \cdot z)$$

[where $W_{a,\Lambda}(z) \equiv \pi_{a,\Lambda}(\delta_z)$].

It is not difficult to show that

$$Y_{a,\Lambda}Y_{a',\Lambda'} = Y_{(a,\Lambda)(a',\Lambda')}.$$

We note also that not all the spaces $K_{a,\Lambda}$ are distinct, for, as we observed in the previous section, certain automorphisms $\theta(a, \Lambda)$ are unitarily implemented in $K_{0,1}$. So for these automorphisms $Y_{a,\Lambda}$ reduces to a unitary transformation of $K_{0,1}$ to itself.

Following a suggestion of Wightman (private communication) we consider the direct integral

$$K = \int_X K_{\Lambda(p)}\, d\mu(p),$$

where X is the homogeneous space SO(3, 1)/SO(3) having Lorentz invariant measure $d\mu$ and parametrized by

"momentum" vectors p, That is, we identify X with a mass shell such that $p_0^2 - \mathbf{p}^2 = 1$ and $\Lambda(p)$ is a Lorentz transformation mapping the vector $\hat{p} = (0, 0, 0, 1)$ to $(\mathbf{p}, p_0) = p$. Note that $K_{\Lambda(\hat{p})} \cong K_{0,1}$. Here we are using the fact that $K_{a,\Lambda} = K_{a',\Lambda'}$ if and only if $(a, \Lambda)^{-1}(a', \Lambda')$ is an element of the subgroup G of P consisting of translations and spatial rotations.

Elements of K are equivalence classes of μ-measurable functions F on X such that $F(p) \in K_{\Lambda(p)}$ and satisfying

$$\int_X \| F(p) \|_p^2 \, d\mu(p) < \infty,$$

where $\| \ \|_p$ denotes the norm in $K_{\Lambda(p)}$.

A representation of $\mathcal{U}(M, B) \circledS P$ in K is defined by

$$\pi(A; a, \Lambda) F(p) = \pi_p(A) Y_{a,\Lambda} F(\Lambda^{-1} \cdot p)$$

where π_p denotes the representation of $\mathcal{U}(M, B)$ in $K_{\Lambda(p)}$. Let us record some observations about this definition. Firstly it is well defined since $F(\Lambda^{-1}p)$ is in $K_{\Lambda(p')}$ where $p' = \Lambda^{-1}\Lambda(p) \cdot \hat{p}$, hence $Y_{a,\Lambda} F(\Lambda^{-1}p)$ is in $K_{\Lambda(p)}$ (noting that translations are unitarily implemented in all the spaces $K_{\Lambda(p)}$, $p \in X$) and thus $\pi_p(A) Y_{a,\Lambda} F(\Lambda^{-1}p)$ is in $K_{\Lambda(p)}$. Secondly, on restriction to $\mathcal{U}(M, B)$ we obtain a direct integral of *inequivalent* representations π_p [this is a corollary of the nonimplementability of the $\theta(0, \Lambda(p))$]. Thirdly, on restriction to the Poincaré group we obtain a unitary representation which may be regarded as an induced representation. In fact in $K_{0,1}$, we have a unitary representation R of G. We observe that the properties of the $Y_{a,\Lambda}$ endow K with the structure of a Hilbert P-bundle (see Varadarajan[24]) and hence we can identify the representation of P defined by

$$(a, \Lambda) \cdot F(p) = Y_{a,\Lambda} F(\Lambda^{-1} \cdot p)$$

as that induced by the representation R of G.

The question naturally arises as to whether π is an irreducible representation of the extended object $\mathcal{U}(M, B) \circledS P$. Denote by π_0 the representation of $\mathcal{U}(M, B)$ defined by

$$\pi_0(A) F(p) = \pi_p(A) F(p), \quad A \in \mathcal{U}(M, B).$$

Consider the central decomposition of the commuting algebra of π_0. Clearly this is precisely the direct integral decomposition $K = \int K_p \, d\mu(p)$, since all the π_p are inequivalent. Thus the operators of the form $(BF)(p) = b(p)F(p)$ where $b(p)$ is a complex number for all p form the center $\mathcal{Z}$ of the commuting algebra. But any operator which commutes with these multiplication operators is itself a multiplication operator of this type (remembering that each π_p is irreducible). Thus the commuting algebra of π_0 is just $\mathcal{Z}$ itself. Now we use the representation of the Poincaré group to show that the only operator B of the above type which commutes with every Poincaré transformation is a multiple of the identity operator. Hence the representation of $\mathcal{U}(M, B) \circledS P$ is irreducible. Thus we have introduced no new constants of the motion. In any calculation in which one wishes to use the Fermi method one simply needs to carry along a momentum label p which indicates which representation of the Weyl algebra one is using. This label could be interpreted as labeling an observer in a particular Lorentz frame, however, we hope to investigate other possible interpretations of this construction in the future.

It is clear that there is some connection between this construction and that used in discussions of rotational symmetry breaking in say the Ising model. However the fact that X is noncompact leads to the important difference that there is no Lorentz invariant vacuum in K. Equivalently, this representation of $\mathcal{U}(M, B)$ is not obtained from a generating functional

$$\rho(A) = \int_X \rho_p(A) \, d\mu(p)$$

as this expression diverges in general. Nevertheless, there is a Poincaré invariant subspace of K which contains all the "vacua." It consists of all functions of the form

$$F(p) = c(p) \Omega_p,$$

where $c(p)$ is a complex number for each $p \in X$, Ω_p is the vacuum for $K_{\Lambda(p)}$ and we require

$$\int_X |c(p)|^2 \, d\mu(p) < \infty.$$

7. CONCLUDING REMARKS

The most significant aspect of these investigations is that they present a challenge to the point of view that the Fermi method of quantizing the electromagnetic field is inconsistent. This point of view is so widely accepted that in investigations in field theory it is not considered necessary to take account of the Fermi method. It is, however, mathematically consistent and should be considered in conjunction with other quantization schemes in any analysis of a general nature. In particular, the techniques described here apply equally well to other field theories which are presented in terms of fields with redundant components and concomitant supplementary conditions, for example, tensor field descriptions of higher spin particles.

Also, it may well be the case that the Fermi method is a convenient representation for particular applications. We have shown that it can be easily included in the Segal–Manuceau formalism and hence should be better adapted than the Gupta–Bleuler representation for applications requiring the techniques of C^*-algebras.

ACKNOWLEDGMENT

One of us (A. L. C.) gratefully acknowledges the support of a Rothman's Fellowship.

*Present address: Center for Applied Mathematics, Michigan State University, E. Lansing, Michigan 48823.

[1] F. Strocchi, Phys. Rev. 162, 1429 (1967).
[2] E. Fermi, Atti. Accad. Lincei 9, 881 (1929); 12, 431 (1930); Rev. Mod. Phys. 4, 125 (1932).
[3] S. T. Ma, Phys. Rev. 75, 535 (1949); 80, 729 (1950); F. J. Belinfante, Phys. Rev. 76, 76 (1949).
[4] C. A. Hurst, Nuovo Cimento 21, 274 (1961).
[5] I. E. Segal, Mat. Fys. Medd. Dan. Vid. Selsk. 31, (12) (1959); Can. J. Math. 13, 1 (1962); Ill. J. Math. 6, 500 (1962).

[6]J. Manuceau, Ann. Inst. Henri Poincaré 8, 139—161 (1968).
[7]W. Heisenberg and W. Pauli, Z. Phys. 56, 1 (1929); 59, 168 (1930).
[8]S. Gupta, Proc. Phys. Soc. London A 63, 681 (1950); K. Bleuler, Helv. Phys. Acta. 23, 567 (1950).
[9]G.G. Emch, *Algebraic Methods in Statistical Mechanics and Quantum Field Theory* (Wiley-Interscience, New York, 1972).
[10]I.E. Segal, *Mathematical Problems of Relativistic Physics* (Am. Math. Soc. Providence, R.I. 1963); in *Cargese Lectures in Theoretical Physics* 1965, edited by F. Lurcat (Gordon and Breach, New York, 1967).
[11]M.H. Stone, Ann. Math. 33, 643 (1932).
[12]J. von Neumann, Math. Ann. 104, 570 (1931).
[13]Z. Takeda, Tohoku Math. J. 7, 68 (1955).
[14]I. Gel'fand and M. Naimark, Mat. Sb. N.S. 12, 197 (1943).
[15]I.E. Segal, Ann. Math. 48, 930 (1947).
[16]P.A.M. Dirac, *The Principles of Quantum Mechanics* (Oxford U.P., Oxford, 1930), 1st ed.; (Clarenden, Oxford, 1958), 4th ed.
[17]A.O. Barut and R. Raczka, Ann. Inst. Henri Poincaré 17, 111 (1972).
[18]S. Sakai, *C*-Algebras and W*-Algebras* (Springer, Berlin, 1971).
[19]J. Dixmier, *Les C*-algèbres et leurs representations* (Gauthier-Villars, Paris, 1964).
[20]M.A. Naimark, *Normed Algebras* (Wolters-Noordhoff, Groningen, 1972).
[21]D. Shale, Trans. Am. Math. Soc. 103, 149 (1962).
[22]R. Schatten, *Norm Ideals of Completely Continuous Operators* (Springer, Berlin, 1960).
[23]J.M. Gaffney, Ph.D. thesis, University of Adelaide, Adelaid (1974).
[24]V.S. Varadarajan, *Geometry of Quantum Theory II* (Van Nostrand, Princeton, 1970).

Gauge equivalence of the electrodynamics of charged bosons

R. Sohn* and K. Haller

Department of Physics, University of Connecticut, Storrs, Connecticut 06268
(Received 12 July 1976)

The quantum electrodynamics of charged scalar and vector bosons is formulated in the Lorentz gauge, and the effect of the charged particle–photon interaction on the subsidiary condition is explicitly taken into account. The results are extensions of earlier work on spinor quantum electrodynamics, but the presence of seagull vertices and anomalous current commutators in the case of the charged bosons make the extensions nontrivial. An operator gauge transformation that encompasses equations of motion as well as the commutator algebra of the field operators is developed; it is used to transform the theory from the Lorentz gauge to the Coulomb gauge.

I. INTRODUCTION

The role of the subsidiary condition in the Lorentz gauge formulation of quantum electrodynamics (qed) has previously been studied[1] for the case of charged spinors interacting with the electromagnetic field. As a result of that work, a very close relationship was demonstrated between the Coulomb gauge and Lorentz gauge versions of the theory, when care was taken in the latter to include the interactions of the charged particles and the electromagnetic field in formulating the subsidiary condition that defines the physically admissible states. One of the outgrowths of this work was the establishment of an operator transformation that relates the Lorentz and Coulomb gauges; another was the demonstration that all transition amplitudes, on and off shell, including all renormalization constants, are identical in the Coulomb gauge and the so-called "interacting subsidiary condition" (ISC) version of the Lorentz gauge.[2] It is the purpose of the present work to extend these results to the electrodynamics of charged scalar and vector bosons. There are two reasons why this work was undertaken. In the first place, the qed of charged spinors has some very special simplifying features which are absent in the qed of charged bosons. These include the vanishing of the equal time commutators of j_4 with the spatial currents $\mathbf{j}$ and the absence of seagull terms from the charged particle–photon vertices. It seemed desirable to examine whether the relationships that were established for spinor qed would survive when the "anomalous" current commutators and the seagull terms of boson qed were included. One of the results of this work has been that the relationships established for spinor qed obtain for boson qed and that the effects of the anomalous current commutators and the seagull components compensate each other so as to protect the identities that were previously demonstrated for spinor qed only. The other reason for this study was the need to have a point of departure for studying the Guralnik–Higgs–Kibble effect by developing a canonical formulation of the electrodynamics of a charged scalar field that has a nonvanishing vacuum expectation value.[3]

II. THE ELECTRODYNAMICS OF CHARGED SCALAR BOSONS

The Lagrangian density for charged scalar bosons interacting with an electromagnetic field in the so-called "generalized Lorentz gauge" (GLG)[4] is

$$\mathcal{L} = -\tfrac{1}{4}F_{\mu\nu}F_{\mu\nu} - G(x)\partial_\mu A_\mu + \tfrac{1}{2}(1-\gamma)G^2(x) - D_\mu^\dagger\phi^\dagger D_\mu\phi - m^2\phi^\dagger\phi, \tag{2.1}$$

where $D_\mu = \partial_\mu - ieA_\mu$ is the covariant derivative of qed; $G(x)$ is a Lagrange multiplier field which prevents Π_4, the momentum conjugate to A_4, from vanishing; and γ is a parameter which specifies the particular member of the GLG. $\gamma = 0,\ 1,\ -2$ correspond to the Feynman, Landau, and Yennie–Fried gauges, respectively. In this paper, we will consider the Feynman gauge, for which $\gamma = 0$. Expanding the Lagrangian density, Eq. (2.1), into free and interacting parts, $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_I$, we find

$$\mathcal{L}_0 = -\tfrac{1}{4}F_{\mu\nu}F_{\mu\nu} - G(x)\,\partial_\mu A_\mu + \tfrac{1}{2}G^2(x) - \partial_\mu\phi^\dagger\partial_\mu\phi - m^2\phi^\dagger\phi \tag{2.2a}$$

and

$$\mathcal{L}_I = ieA_\mu[\phi\overleftrightarrow{\partial}_\mu\phi^\dagger] - e^2A_\mu A_\mu\phi^\dagger\phi. \tag{2.2b}$$

The conserved electromagnetic current, determined by the U(1) invariance of the total Lagrangian density (in the subspace in which $G = 0$) Eq. (2.1) is

$$j_\mu = ie\phi\overleftrightarrow{\partial}_\mu\phi^\dagger - 2e^2A_\mu\phi^\dagger\phi. \tag{2.3}$$

j_μ may also be calculated by[5]

$$j_\mu = \partial\mathcal{L}_I/\partial A_\mu. \tag{2.4}$$

Thus, we may also calculate the interaction Lagrangian as

$$\mathcal{L}_I = \int j_\mu[A]\,dA_\mu, \tag{2.5}$$

where $j_\mu[A]$ is treated as a functional of A_μ; this gives

$$\mathcal{L}_I = A_\mu j_\mu + e^2A_\mu A_\mu\phi^\dagger\phi$$

in agreement with Eq. (2.2b).

The Euler–Lagrange equations of motion generated by the Lagrangian density are

$$D_\mu^\dagger D_\mu^\dagger\phi^\dagger - m^2\phi^\dagger = 0,\quad D_\mu D_\mu\phi - m^2\phi = 0,$$
$$G = \partial_\mu A_\mu,\quad \text{and}\quad \partial_\mu F_{\mu\nu} + \partial_\nu G = \Box A_\nu = -j_\nu. \tag{2.6}$$

By the conservation of the current j_μ, Eq. (2.3), G is seen to be a free field:

$$\Box G = 0. \tag{2.7}$$

Copyright © 1977 American Institute of Physics

The conjugate momenta corresponding to A_j, A_4, ϕ, and $\phi^\dagger$ respectively are

$$\Pi_j = \partial_0 A_j - i\partial_j A_4, \quad \Pi_4 = iG, \qquad P = iD_4^\dagger \phi^\dagger, \quad \text{and} \quad P^\dagger = iD_4 \phi. \tag{2.8}$$

The Hamiltonian density that results from the canonical formalism when time derivatives are eliminated by using Eqs. (2.8) is given by

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_I$$

with

$$\mathcal{H}_0 = \tfrac{1}{4}F_{jk}F_{jk} + \tfrac{1}{2}\Pi_\mu^2 - i[A_4 \nabla \cdot \Pi + \Pi_4 \nabla \cdot \mathbf{A}] + P^\dagger P + \nabla\phi^\dagger \cdot \nabla\phi + m^2\phi^\dagger\phi \tag{2.9a}$$

and

$$\mathcal{H}_I = eA_4[P^\dagger\phi^\dagger - P\phi] - ie\mathbf{A} \cdot [\phi \overleftrightarrow{\nabla} \phi^\dagger] + e^2 \mathbf{A} \cdot \mathbf{A}\phi^\dagger\phi. \tag{2.9b}$$

Equations (2.8) can also be used to eliminate the time derivatives of field variables in the current density j_μ. The components of j_μ then take the form

$$\mathbf{j} = ie\phi \overleftrightarrow{\nabla} \phi^\dagger - 2e^2 \mathbf{A}\phi^\dagger\phi, \tag{2.10a}$$

$$\rho = ie[P^\dagger\phi^\dagger - P\phi]. \tag{2.10b}$$

We notice that although $\mathbf{j}$ contains the electromagnetic potentials explicitly, ρ does not. The A_4's in ρ are buried in the definitions of the P's. Since A_4 occurs in $\mathcal{H}_I$ only in the form $-iA_4\rho$ [Eq. (2.9b)], this absence of explicit A_4 dependence in ρ manifests itself in the fact that the Hamiltonian contains no A_4 seagull terms; i.e., no terms of the form $e^2A_4^2\phi^\dagger\phi$, etc. Since in the Lorentz gauge all field quantities ξ have a well-defined conjugate momentum Π_ξ and this formulation therefore leads to canonical equal time commutation rules

$$[\xi(\mathbf{x}), \Pi_\xi(\mathbf{y})] = i\delta(\mathbf{x} - \mathbf{y}),$$

ρ trivially commutes with all components of A_μ. However, the commutation rules of ρ with $\mathbf{j}$ are different in spinor and scalar qed. In spinor qed ρ commutes with all components of $\mathbf{j}$, but in scalar qed we have that

$$[\rho(\mathbf{x}), j_l(\mathbf{y})] = -2ie^2\partial[\phi^\dagger(\mathbf{x})\phi(\mathbf{x})\delta(\mathbf{x} - \mathbf{y})]/\partial x_l. \tag{2.11}$$

This is an illustration of a general result, due to Schwinger and to Brown,[6] which states that

$$[\rho(\mathbf{x}), j_l(\mathbf{y})] = i\partial[\delta j_l(\mathbf{y})/\delta A_k(\mathbf{x})]/\partial x_k. \tag{2.12}$$

However, the total charge $Q = \int d\mathbf{x}\,\rho(\mathbf{x})$ still commutes with $\mathbf{j}(\mathbf{x})$.

There is an additional difference between scalar and spinor qed in the form of the interaction Hamiltonian. For scalar qed, we have

$$\mathcal{H}_I = -j_\mu A_\mu - e^2 \mathbf{A} \cdot \mathbf{A}\phi^\dagger\phi \tag{2.13}$$

rather than $-j_\mu A_\mu$, as in the case of spinor qed. The terms in $\mathcal{H}_I$ that are quadratic in e, which include the contributions from the 3-space part of $-j_\mu A_\mu$, i.e., from $-\mathbf{j} \cdot \mathbf{A}$, correspond to the seagull vertices; as noted, seagull terms for A_4 do not occur.

Since $\Pi_4(x)$ is a free field, as seen from Eqs. (2.7) and (2.8), it can be decomposed into invariant positive and negative frequency parts within the indefinite metric space, $\Pi_4^\pm(x)$, and a Gupta–Bleuler type subsidiary condition[7] can be imposed, defining the states within the physical subspace:

$$\Pi_4^+(x)\,|\,\nu\rangle = 0 \tag{2.14a}$$

or, equivalently,

$$\langle \nu^\star \,|\, \Pi_4^-(x) = 0 \tag{2.14b}$$

since $[\Pi_4^+(x)]^\star = -\Pi_4^-(x)$,[8] where $\star$ denotes indefinite metric space adjoint. Π_4^+ and Π_4^- are evaluated by using the Cauchy solution to d'Alembert's equation[9]:

$$\Pi_4^\pm(x) = \int_{x_0'=0} d\mathbf{x}'\{[\partial\Delta^\pm(x - x')/\partial x_0'][\Pi_4(\mathbf{x}') - i\Delta^\pm(x - x')[H, \Pi_4(\mathbf{x}')]\}, \tag{2.15}$$

where $\Delta^\pm(x - x')$ are the positive and negative frequency parts of the invariant Green's function for massless fields:

$$\Delta^\pm(x - x') = \mp i(2\pi)^{-3} \int d\mathbf{p}\,[2\,|\,\mathbf{p}\,|\,]^{-1} \exp[\pm ip \cdot (x - x')]. \tag{2.16}$$

The time derivative of Π_4, $i[H, \Pi_4]$, is given by

$$i[H, \Pi_4] = i\nabla \cdot \Pi + i\rho. \tag{2.17}$$

As in the case of spinor qed, the positive and negative frequency parts of Π_4 are

$$\Pi_4^\pm(x) = \mp(2\pi)^{-3/2} \int d\mathbf{k}\,[\,|\,\mathbf{k}\,|\,]^{1/2} \exp[\pm ik \cdot x]\,\Omega^\pm(\mathbf{k}) \tag{2.18}$$

with

$$\Omega^+(\mathbf{k}) = a_Q(\mathbf{k}) + \rho(\mathbf{k})/(2\,|\,\mathbf{k}\,|^{3/2}) \tag{2.19a}$$

and

$$\Omega^-(\mathbf{k}) = a_R^\dagger(\mathbf{k}) + \rho(-\mathbf{k})/(2\,|\,\mathbf{k}\,|^{3/2}). \tag{2.19b}$$

We have used the Q and R formalism of Haller *et al.*[10] $a_R^\dagger$ is the indefinite metric space adjoint of a_Q, and commutes with it. The Gupta–Bleuler subsidiary condition defining the physical subspace is

$$\Omega^+(\mathbf{k})\,|\,\nu\rangle = 0. \tag{2.20}$$

In establishing that Ω^+ and Ω^- are the invariant positive and negative frequency parts of $\Pi_4(x)$, we note that, unlike spinor qed, in scalar qed $[H_I, \Omega^\pm(\mathbf{k})]$ has contributions from the anomalous current commutators, since $\Omega^\pm(\mathbf{k})$ contains ρ, and from the seagull terms; however, these contributions cancel each other, so that

$$[H, \Omega^\pm(\mathbf{k})] = \mp\,|\,\mathbf{k}\,|\,\Omega^\pm(\mathbf{k}). \tag{2.21}$$

III. THE LORENTZ GAUGE IN THE TRANSFORMED REPRESENTATION

As was the case in spinor qed, the subsidiary condition can be simplified by carrying out a transformation, unitary within the indefinite metric space (pseudounitary), that relates Ω^+ and Ω^- to a_Q and $a_R^\dagger$, respectively:

$$\Omega^+(\mathbf{k}) = e^{-D} a_Q(\mathbf{k})\, e^D \tag{3.1a}$$

and

$$\Omega^-(\mathbf{k}) = e^{-D} a_R^\dagger(\mathbf{k})\, e^D, \tag{3.1b}$$

where

$$D = -\tfrac{1}{2} \int d\mathbf{k}\,[\,|\,\mathbf{k}\,|\,]^{-3/2} \rho(\mathbf{k})[a_R(-\mathbf{k}) - a_Q^\dagger(\mathbf{k})]. \tag{3.2}$$

D is anti-self-adjoint within the indefinite metric space:

$$D^\star = -D \tag{3.3}$$

so that

$$[e^D]^\star [e^D] = 1. \tag{3.4}$$

By transforming all operators by

$$\hat{\xi} = e^D \xi e^{-D} \tag{3.5}$$

and all states by

$$|n\rangle = e^D |\nu\rangle \tag{3.6}$$

(this new representation of the Lorentz gauge is called the "hatted" representation), so that states have the same norm and operators have the same matrix elements before and after the transformation, the Gupta–Bleuler subsidiary condition becomes

$$a_Q(\mathbf{k})\,|n\rangle = 0 \tag{3.7a}$$

and

$$\langle n^\star |\, a_R^\dagger(\mathbf{k}) = 0. \tag{3.7b}$$

Since $\Omega^*(\mathbf{k})$ has the form $a_Q(\mathbf{k})$ in the hatted representation, the "forbidden" states in that case are the states with any number of Q-type photons; i. e., states of the form $[a_Q^\dagger]^n |\,0\rangle$. These states are zero norm "ghosts." The allowed physical states are the transverse photon states and the R-type ghost states; i. e., the states $a_R^\dagger |\,0\rangle$ and their many particle analogs. The allowed states form a positive semidefinite metric space; this space is degenerate, since the R-type photon states have zero norm and zero inner product with all transverse photons. The transformed (hatted) Hamiltonian is

$$\hat{H} = e^D H e^{-D} = H + [D, H] + (1/2!)[D, [D, H]]. \tag{3.8}$$

$[D, [D, [D, H]]]$ and all higher commutator terms vanish. $\hat{H}$ is given by

$$\hat{H} = H_0 + H_C + H_{I,T} + H_Q, \tag{3.9}$$

where

$$H_0 = \int d\mathbf{k}\,|\mathbf{k}| \left\{ \sum_{i=1}^{2} a_i^\dagger(\mathbf{k})\, a_i(\mathbf{k}) + a_R^\dagger(\mathbf{k})\, a_R(\mathbf{k}) + a_Q^\dagger(\mathbf{k})\, a_Q(\mathbf{k}) \right\} + \int d\mathbf{p}\, p_0 \{\alpha^\dagger(\mathbf{p})\,\alpha(\mathbf{p}) + \beta^\dagger(\mathbf{p})\,\beta(\mathbf{p})\}, \tag{3.10a}$$

$$H_C = (8\pi)^{-1} \int d\mathbf{x}\, d\mathbf{y}\, \rho(\mathbf{x})\, \rho(\mathbf{y})\, |\mathbf{x} - \mathbf{y}|^{-1}, \tag{3.10b}$$

$$\begin{aligned} H_{I,T} = {} & -e(2\pi)^{-3/2} \int d\mathbf{k}\, d\mathbf{p}\, d\mathbf{p}' \,[8\,|\mathbf{k}|\; p_0 p_0']^{-1/2} \sum_{i=1}^{2} \hat{e}^i(\mathbf{k}) \cdot [\mathbf{p} + \mathbf{p}']\, \delta(\mathbf{k} + \mathbf{p} - \mathbf{p}')[a_i(\mathbf{k}) + (-1)^i\, a_i^\dagger(-\mathbf{k})] \\ & \times [\alpha^\dagger(\mathbf{p}') + \beta(-\mathbf{p}')][\alpha(\mathbf{p}) + \beta^\dagger(-\mathbf{p})] + e^2(2\pi)^{-3} \int d\mathbf{k}\, d\mathbf{k}'\, d\mathbf{p}\, d\mathbf{p}' [4(|\mathbf{k}|\;|\mathbf{k}'|\; p_0 p_0')^{1/2}]^{-1} \\ & \times \delta(\mathbf{k} + \mathbf{k}' + \mathbf{p} - \mathbf{p}') \sum_{i,j=1}^{2} \hat{e}^i(\mathbf{k}) \cdot \hat{e}^j(\mathbf{k}')[\alpha^\dagger(\mathbf{p}') + \beta(-\mathbf{p}')][\alpha(\mathbf{p}) + \beta^\dagger(-\mathbf{p})][a_i(\mathbf{k}) + (-1)^i\, a_i^\dagger(-\mathbf{k}')][a_j(\mathbf{k}') + (-1)^j\, a_j^\dagger(-\mathbf{k}')], \end{aligned} \tag{3.10c}$$

$$\begin{aligned} H_Q = {} & -\int d\mathbf{k}[2\,|\mathbf{k}|^{1/2}]^{-1} \rho(\mathbf{k})[a_Q(-\mathbf{k}) + a_R^\dagger(\mathbf{k})] - e(2\pi)^{-3/2} \int d\mathbf{k}\, d\mathbf{p}\, d\mathbf{p}' [4(|\mathbf{k}|\; p_0 p_0')^{1/2}]^{-1}\, \hat{k} \cdot [\mathbf{p} + \mathbf{p}'] \\ & \times \delta(\mathbf{k} + \mathbf{p} - \mathbf{p}')[a_Q(\mathbf{k}) + a_R^\dagger(\mathbf{k})][\alpha^\dagger(\mathbf{p}') + \beta(-\mathbf{p}')][\alpha(\mathbf{p}) + \beta^\dagger(-\mathbf{p})] + e^2(2\pi)^{-3} \int d\mathbf{k}\, d\mathbf{k}'\, d\mathbf{p}\, d\mathbf{p}' \\ & \times [8(|\mathbf{k}|\;|\mathbf{k}'|\; p_0 p_0')^{1/2}]^{-1}\, \hat{k} \cdot \hat{k}'\, \delta(\mathbf{k} + \mathbf{k}' + \mathbf{p} - \mathbf{p}')[\alpha^\dagger(\mathbf{p}') + \beta(-\mathbf{p}')][\alpha(\mathbf{p}) + \beta^\dagger(-\mathbf{p})] \\ & \times [a_Q(\mathbf{k})\, a_Q(\mathbf{k}') - 2a_Q(\mathbf{k})\, a_R^\dagger(-\mathbf{k}') + a_R^\dagger(-\mathbf{k})\, a_R^\dagger(-\mathbf{k}')] + e^2(2\pi)^{-3} \int d\mathbf{k}\, d\mathbf{k}'\, d\mathbf{p}\, d\mathbf{p}' [8\,|\mathbf{k}|\;|\mathbf{k}'|\; p_0 p_0']^{-1/2} \\ & \times \sum_{i=1}^{2} \hat{e}^i(\mathbf{k}) \cdot \hat{k}'\, \delta(\mathbf{k} + \mathbf{k}' + \mathbf{p} - \mathbf{p}')[\alpha^\dagger(\mathbf{p}') + \beta(-\mathbf{p}')][\alpha(\mathbf{p}) + \beta^\dagger(-\mathbf{p})][a_i(\mathbf{k}) + (-1)^i\, a_i^\dagger(-\mathbf{k})][a_Q(\mathbf{k}') + a_R^\dagger(\mathbf{k}')]. \end{aligned} \tag{3.10d}$$

We have used the following representation for ϕ and P:

$$\phi(\mathbf{x}) = (2\pi)^{-3/2} \int d\mathbf{p}\,[2p_0]^{-1/2} [\alpha(\mathbf{p}) \exp(i\mathbf{p} \cdot \mathbf{x}) + \beta^\dagger(\mathbf{p}) \exp(-i\mathbf{p} \cdot \mathbf{x})] \tag{3.11a}$$

and

$$P(\mathbf{x}) = i(2\pi)^{-3/2} \int d\mathbf{p}\,(p_0/2)^{1/2} [\alpha^\dagger(\mathbf{p}) \exp(-i\mathbf{p} \cdot \mathbf{x}) - \beta(\mathbf{p}) \exp(+i\mathbf{p} \cdot \mathbf{x})], \tag{3.11b}$$

where α, $\alpha^\dagger$, etc., are the annihilation and creation operators for the elementary excitations of the charged scalar field, and $p_0 = [|\mathbf{p}|^2 + m^2]^{1/2}$.

Physical states do not contain Q-type photons, as guaranteed by the subsidiary condition, Eq. (3.7a), and $\hat{H}$ does not contain any mechanism for generating them. In particular, of all the seagull terms contained in the original Hamiltonian H, those which could generate Q-type photons in final states must be eliminated by the transformation of H into $\hat{H}$. Inspection of Eqs. (3.10) shows that they indeed have disappeared.

In the case of boson qed the seagull terms in H_I and the anomalous current commutators, Eq. (2.12), cooperate to eliminate all terms from $\hat{H}_I$ (with $\hat{H}_I = \hat{H} - H_0$) that contain either $a_Q^\dagger$ or a_R operators. In transforming to $\hat{H}$, the anomalous current commutators arise in the term $\frac{1}{2}[D, [D, H]]$ of $e^D H e^{-D}$, and serve to cancel off the forbidden seagull terms that remain in $H + [D, H]$.

As in related work on spinor qed, H_0 is the Hamiltonian for noninteracting photons and charged scalar bosons. H_C is the nonlocal contact term associated with the Coulomb gauge; ρ is the charge density for charged scalar particles. $H_{I,T}$ is the interaction term for transverse photons and charged scalar bosons. It contains transverse photon vertices and purely transverse seagull terms. H_Q contains terms to destroy one Q or create one R type photon; seagull terms to destroy a Q and create a transverse photon, to destroy a transverse and create an R photon, and to destroy a Q and create an R photon. However, H_Q contains no mechanism to create and destroy a Q photon (a Q-seagull) or to create and destroy an R photon (an R-seagull) or to destroy an R and create a Q-type photon.

H_Q, the part of $\hat{H}$ which contains the photon operators a_Q and $a_R^\dagger$, but not the operators $a_Q^\dagger$ or a_R, is necessary

for the preservation of the equations of motion for qed in the Lorentz gauge, as well as for the manifest covariance of the theory. However, H_Q plays no role in the calculation of scattering amplitudes. The temporal evolution of wavefunctions in the physical subspace is completely unaffected by H_Q except for the fact that the latter generates R-type ghosts as components of wavefunctions. H_Q may in principle destroy the unphysical Q-type ghosts, but there are none in any wavefunction that represents a physical system, since such a wavefunction would not have Q-type ghosts initially, and since there is no possibility that $\exp(-i\hat{H}t)$ could generate them later. H_Q can also never mediate any virtual photon exchange processes in which it first creates and then destroys a photon, or creates a photon which is then destroyed by some other part of $\hat{H}$, or destroys a photon created by another part of $\hat{H}$. This is because $[a_Q(\mathbf{k}), a_R^\dagger(\mathbf{k}')] = 0$, and because only $a_Q(\mathbf{k})$ and $a_R^\dagger(\mathbf{k})$ operators occur in H_Q. The $a_Q^\dagger(\mathbf{k})$ and $a_R(\mathbf{k})$ operators with which H_Q could combine to mediate virtual photon exchange processes do not arise in either H_Q or in any other part of H_I. They occur in H_0 only bilinearly and only to count Q and R photon energy and there also cannot mediate exchange processes. This feature of the theory makes it manifestly unitary within the physical subspace. The S matrix must always be formally unitary because of the self-adjointness, in the indefinite metric space, of H (and $\hat{H}$).

For these reasons, H_Q may be completely amputated from $\hat{H}$ without in any way affecting the dynamics. The hatted representation of the ISC is then dynamically equivalent to the Coulomb gauge. For instance, the probability amplitude of observing any physically observable state $|n_\alpha\rangle$, i.e., the matrix element $\langle n_\alpha | \psi(t)\rangle$, is identical when the time displacement operator is $\exp(-i\hat{H}t)$ or when it is $\exp(-iH_\xi t)$, where H_ξ is the Hamiltonian for the Coulomb gauge formulation.

IV. THE COULOMB GAUGE

To transform the theory to the Coulomb gauge, so that the fields satisfy the Coulomb gauge commutation relations, we first rotate all operators into their hatted representation.[11] This does not constitute a gauge transformation, but merely a pseudounitary transformation of the Lorentz gauge variables. Since the charge density commutes with D, $\hat{\rho} = \rho$. Some examples of hatted variables are

$$\begin{aligned}\hat{A}_4(\mathbf{x}) &= i(2\pi)^{-3/2}\int d\mathbf{k}\,[2|\mathbf{k}|^{1/2}]^{-1}\\ &\times\{[a_R(\mathbf{k}) - a_Q(\mathbf{k})]\exp(i\mathbf{k}\cdot\mathbf{x}) + [a_Q^\dagger(\mathbf{k})\\ &- a_R^\dagger(\mathbf{k})]\exp(-i\mathbf{k}\cdot\mathbf{x})\} + (2\pi)^{-3/2}\int d\mathbf{k}\\ &\times[2|\mathbf{k}|^2]^{-1} j_4(\mathbf{k})\exp(i\mathbf{k}\cdot\mathbf{x}),\\ \hat{\mathbf{A}}(\mathbf{x}) &= (2\pi)^{-3/2}\int d\mathbf{k}\,[2|\mathbf{k}|]^{-1/2}\sum_{i=1}^{2}\hat{e}^i(\mathbf{k})\\ &\times[a_i(\mathbf{k})\exp(i\mathbf{k}\cdot\mathbf{x}) + a_i^\dagger(\mathbf{k})\exp(-i\mathbf{k}\cdot\mathbf{x})]\\ &+ (2\pi)^{-3/2}\int d\mathbf{k}\,[2|\mathbf{k}|^{1/2}]^{-1}\hat{k}\\ &\times\{[a_R(\mathbf{k}) + a_Q(\mathbf{k})]\exp(i\mathbf{k}\cdot\mathbf{x}) + [a_Q^\dagger(\mathbf{k}) + a_R^\dagger(\mathbf{k})]\\ &\times\exp(-i\mathbf{k}\cdot\mathbf{x})\} \qquad (4.1)\end{aligned}$$

and

$$\hat{\phi}(\mathbf{x}) = \exp[g(\mathbf{x})]\,\phi(\mathbf{x})$$

with

$$\begin{aligned}g(\mathbf{x}) &= e(2\pi)^{-3/2}\int d\mathbf{k}\,[2|\mathbf{k}|^{3/2}]^{-1}[a_R(\mathbf{k})\exp(i\mathbf{k}\cdot\mathbf{x})\\ &- a_Q^\dagger(\mathbf{k})\exp(-i\mathbf{k}\cdot\mathbf{x})]\end{aligned}$$

since

$$[\rho(\mathbf{x}), \phi(\mathbf{y})] = -e\phi(\mathbf{x})\,\delta(\mathbf{x}-\mathbf{y}).$$

We now seek an operator gauge transformation

$$\mathbf{A}^C = \hat{\mathbf{A}} + \nabla X^C, \quad A_0^C = \hat{A}_0 - i[\hat{H}, X^C] \qquad (4.2)$$

such that

$$\nabla\cdot\mathbf{A}^C = 0. \qquad (4.3)$$

Equations (4.1), (4.2), and (4.3) are solved to give

$$\begin{aligned}X^C &= i(2\pi)^{-3/2}\int d\mathbf{k}\,[2|\mathbf{k}|^{3/2}]^{-1}\{[a_R(\mathbf{k}) + a_Q(\mathbf{k})]\exp(i\mathbf{k}\cdot\mathbf{x})\\ &- [a_Q^\dagger(\mathbf{k}) + a_R^\dagger(\mathbf{k})]\exp(-i\mathbf{k}\cdot\mathbf{x})\}.\end{aligned} \qquad (4.4)$$

To gauge transform the charged scalar field such that

$$(\partial_\mu - ieA_\mu^C)(\partial_\mu - jeA_\mu^C)\,\phi^C - m^2\phi^C = 0, \qquad (4.5)$$

we seek an operator $G^{C,L}(x)$ such that

$$\phi^C(\mathbf{x}) = G^{C,L}(x)\,\hat{\phi}(\mathbf{x}); \qquad (4.6)$$

$G^{C,L}$ must satisfy

$$\partial_\mu G^{C,L} = ie[A_\mu^C G^{C,L} - G^{C,L}\hat{A}_\mu]. \qquad (4.7)$$

$G^{C,L}$ is given by

$$G^{C,L}(x) = \lim_{\epsilon\to 0} T\{\exp[ie\int dt'\,\dot{X}^C(x,t')\exp(-\epsilon|t'|)]\} \qquad (4.8)$$

so that

$$\phi^C(x) = \exp(ieX^C + eg)\exp[if(t)]\,\phi(x) \qquad (4.9)$$

or, equivalently,

$$\phi^C(x) = \exp(ie\Lambda^C)\exp[if(t)]\,\phi(x)$$

with

$$\begin{aligned}\Lambda^C &= i(2\pi)^{-3/2}\int d\mathbf{k}\,[2|\mathbf{k}|^{3/2}]^{-1}[a_Q(\mathbf{k})\exp(i\mathbf{k}\cdot\mathbf{x})\\ &- a_R^\dagger(\mathbf{k})\exp(-i\mathbf{k}\cdot\mathbf{x})].\end{aligned} \qquad (4.10)$$

$f(t)$ can be formally represented by

$$f(t) = \lim_{\epsilon\to 0} ie(2\pi^2)^{-1}\int_0^\infty dk\int_{-\infty}^{t} dt'\exp(-2\epsilon|t'|) \qquad (4.11)$$

and is necessary to compensate for the fact that in

$$\begin{aligned}[H_C, \phi(\mathbf{x})] &= ie(8\pi)^{-1}\left[\int d\mathbf{y}\, j_4(\mathbf{y})|\mathbf{x}-\mathbf{y}|^{-1}\right]\phi(\mathbf{x})\\ &+ ie(8\pi)^{-1}\phi(\mathbf{x})\int d\mathbf{y}\, j_4(\mathbf{y})|\mathbf{x}-\mathbf{y}|^{-1}\end{aligned} \qquad (4.12)$$

the second term on the right-hand side makes a contribution which has $\phi(\mathbf{x})$ on the left of $A_4^C(\mathbf{x})$ instead of on the right.

The Coulomb gauge operators obtained by this transformation obey the commutation rules

$$\begin{aligned}&[A_0^C(\mathbf{x}), \phi^C(\mathbf{y})] = -e\phi^C(\mathbf{y})\,[4\pi|\mathbf{x}-\mathbf{y}|]^{-1},\\ &[\mathbf{A}_T^C(\mathbf{x}), \phi^C(\mathbf{y})] = 0\end{aligned} \qquad (4.13)$$

and the equations of motion

$$\Box \mathbf{A}_T^C = -\mathbf{j}_T, \quad \mathbf{A}_{\text{Long}}^C = 0 \tag{4.14}$$

and

$$\nabla^2 A_0^C = i j_4,$$

where

$$\Box = \nabla^2 + [\hat{H}, [\hat{H}, \ ]].$$

The identity of all dynamical consequences for the Coulomb and Lorentz gauge versions is trivial in this theory. In the transformation from the Lorentz to the Coulomb gauge neither $\hat{H}$ nor the definition of the physical states is in any way altered. Only the potentials are redefined; all other quantities remain unaltered.

V. THE QUANTUM ELECTRODYNAMICS OF CHARGED VECTOR BOSONS IN THE LORENTZ GAUGE

In this section we quote, without much detailed explanation, results for charged vector bosons analogous to those derived above for charged scalars.

The Lagrangian density, in the Feynman gauge, obtained by minimal coupling, is

$$\begin{aligned} \mathcal{L} = &-\tfrac{1}{4} F_{\mu\nu} F_{\mu\nu} - G(x)\, \partial_\mu A_\mu + \tfrac{1}{2} G^2(x) \\ &- \tfrac{1}{2} S_{\mu\nu}^\dagger S_{\mu\nu} - M^2 W_\mu^\dagger W_\mu, \end{aligned} \tag{5.1}$$

where

$$S_{\mu\nu} = D_\mu W_\nu - D_\nu W_\mu \tag{5.2a}$$

and

$$S_{\mu\nu}^\dagger = D_\mu^\dagger W_\nu^\dagger - D_\nu^\dagger W_\mu^\dagger. \tag{5.2b}$$

The equations of motion generated by this Lagrangian density are

$$\begin{aligned} &D_\mu S_{\mu\nu} - M^2 W_\nu = 0, \quad D_\mu^\dagger S_{\mu\nu}^\dagger - M^2 W_\nu^\dagger = 0, \\ &\Box A_\nu = -j_\nu, \quad \text{and} \quad G(x) = \partial_\mu A_\mu \end{aligned} \tag{5.3}$$

with

$$j_\nu = ie(W_\mu^\dagger S_{\mu\nu} - W_\mu S_{\mu\nu}^\dagger) \tag{5.4}$$

and

$$\partial_\mu j_\mu = 0. \tag{5.5}$$

Separating $\mathcal{L}$ into free and interacting parts, $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_I$, we find

$$\mathcal{L}_0 = -(1/4) F_{\mu\nu} F_{\mu\nu} - G \partial_\mu A_\mu + (1/2) s_{\mu\nu}^\dagger s_{\mu\nu} - M^2 W_\mu^\dagger W_\mu \tag{5.6a}$$

$$\begin{aligned} \mathcal{L}_I = (1/2)\{&ie s_{\mu\nu}^\dagger [A_\mu W_\nu - A_\nu W_\mu] - ie s_{\mu\nu} [A_\mu W_\nu^\dagger - A_\nu W_\mu^\dagger] \\ &- e^2 [A_\mu W_\nu^\dagger - A_\nu W_\mu^\dagger][A_\mu W_\nu - A_\nu W_\mu]\}, \end{aligned} \tag{5.6b}$$

where

$$s_{\mu\nu} = \partial_\mu W_\nu - \partial_\nu W_\mu \quad \text{and} \quad s_{\mu\nu}^\dagger = \partial_\mu W_\nu^\dagger - \partial_\nu W_\mu^\dagger. \tag{5.7}$$

Using our prescription, Eq. (2.5), that $\mathcal{L}_I = \int j_\mu[A]\, dA_\mu$, we find that

$$\mathcal{L}_I = A_\mu j_\mu + e^2 [A_\mu A_\mu W_\nu^\dagger W_\nu - A_\mu A_\nu W_\mu^\dagger W_\nu] \tag{5.8}$$

in agreement with Eq. (5.6b).

The momenta conjugate to the fields A_j, A_4, W_μ, $W_\mu^\dagger$ are

$$\begin{aligned} &\Pi_j = \partial_0 A_j - i\partial_j A_4, \quad \Pi_4 = iG, \\ &P_\mu = iS_{4\mu}^\dagger, \quad \text{and,} \quad P_\mu^\dagger = iS_{4\mu}. \end{aligned} \tag{5.9}$$

Since P_4 and $P_4^\dagger$ both vanish, W_4 and $W_4^\dagger$ are not independent dynamical degrees of freedom; rather, they are eliminated through the equations of motion for W_4 and $W_4^\dagger$:

$$W_4 = (i/M^2)\, \mathbf{D} \cdot \mathbf{P}^\dagger \quad \text{and} \quad W_4^\dagger = (i/M^2)\, \mathbf{D}^\dagger \cdot \mathbf{P}. \tag{5.10}$$

A representation of W_j and P_j which leads to the commutation relations

$$[W_j(\mathbf{x}), P_k(\mathbf{x}')] = [W_j^\dagger(\mathbf{x}), P_k^\dagger(\mathbf{x}')] = i\delta_{jk}\delta(\mathbf{x} - \mathbf{x}') \tag{5.11}$$

is

$$\begin{aligned} W_j(\mathbf{x}) = (2\pi)^{-3/2} \int d\mathbf{p}\, [2p_0]^{-1/2} \sum_{\lambda=1}^{3} \epsilon_j^\lambda(\mathbf{p}) [&b_\lambda(\mathbf{p}) \exp(i\mathbf{p}\cdot\mathbf{x}) \\ &+ d_\lambda^\dagger(\mathbf{p}) \exp(-i\mathbf{p}\cdot\mathbf{x})], \end{aligned} \tag{5.12a}$$

$$\begin{aligned} P_j(\mathbf{x}) = i(2\pi)^{-3/2} &\int d\mathbf{p}\, [2p_0]^{-1/2} \\ &\times \sum_{\lambda=1}^{3} [\epsilon_j^\lambda(\mathbf{p})\, p_0 - \epsilon_0^\lambda(\mathbf{p})\, p_j][b_\lambda^\dagger(\mathbf{p}) \exp(-i\mathbf{p}\cdot\mathbf{x}) \\ &- d_\lambda(\mathbf{p}) \exp(i\mathbf{p}\cdot\mathbf{x})], \end{aligned} \tag{5.12b}$$

where $p_0 = [|\mathbf{p}|^2 + M^2]^{1/2}$ and the polarization vectors $\epsilon_j^\lambda(\mathbf{p})$ satisfy the relation[12]

$$\sum_{\lambda=1}^{3} \epsilon_j^\lambda(\mathbf{p})\, \epsilon_k^\lambda(\mathbf{p}) = \delta_{jk} + p_j p_k / M^2. \tag{5.13}$$

$b_\lambda^\dagger$, b_λ, etc., are the creation and annihilation operators for the elementary excitations of the charged vector field of polarization λ.

The Hamiltonian density is

$$\begin{aligned} \mathcal{H} = &\tfrac{1}{4} F_{jk} F_{jk} + \tfrac{1}{2} \Pi_\mu^2 - i[A_4 \nabla \cdot \mathbf{\Pi} + \Pi_4 \nabla \cdot \mathbf{A}] + \mathbf{P}^\dagger \cdot \mathbf{P} \\ &+ M^2 \mathbf{W}^\dagger \cdot \mathbf{W} + [\mathbf{D}^\dagger \times \mathbf{W}^\dagger] \cdot [\mathbf{D} \times \mathbf{W}] - ieA_0 \\ &\times [\mathbf{P}^\dagger \cdot \mathbf{W}^\dagger - \mathbf{P} \cdot \mathbf{W}] + (1/M^2)[\mathbf{D} \cdot \mathbf{P}^\dagger][\mathbf{D}^\dagger \cdot \mathbf{P}], \end{aligned} \tag{5.14}$$

where, following standard practice,[13] we have added the total divergence

$$-i\partial_j [P_j W_4 + P_j^\dagger W_4^\dagger], \tag{5.15}$$

which vanishes when integrated over all space to give the total Hamiltonian.

The charge density is

$$\rho = ie[\mathbf{P}^\dagger \cdot \mathbf{W}^\dagger - \mathbf{P} \cdot \mathbf{W}], \tag{5.16}$$

and the current density is

$$j_k = ie[S_{kl}^\dagger W_l - S_{kl} W_l^\dagger] + (ie/M^2)[P_k^\dagger \mathbf{D}^\dagger \cdot \mathbf{P} - \mathbf{P}_k \mathbf{D} \cdot \mathbf{P}^\dagger]. \tag{5.17}$$

Once again, $\mathbf{j}$ contains the electromagnetic potentials explicitly, while ρ does not. Hence, the Hamiltonian contains seagull terms in A_j, but not in A_4. The commutator of ρ and $\mathbf{j}$ is given, following Eq. (2.12), by

$$\begin{aligned} [\rho(\mathbf{x}), j_l(\mathbf{x}')] = -ie^2 (\partial/\partial x_k) \{&[2\delta_{kl} \mathbf{W}^\dagger \cdot \mathbf{W} - W_l^\dagger W_k \\ &- W_k^\dagger W_l]\, \delta(\mathbf{x} - \mathbf{x}') + (1/M^2) \\ &\times [P_l^\dagger P_k + P_k^\dagger P_l]\, \delta(\mathbf{x} - \mathbf{x}')\}. \end{aligned} \tag{5.18}$$

The equations of motion and current conservation again demand that Π_4 be free:

$$\Box \Pi_4 = 0. \tag{5.19}$$

The invariant positive and negative frequency parts, calculated using the Cauchy solution, are

$$\Pi_4^{\pm}(x) = \mp(2\pi)^{-3/2} \int d\mathbf{k}\, |\mathbf{k}|^{1/2} \exp(\pm i\mathbf{k}\cdot\mathbf{x})\, \Omega^{\pm}(\mathbf{k}), \tag{5.20}$$

where

$$\Omega^{+}(\mathbf{k}) = a_Q(\mathbf{k}) + \rho(\mathbf{k})/[2|\mathbf{k}|^{3/2}] \tag{5.21a}$$

and

$$\Omega^{-}(\mathbf{k}) = a_R^{\dagger}(\mathbf{k}) + \rho(-\mathbf{k})/[2|\mathbf{k}|^{3/2}]. \tag{5.21b}$$

a_Q and $a_R^{\dagger}$ are the same as in the scalar case, but $\rho(\mathbf{k})$ is now the charge density, in momentum space, of charged vector bosons. The subsidiary condition and its adjoint, defining the physical subspace, are

$$\Omega^{+}(\mathbf{k})\,|\nu\rangle = 0 \tag{5.22a}$$

and

$$\langle \nu^{\star}|\,\Omega^{-}(\mathbf{k}) = 0. \tag{5.22b}$$

As before, we perform a rotation within the indefinite metric space so that the hatted form of the subsidiary condition is

$$a_Q(\mathbf{k})\,|n\rangle = 0 \tag{5.23}$$

and the hatted Hamiltonian is

$$\hat{H} = e^{D} H e^{-D} = H + [D, H] + \tfrac{1}{2}[D,[D,H]]. \tag{5.24}$$

$[D,[D,[D,H]]]$ and all higher commutator terms vanish. $\hat{H}$ is given by

$$\hat{H} = H_0 + H_C + H_{I,T} + H_Q \tag{5.25}$$

where

$$H_0 = \int d\mathbf{k}|\mathbf{k}| \left\{ \sum_{i=1}^{2} a_i^{\dagger}(\mathbf{k}) a_i(\mathbf{k}) + a_Q^{\dagger}(\mathbf{k})\, a_Q(\mathbf{k}) + a_R^{\dagger}(\mathbf{k})\, a_R(\mathbf{k}) \right\} + \int d\mathbf{p}\, p_0 \sum_{\lambda=1}^{3} \{ b_\lambda^{\dagger}(\mathbf{p})\, b_\lambda(\mathbf{p}) + d_\lambda^{\dagger}(\mathbf{p})\, d_\lambda(\mathbf{p}) \}, \tag{5.26a}$$

$$H_C = (8\pi)^{-1} \int d\mathbf{x}\, d\mathbf{y}\, \rho(\mathbf{x})\, \rho(\mathbf{y})\, |\mathbf{x}-\mathbf{y}|^{-1}, \tag{5.26b}$$

$$H_{I,T} = -\int d\mathbf{x}\, A_k^T(\mathbf{x}) [ie j_k^M(\mathbf{x}) - e^2 A_k^T(\mathbf{x})\, \alpha(\mathbf{x}) + (e^2/2) A_j^T(\mathbf{x})\, Y_{kl}(\mathbf{x})], \tag{5.26c}$$

$$\begin{aligned} H_Q = {} & -\tfrac{1}{2} \int d\mathbf{k} |\mathbf{k}|^{-1/2} [\rho(-\mathbf{k})\, a_Q(\mathbf{k}) + \rho(\mathbf{k})\, a_R^{\dagger}(\mathbf{k})] - ie \int d\mathbf{k} [2|\mathbf{k}|^{3/2}]^{-1} \mathbf{k}\cdot[\mathbf{j}^M(-\mathbf{k})\, a_Q(\mathbf{k}) + \mathbf{j}^M(\mathbf{k})\, a_R^{\dagger}(\mathbf{k})] \\ & + e^2 (2\pi)^{-3/2} \int d\mathbf{k}\, d\mathbf{k}'\, d\mathbf{k}''\, (16|\mathbf{k}|\,|\mathbf{k}'|)^{-1/2}\, \hat{\mathbf{k}}\cdot\hat{\mathbf{k}}'\, \alpha(\mathbf{k}'') \{ a_Q(\mathbf{k})\, a_Q(\mathbf{k}')\, \delta(\mathbf{k}+\mathbf{k}'+\mathbf{k}'') + a_R^{\dagger}(\mathbf{k})\, a_R^{\dagger}(\mathbf{k}'')\, \delta(\mathbf{k}+\mathbf{k}'-\mathbf{k}'') \\ & + 2 a_R^{\dagger}(\mathbf{k})\, a_Q(\mathbf{k}')\, \delta(\mathbf{k}-\mathbf{k}'-\mathbf{k}'') \} + e^2 (2\pi)^{-3/2} \int d\mathbf{k}\, d\mathbf{k}'\, d\mathbf{k}''\, [2|\mathbf{k}|\,|\mathbf{k}'|]^{-1/2} \sum_{i=1}^{2} \hat{e}^i(\mathbf{k})\cdot\hat{\mathbf{k}}'\, \alpha(\mathbf{k}'') \\ & \times \{ a_i(\mathbf{k})\, a_Q(\mathbf{k}')\, \delta(\mathbf{k}+\mathbf{k}'+\mathbf{k}'') + a_i^{\dagger}(\mathbf{k})\, a_R^{\dagger}(\mathbf{k}')\, \delta(\mathbf{k}+\mathbf{k}'-\mathbf{k}'') + a_i^{\dagger}(\mathbf{k})\, a_Q(\mathbf{k}')\, \delta(\mathbf{k}-\mathbf{k}'-\mathbf{k}'') + a_i(\mathbf{k})\, a_R^{\dagger}(\mathbf{k}')\, \delta(\mathbf{k}-\mathbf{k}'+\mathbf{k}'') \} \\ & - e^2 (2\pi)^{-3/2} \int d\mathbf{k}\, d\mathbf{k}'\, d\mathbf{k}''\, [64|\mathbf{k}|\,|\mathbf{k}'|]^{-1/2}\, \hat{\mathbf{k}}\cdot Y(\mathbf{k}'')\cdot\hat{\mathbf{k}}' \{ a_Q(\mathbf{k})\, a_Q(\mathbf{k}')\, \delta(\mathbf{k}+\mathbf{k}'+\mathbf{k}'') + a_R^{\dagger}(\mathbf{k})\, a_R^{\dagger}(\mathbf{k}')\, \delta(\mathbf{k}+\mathbf{k}'-\mathbf{k}'') \\ & + 2 a_R^{\dagger}(\mathbf{k})\, a_Q(\mathbf{k}')\, \delta(\mathbf{k}-\mathbf{k}'-\mathbf{k}'') \} - e^2 (2\pi)^{-3/2} \int d\mathbf{k}\, d\mathbf{k}'\, d\mathbf{k}''\, [8|\mathbf{k}|\,|\mathbf{k}'|]^{-1/2} \sum_{i=1}^{2} \hat{e}^i(\mathbf{k})\cdot Y(\mathbf{k}'')\cdot\hat{\mathbf{k}}' \\ & \times \{ a_i(\mathbf{k})\, a_Q(\mathbf{k}')\, \delta(\mathbf{k}+\mathbf{k}'+\mathbf{k}'') + a_i^{\dagger}(\mathbf{k})\, a_R^{\dagger}(\mathbf{k}')\, \delta(\mathbf{k}+\mathbf{k}'-\mathbf{k}'') + a_i^{\dagger}(\mathbf{k})\, a_Q(\mathbf{k}')\, \delta(\mathbf{k}-\mathbf{k}'-\mathbf{k}'') + a_i(\mathbf{k})\, a_R^{\dagger}(\mathbf{k}')\, \delta(\mathbf{k}-\mathbf{k}'-\mathbf{k}'') \}. \end{aligned} \tag{5.26d}$$

$A_j^T(\mathbf{x})$ is the transverse part of the electromagnetic potential. α, $\mathbf{j}^M$, and $\overleftrightarrow{Y}$ are the parts of the current density which do not depend on A_μ, and are given by

$$\alpha = \mathbf{W}^{\dagger}\cdot\mathbf{W}, \tag{5.27a}$$

$$j_l^M = [S_{lk}^{\dagger} W_k - S_{lk} W_k^{\dagger}] + (1/M^2)[\Pi_l^{\dagger} \nabla\cdot\Pi - \Pi_l \nabla\cdot\Pi^{\dagger}], \tag{5.27b}$$

$$Y_{kl} = W_k^{\dagger} W_l + W_l^{\dagger} W_k - (1/M^2)[\Pi_k^{\dagger}\Pi_l + \Pi_l^{\dagger}\Pi_k]. \tag{5.27c}$$

They are related to $\mathbf{j}$ by

$$\mathbf{j} = ie\mathbf{j}^M - 2e^2\alpha\mathbf{A} + e^2\mathbf{A}\cdot\overleftrightarrow{Y}. \tag{5.28}$$

As in the case of charged scalars, H_Q may be completely amputated from $\hat{H}$ without in any way affecting the calculation of dynamical processes. An operator gauge transformation can be effected upon the potentials, with the vector field transformed to

$$W_j^C(\mathbf{x}) = \exp(ie\Lambda^C) \exp[if(t)]\, W_j(\mathbf{x}), \tag{5.29}$$

and with the Coulomb gauge commutation relations

$$\begin{aligned} & [A_0^C(\mathbf{x}),\, W_j^C(\mathbf{y})] = -e W_j^C(\mathbf{y}) [4\pi |\mathbf{x}-\mathbf{y}|]^{-1}, \\ & [\mathbf{A}_T^C(\mathbf{x}),\, W_j^C(\mathbf{y})] = 0, \quad \text{and} \quad \mathbf{A}_{\text{Long}}^C = 0. \end{aligned} \tag{5.30}$$

The relationship between the transition amplitude in the ISC formulation of Lorentz gauge qed, in which the subsidiary condition is exact, and the NISC formulation, in which the photon-charged particle interaction is ignored in formulating the subsidiary condition, is identical in boson and spinor qed. That relationship is given by[14]

$$\begin{aligned} \overline{T}_{f,i}^{\epsilon} = {} & T_{f,i}^{\epsilon} + \langle \varphi_f^{\star} | (1 - e^{-D}) | \hat{\Psi}_i^{\epsilon} \rangle (E_i - E_f) \\ & + i\epsilon \langle \varphi_f^{\star} | (1 - e^{-D}) [E_i - \hat{H} + i\epsilon]^{-1} \hat{H}_I \\ & - H_I [E_i - H + i\epsilon]^{-1} (1 - e^{-D}) | \varphi_i \rangle, \end{aligned} \tag{5.31}$$

where $\overline{T}_{f,i}^{\epsilon}$ and $T_{f,i}^{\epsilon}$ are the transition amplitudes in the ISC and NISC respectively, where $|\varphi_i\rangle$ and $|\varphi_f\rangle$ are eigenstates of H_0 and $\hat{\Psi}_i^{\epsilon}$ is the outgoing eigenstate of $\hat{H}$.

*Present address: Department of Physics and Astronomy, University of Maryland, College Park, Maryland 20742.

[1] A complete review of the subject, with references to earlier work, is given in K. Haller, Acta Physica Austriaca **42**, 163 (1975).

[2]S. P. Tomczak and K. Haller, Nuovo Cimento B **8**, 1 (1972).
[3]K. Haller and R. Sohn, Phys. Rev. D **14**, 479 (1976).
[4]Reference 1.
[5]D. Lurie, *Particles and Fields* (Interscience, New York, 1968), Chap. 5.
[6]J. Schwinger, Phys. Rev. **130**, 406 (1963); L. Brown, Phys. Rev. **150**, 1338 (1966).
[7]S. N. Gupta, Proc. Phys. Soc. A **63**, 681 (1950); K. Bleuler, Helv. Phys. Acta **23**, 567 (1950).
[8]Reference 1.
[9]Reference 2.
[10]References 1 and 2.
[11]Many of the results quoted here are the same as those for spinor qed; see K. Haller, Nucl. Phys. B **57**, 589 (1973).
[12]S. Gasiorowicz, *Elementary Particle Physics* (Wiley, New York, 1967), see Chap. 3.
[13]G. Wentzel, *Quantum Theory of Fields* (Interscience, New York, 1949).
[14]K. Haller and L. Landovitz, Phys. Rev. D **2**, 1498 (1970); also, Refs. 1 and 2.

A simple derivation of a closed formula for Bogoliubov boson transformations

A. A. Raduta*

Institute for Theoretical Physics, University of Frankfurt/M., W. Germany

M. Badea and E. Badralexe

Institute for Atomic Physics, Bucharest, P.O.B. 5026, Romania
(Received 30 April 1976)

Considering the states with an arbitrary number of bosons and their transformed states under Bogoliubov transformations as wavefunctions of an oscillator type, a very simple derivation for the matrix elements of the Bogoliubov transformations is given.

1. INTRODUCTION

It is well known that the Bogoliubov (B) transformation is very useful in treating many body systems of fermions as well as of bosons. Under such a transformation the system of physical particles goes over into a system of quasiparticles which are "dressed" with a part of the two-body interaction.[1] In the new picture some phenomena like superconductivity (if we deal with fermions) and superfluidity (in the case of bosons) can be easily explained. Also, in the case of a system of fermions in certain circumstances a pair of fermions is approximated as a quasiboson, and subsequently by means of a transformation of a B type one defines a new type of excitation named "phonons." This approach is usually called the random phase approximation (RPA).[2]

For such applications it is useful to know how to express the transformed states in terms of the initial states.

Recently, there appeared several papers[3-5] which treat this problem by different methods. Thus in Ref. 3 the very lengthy method of "series expansion" is used. In Ref. 4, the recursion formulas of the coefficients of the B transformations constitute defining equations which are solved by the generating function technique. In Ref. 5 one employs the fact that the operators a^{+2}, a^2, $2a^+a+1$ form a closed algebra.

The solution we give here is particularly simple and is based on the fact that the boson states are related to the motion of a harmonic oscillator.

In Sec. 2 we treat the case of zero-angular momentum bosons. Section 3 deals with the nonzero momentum bosons.

2. THE TREATMENT OF ZERO-ANGULAR MOMENTUM BOSONS

Let us denote by a^+, a the creation and the annihilation operators of one particle of a boson type:

$$[a, a^+]=1. \tag{2.1}$$

The Fock space for the given system of bosons is defined as follows:

$$|m\rangle=(1/\sqrt{m!})(a^+)^m|0\rangle, \tag{2.2}$$

where $|0\rangle$ is the vacuum state and satisfies the equation:

$$a|0\rangle=0. \tag{2.3}$$

Under the B transformation

$$T=e^S, \quad S=\tfrac{1}{2}y(a^{+2}-a^2), \quad y \text{ real}, \tag{2.4}$$

the operators a^+, a change linearly

$$\begin{aligned} Ta^+T^+ &= Ua^+ + Va \equiv b^+, \\ TaT^+ &= Ua + Va^+ \equiv b, \end{aligned} \tag{2.5}$$

where $U=\cosh y$, $V=\sinh y$. One easily verifies that b^+, b are boson operators:

$$[b, b^+]=1. \tag{2.6}$$

The B transformation changes the state $|m\rangle$ into

$$|\tilde{m}\rangle=T|m\rangle=(1/\sqrt{m!})(b^+)^m|\tilde{0}\rangle, \quad b|\tilde{0}\rangle=0. \tag{2.7}$$

Here, we are interested in evaluating the overlap:

$$G_{n,m}(y)=\langle n|\tilde{m}\rangle. \tag{2.8}$$

To this end we define first the Hamiltonian for a one-dimensional harmonic oscillator.

$$H=p^2/2M+Kx^2. \tag{2.9}$$

The variables (x,p) are related to the boson operators a^+, a in the following way:

$$x=\alpha(a^+ + a), \quad p=(i/2\alpha)(a^+ - a) \tag{2.10}$$

with $\alpha=(8KM)^{-1/4}$.

In terms of a^+, a operators, H has the form

$$H=\omega(a^+a+\tfrac{1}{2}), \quad \omega=(2K/M)^{1/2}. \tag{2.11}$$

Now, it becomes clear that the state $|n\rangle$ is an eigenstate of H corresponding to the eigenvalue $\omega(n+\frac{1}{2})$. Considering this remark in connection with Eq. (2.9), we have

$$\begin{aligned} |n\rangle &= N_n(\alpha)\exp(-x^2/4\alpha^2)H_n(x/\sqrt{2}\alpha), \\ N_n(\alpha) &= (2^n n!\alpha\sqrt{2\pi})^{-1/2}. \end{aligned} \tag{2.12}$$

Here, H_n stands for the Hermite polynomial. The image of H [(2.11)] through the B transformation is

$$H' \equiv THT^+ = \omega[(U^2+V^2)a^+a + UV(a^{+2}+a^2) + (V^2+\tfrac{1}{2})]. \tag{2.13}$$

or, alternatively,

$$H'=\omega(b^+b+\tfrac{1}{2}). \tag{2.14}$$

Making use of the transformation (2.10), we can express H' in terms of (x,p) variables:

Copyright © 1977 American Institute of Physics

$$H' = \omega\alpha^2(U-V)^2p^2 + [\omega(U+V)^2/4\alpha^2]x^2. \tag{2.15}$$

This Hamiltonian describes the motion of a one-dimensional oscillator having the characteristics

$$M_1 = [2\omega\alpha^2(U-V)^2]^{-1}, \quad K_1 = \omega(U+V)^2/4\alpha^2, \quad \alpha_1 = e^{-\gamma}\alpha, \quad \omega_1 = \omega. \tag{2.16}$$

From Eq. (2.14) it results that $|\tilde{m}\rangle$ is an eigenstate of H' corresponding to the eigenvalue $\omega(m+\frac{1}{2})$. Then, from Eq. (2.15) follows

$$|\tilde{m}\rangle = N_m(\alpha_1)\exp(-x^2/4\alpha_1^2)H_m(x/\sqrt{2}\alpha_1). \tag{2.17}$$

Now, taking into account the relations (2.12) and (2.17), one obtains

$$G_{n,m}(x) = N_n(\alpha)N_m(\alpha_1)\alpha\sqrt{2}I_{nm}, \tag{2.18}$$

where we have denoted

$$I_{nm} = \int_{-\infty}^{\infty}\exp[-z^2(\beta^2+1)/2]H_n(z)H_m(\beta z)\,dz, \quad \beta = e^{\gamma}. \tag{2.19}$$

In order to perform the integral I_{nm} we shall use the generating function technique. Thus, the Hermite polynomials $H_n(z)$, $H_m(\beta z)$ are the coefficients of the following expansions:

$$\exp(-t^2+2tz) = \sum_{n=0}^{\infty}\frac{t^n}{n!}H_n(z),$$
$$\exp(-s^2+2s\beta z) = \sum_{m=0}^{\infty}\frac{s^m}{m!}H_m(\beta z). \tag{2.20}$$

Multiplying the relations (2.20) and then integrating the results with the weight $\exp[-z^2(\beta^2+1)/2]$, we find

$$(\pi/\beta U)^{1/2}\exp\left(-\frac{V}{U}t^2 - \frac{V}{U}s^2 + \frac{2}{U}ts\right) = \sum\frac{t^n s^m}{n!\,m!}I_{nm}. \tag{2.21}$$

Expanding now the left-hand side in terms of t and s and then identifying the coefficients of the two series one easily finds the value of I_{nm}. Inserting this into (2.18), one obtains

$$G_{n,m}(y) = \sqrt{m!\,n!}\,(\cosh y)^{-[(m+n+1)/2]} \times \sum_q \frac{(-1)^{(n-q)/2}}{q![(n-q)/2]![(m-q)/2]!}\left(\tfrac{1}{2}\sinh y\right)^{(m+n)/2-q}. \tag{2.22}$$

The summation in the above expression is restricted by the following conditions:

$$n-q = \text{even}, \quad m-q = \text{even}, \quad q < \min(m,n). \tag{2.23}$$

3. THE CASE OF NONZERO MOMENTUM BOSONS

Now, let us consider the system of boson operators $\{a_\beta^+\}_{\beta=\pm k}$ which create a particle in a state of momentum $\beta \neq 0$. We shall denote by a_β the corresponding annihilation operators.

These operators satisfy the following commutation relations:

$$[a_\beta, a_{\beta'}^+] = \delta_{\beta\beta'}, \quad [a_\beta, a_{\beta'}] = [a_\beta^+, a_{\beta'}^+] = 0, \quad \beta,\beta' = \pm k. \tag{3.1}$$

The Fock space will be generated as follows:

$$|m,n\rangle = [(a_k^+)^m(a_{-k}^+)^n/\sqrt{m!\,n!}]|0,0\rangle,$$
$$a_\beta|0,0\rangle = 0, \quad \beta = \pm k. \tag{3.2}$$

Conventionally we shall define the conjugate states by

$$\langle m,n| = (|n,m\rangle)^*. \tag{3.3}$$

The B transformation is defined by

$$T = e^S, \quad S = -\gamma(a_k^+a_{-k}^+ - a_ka_{-k}), \quad k \neq 0. \tag{3.4}$$

Under this transformation the operators a_β^+, a_β change as follows:

$$Ta_\beta^+T^+ = Ua_\beta^+ + Va_{-\beta} \equiv b_\beta^+,$$
$$Ta_\beta T^+ = Ua_\beta + Va_{-\beta}^+ \equiv b_\beta, \quad \beta = \pm k. \tag{3.5}$$

with $U = \cosh\gamma$, $V = \sinh\gamma$.

Consequently, the image of the states (3.2) through the transformation (3.4) is

$$\widetilde{|m,n\rangle} = T|m,n\rangle = [(b_k^+)^m(b_{-k}^+)^n/\sqrt{m!\,n!}]\widetilde{|0,0\rangle},$$
$$b_\beta\widetilde{|0,0\rangle} = 0, \quad \beta = \pm k. \tag{3.6}$$

In what follows we shall be interested in evaluating the scalar product

$$G_{pl,mn}(\gamma) = \langle p,l|\widetilde{m,n}\rangle. \tag{3.7}$$

For this purpose we shall introduce the Hamiltonian

$$H = [(p_k^2+p_{-k}^2)/2M] + \eta(x_k^2+x_{-k}^2), \tag{3.8}$$

where x_β, p_β ($\beta = \pm k$) are conjugate variables and are related with the boson operators a_β^+, a_β by the equations

$$x_\beta = \alpha(a_\beta^+ + a_\beta),$$
$$p_\beta = (i/2\alpha)(a_\beta^+ - a_\beta), \quad \beta = \pm k, \quad \alpha = (8M\eta)^{-1/4} \tag{3.9}$$

It is convenient to express H in terms of a_β^+, a_β operators. Thus, one obtains

$$H = \omega(a_k^+a_k + a_{-k} + 1), \quad \omega = (2\eta/M)^{1/2}. \tag{3.10}$$

Now, using the equality

$$H|l,p\rangle = \omega(p+l+1)|l,p\rangle \tag{3.11}$$

and taking into account the form (3.8) for H, one obtains

$$|l,p\rangle = N_l(\alpha)N_p(\alpha)\exp[-(1/4\alpha^2)(x_k^2+x_{-k}^2)] \times H_l(x_k/\sqrt{2}\alpha)H_p(x_{-k}/\sqrt{2}\alpha). \tag{3.12}$$

Here H_l, H_p are Hermite polynomials, while N_l, N_p are the corresponding normalization factors.

The transformed Hamiltonian under the B transformation is

$$H' \equiv THT^+ = \omega(U^2+V^2)(a_k^+a_k + a_{-k}^+a_{-k}) + 2\omega UV(a_k^+a_{-k}^+ + a_{-k}a_k) + \omega(2V^2+1), \tag{3.13}$$

or, alternatively,

$$H' = \omega(b_k^+b_k + b_{-k}^+b_{-k} + 1). \tag{3.14}$$

In what follows it is useful to introduce the operators

$$X_\beta = Ux_\beta + Vx_{-\beta},$$
$$P_\beta = Up_\beta - Vp_{-\beta}, \quad \beta = \pm k. \tag{3.15}$$

The Hamiltonian (3.13) can be easily expressed in terms of the new variables:

$$\mathcal{H}' = \omega\alpha^2(P_k^2 + P_{-k}^2) + (\omega/4\alpha^2)(X_k^2 + X_{-k}^2) - 3\omega(U^2 + V^2). \tag{3.16}$$

From the defining Equations (3.15) results

$$[X_\beta, P_{\beta'}] = i\delta_{\beta\beta'}. \tag{3.17}$$

Thus, we can interpret (X_κ, P_κ) and $(X_{-\kappa}, P_{-\kappa})$ as the coordinates and the conjugate momenta of two independent oscillators.

From Eq. (3.14), it is evident that

$$\mathcal{H}'|\widetilde{m,n}\rangle = \omega(m+n+1)|\widetilde{m,n}\rangle. \tag{3.18}$$

Then, from the relation (3.16) it follows that

$$\langle\widetilde{m,n}\rangle = N_m(\alpha)N_n(\alpha)\exp[-(1/4\alpha^2)(X_k^2 + X_{-k}^2)] \times H_m(X_k/\sqrt{2}\alpha)H_n(X_{-k}/\sqrt{2}\alpha). \tag{3.19}$$

Inserting now the wavefunctions (3.12) and (3.19) into Eq. (3.7), one obtains

$$G_{pl\,mn} = N_p(\alpha)N_l(\alpha)N_m(\alpha)N_n(\alpha)2\alpha^2 I_{pl,\,mn}, \tag{3.20}$$

where we denoted

$$I_{pl,\,mn} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp[-(U^2y^2 + U^2z^2 + 2UVz)] \times H_p(z)H_l(y)H_m(Uy + Vz)H_n(Uz + Vy)\,dy\,dz. \tag{3.21}$$

In order to evaluate the integrals (3.21), we use the procedure we described in the previous sections. Thus, one obtains

$$(\pi/U)\exp[(2/U)(\lambda h - V\lambda t + tw + Vwh)] = \sum(\lambda^p t^l w^m h^n/p!\,l!\,m!\,n!)I_{pl,\,mn}. \tag{3.22}$$

Expanding the exponential function in terms of the variables λ, t, w, h had then identifying the series from the two sides, one obtains the coefficients $I_{pl,\,mn}$. The identification is possible only if the following compatibility condition is fulfilled:

$$m - l = n - p.$$

Inserting now the resulting coefficients $I_{pl,\,mn}$ in the relation (3.20) one obtains the final expression for $G_{pl,\,mn}$:

$$G_{pl,\,mn} = \sqrt{p!\,l!\,m!\,n!}(\cosh y)^{-(l+n+1)} \times \sum_q \frac{(-)^{l-m+q}(\sinh y)^{l-m+2q}}{q!\,(n-q)!\,(m-q)!\,(l-m+q)!}. \tag{3.23}$$

Here, the summation over q is restricted by the conditions

$$q \leq \min(m, n), \quad q \geq m - l. \tag{3.24}$$

ACKNOWLEDGMENT

One of us (A.A.R.) would like to thank Professor R. Dreizler for his high hospitality at Institut für Theoretische Physik, Frankfurt.

*Fellow of Humboldt-Stiftung. Permanent address: Institute for Atomic Physics, Bucharest, P.O.B. 5026, Romania.

[1]N.N. Bogoliubov, J. Phys. USSR **11**, 23 (1947); Zh. Eksp. Teor. Fiz. **34**, 58 (1958) [Sov. Phys. JETP **7**, 41 (1958)]; Nuovo Cimento **7**, 794 (1958); J.G. Valatin, Nuovo Cimento **7**, 843 (1958).
[2]See, for instance, G.E. Brown, *Unified Theory of Nuclear Models and Forces* (North-Holland, Amsterdam, 1967).
[3]K. Tanabe, J. Math. Phys. **14**, 618 (1973).
[4]M.A. Rashid, J. Math. Phys. **16**, 378 (1975).
[5]A. Keleman, Z. Physik A **274**, 109 (1975).

Moment-theory approximations for nonnegative spectral densities

C. T. Corcoran*

Department of Chemistry, Indiana University, Bloomington, Indiana 47401

P. W. Langhoff†

Department of Chemistry, Indiana University, Bloomington, Indiana 47401
and Joint Institute for Laboratory Astrophysics, National Bureau of Standards and University of Colorado, Boulder, Colorado 80309
(Received 7 September 1976)

Moment-theory approximations constructed from finite numbers of spectral power moments are described for continuous, nonnegative spectral densities and associated Stieltjes integrals. Derivatives of the mean (Stieltjes) values of the nth-order Tchebycheff bounds on nondecreasing distributions provide the appropriate approximations to the associated spectral densities. The nth-order Tchebycheff density so defined is shown to be real, nonnegative, and continuous on the real axis, to have $2n-4$ continuous derivatives there, and to support $2n-2$ positive-integer power moments. Related approximations to the associated Stieltjes integral are obtained from corresponding principal-value quadratures. The Tchebycheff densities are convergent in the limit of large numbers of spectral moments for determined moment problems, but they are not solutions of reduced moment problems of appropriate finite order. An illustrative application in the case of normal-mode lattice vibrations in a diatomic chain indicates that the Tchebycheff densities are suitably convergent, and provide faithful images of the forbidden band gap and Van Hove singularities present.

1. INTRODUCTION

Nonnegative spectral densities and associated Stieltjes integrals[1,2] arise in a great many connections in mathematical physics,[3] including studies of the absorption and dispersion of electromagnetic radiation in matter,[4,5] equilibrium statistical mechanics,[6,7] crystal lattice vibrations,[8–10] and certain formulations of particle scattering,[11–13] to mention some representative cases. When an infinite number of power moments of the spectral density exist and the moment problem is determined,[2] the classical theory of moments provides convergent bounds on the associated Stieltjes integral at any point in the complex plane off the real axis.[2,14,15] The bounds fail on the real axis in any finite order, however, and, consequently, numerical approximations to the spectral density and its associated principal-value integral are not obtained from the Stieltjes–Tchebycheff–Markoff moment theory. Since both densities and associated Stieltjes integrals can correspond to physical observables, it would be highly desirable to devise appropriate convergent approximations to their values from finite sequences of power moments.[16]

In the present paper, the celebrated Tchebycheff inequalities[2] for determined moment problems are employed in the construction of approximations to a continuous nonnegative spectral density from a given sequence of its power moments.[17] Derivatives of the mean values of the Tchebycheff bounds on the spectral distribution define the appropriate approximations to the associated continuous spectral density. The nth-order Tchebycheff density so constructed is shown to be real, nonnegative, and continuous on the real axis, to have $2n-4$ continuous derivatives there, and to support $2n-2$ positive-integrer power moments. Approximations to the associated Stieltjes integral are obtained from corresponding principal-value quadratures. The Tchebycheff densities are convergent in the limit of large numbers of spectral moments for determined moment problems, but they are not solutions of reduced moment problems of appropriate finite order.

Explicit expressions are obtained for the nth-order Tchebycheff density in terms of appropriate orthogonal polynomials, facilitating the investigation of its analytic properties and its numerical evaluation. Although detailed computational results are presented elsewhere,[18,19] an illustrative application in the case of lattice vibrations in a diatomic chain is given here. The Tchebycheff densities are found to be suitably convergent in this case, and to provide faithful images of the forbidden band gap and van Hove singularities present.

Spectral densities and distributions, and related Stieltjes integrals, are defined in Sec. 2.[1,2] The Stieltjes distributions and densities discussed by Deltour[17] are briefly reviewed in Sec. 3 as a preliminary to the Tchebycheff approach presented in Sec. 4. In Sec. 5, the properties of the Tchebycheff density are established, and the illustrative application is given in Sec. 6. Some concluding remarks are made in Sec. 7.

2. DENSITIES, DISTRIBUTIONS, AND STIELTJES INTEGRALS

The mean value of the resolvent $(z-H)^{-1}$ of a self-adjoint (Hamiltonian) operator H with respect to some test function Φ can be written in the form of a Stieltjes integral[20]

$$R_\Phi(z) = \langle \Phi | (z-H)^{-1} | \Phi \rangle = \int_0^\infty \frac{d\Gamma(\epsilon)}{z-\epsilon} = \int_0^\infty \frac{g(\epsilon)\,d\epsilon}{z-\epsilon}, \tag{1}$$

 Copyright © 1977 American Institute of Physics

where

$$g(\epsilon) = |\langle \Phi | \phi_\epsilon \rangle|^2 \geq 0 \tag{2}$$

is the spectral density function, and $\Gamma(\epsilon)$ is the associated distribution. In Eqs. (1) and (2) we assume that the spectrum of the Hamiltonian H is continuous,

$$(H - \epsilon)\phi_\epsilon = 0, \quad 0 \leq \epsilon < \infty, \tag{3}$$

although the presence of bound states offers no essential difficulty, adding a (possibly infinite) denumerable sequence of step functions having appropriate weights to the distribution $\Gamma(\epsilon)$ and associated delta functions to $g(\epsilon)$.

When spectral moments of the form

$$\mu_k = \langle \Phi | H^k | \Phi \rangle = \int_0^\infty \epsilon^k \, d\Gamma(\epsilon), \quad k \geq 0, \tag{4}$$

exist and the moment problem is determined, well-known theorems from the theory of moments provide bounds on $R_\Phi(z)$ at any point in the complex z plane off the real axis.[2] The bounds fail, however, on the real axis in any finite order. Consequently, approximations to the spectral density,

$$\begin{aligned} g(\epsilon) &= (1/2\pi i)[R_\Phi(\epsilon - i0) - R_\Phi(\epsilon + i0)] \\ &= \pm (1/\pi) \operatorname{Im} R_\Phi(\epsilon \mp i0), \end{aligned} \tag{5}$$

and to the related principal-value integral

$$\operatorname{Re} R_\Phi(\epsilon \pm i0) = P \int_0^\infty \frac{g(\omega)\, d\omega}{\epsilon - \omega}, \tag{6}$$

are not obtained from the classical theory of moments employing a finite sequence of spectral power moments [Eq. (4)]. Since both the density and the integral of Eqs. (5) and (6), respectively, or their ratio, can correspond to physical observables,[3-13] approximations to $R_\Phi(\epsilon \pm i0)$ constructed from spectral power moments are useful in various connections.

3. STIELTJES DISTRIBUTIONS AND DENSITIES

The spectral moments of Eq. (4) provide an asymptotic expansion of Eq. (1) in the form (S & T, p. 23)[21]

$$R_\Phi(z) \sim \frac{\mu_0}{z} + \frac{\mu_1}{z^2} + \frac{\mu_2}{z^3} + \cdots. \tag{7}$$

A convergent summation of the series of Eq. (7) in the case of a determined moment problem is obtained in terms of the continued fraction (S & T, p. 31)

$$A(z) = \cfrac{\beta_0}{z - \alpha_1 - \cfrac{\beta_1}{z - \alpha_2 - \cdots}}, \tag{8}$$

which can be regarded as the $(n - \infty)$ limit of its rational approximants (S & T, p. 33)

$$\begin{aligned} A^{(n)}(z) &= \cfrac{\beta_0}{z - \alpha_1 - \cfrac{\beta_1}{z - \alpha_2 - \cdots \cfrac{\beta_{n-1}}{z - \alpha_n}}} \\ &= p_{n-1}(z)/q_n(z). \end{aligned} \tag{9}$$

The coefficients $\beta_0, \beta_1, \cdots, \alpha_1, \alpha_2, \cdots$ of Eq. (9) are obtained from the moments μ_k by equating Eqs. (7) and (9) term by term in $(1/z)$ to order $2n$, resulting in well-known computational algorithms.[22] The polynomials $p_{n-1}(z)$, $q_n(z)$ are then conveniently obtained from the recurrence relations (S & T, p. 47)

$$p_{n-1}(z) = (z - \alpha_n)p_{n-2}(z) - \beta_{n-1}p_{n-3}(z),$$
$$p_{-1}(z) = 0, \quad p_0(z) = \beta_0, \tag{10a}$$
$$q_n(z) = (z - \alpha_n)q_{n-1}(z) - \beta_{n-1}q_{n-2}(z),$$
$$q_{-1}(z) = 0, \quad q_0(z) = 1. \tag{10b}$$

Here, the $q_n(z)$ are the nth-degree polynomials orthogonal with respect to the distribution $\Gamma(\epsilon)$,

$$\begin{aligned} \int_0^\infty q_n(\epsilon) q_m(\epsilon)\, d\Gamma(\epsilon) &= \beta_0\beta_1 \cdots \beta_n \delta_{nm} \\ &\equiv N_n \delta_{nm}, \end{aligned} \tag{11}$$

the $p_{n-1}(z)$ are associated numerator polynomials, and $A^{(n)}(z)$ is equivalent to the $[n, n-1](z)$ Padé approximant.[23] The latter can be written in the alternative form

$$A^{(n)}(z) = \sum_{i=1}^{n} \frac{f_i^{(n)}}{z - \epsilon_i^{(n)}}, \tag{12}$$

where the $\epsilon_i^{(n)}$ are real, distinct, and positive and satisfy

$$q_n(\epsilon_i^{(n)}) = 0, \tag{13a}$$

and the residues, or Christoffel numbers, $f_i^{(n)}$ are positive and are given by (S & T, pp. 39–41)

$$\begin{aligned} f_i^{(n)} &= p_{n-1}(\epsilon_i^{(n)})/q_n'(\epsilon_i^{(n)}) \\ &= \left[\sum_{k=0}^{n-1} N_k^{-1} q_k^2(\epsilon_i^{(n)})\right]^{-1}. \end{aligned} \tag{13b}$$

Note that the roots $\epsilon_i^{(n)}$ and residues $f_i^{(n)}$ of the approximant of Eq. (12) satisfy the moment equations (S & T, p. 38)

$$\mu_k = \sum_{i=1}^{n} (\epsilon_i^{(n)})^k f_i^{(n)}, \quad k = 0, 1, \ldots, 2n-1, \tag{14}$$

as a consequence of the definition of $A^{(n)}(z)$. Since $A^{(n)}(z)$ has simple poles on the real axis it is a poor approximation there to $R_\Phi(z)$, which has an imaginary discontinuity across the real axis. However, $A^{(n)}(z)$ can provide an accurate approximation to $R_\Phi(z)$ for complex z off the real axis, and convergence to $R_\Phi(z)$ is achieved in the limit $n \to \infty$.

Stieltjes histogram approximations to the distribution $\Gamma(\epsilon)$ are obtained from the roots $\epsilon_i^{(n)}$ and residues $f_i^{(n)}$ of the nth-order approximant of Eqs. (9) and (12) in the form (S & T, p. 42)

$$\Gamma_s^{(n)}(\epsilon) = 0, \quad 0 \leq \epsilon < \epsilon_1^{(n)}, \tag{15a}$$

$$\Gamma_s^{(n)}(\epsilon) = \sum_{i=1}^{j} f_i^{(n)}, \quad \epsilon_j^{(n)} < \epsilon < \epsilon_{j+1}^{(n)}, \tag{15b}$$

$$\Gamma_s^{(n)}(\epsilon) = \sum_{i=1}^{n} f_i^{(n)} = \mu_0, \quad \epsilon_n^{(n)} < \epsilon. \tag{15c}$$

Equations (15) provide bounds on $\Gamma(\epsilon)$ at the points of increase $\epsilon_i^{(n)}$ in the form of the Tchebycheff inequalities $(m \geq n)$ (S & T, p. 43)

$$\begin{aligned} \Gamma_s^{(n)}(\epsilon_i^{(n)} - 0) \leq \Gamma_s^{(m)}(\epsilon_i^{(n)} - 0) \leq \Gamma(\epsilon_i^{(n)}) \\ \leq \Gamma_s^{(m)}(\epsilon_i^{(n)} + 0) \leq \Gamma_s^{(n)}(\epsilon_i^{(n)} + 0), \end{aligned} \tag{16}$$

and convergent approximations to $\Gamma(\epsilon)$ at the points of

increase $\epsilon_i^{(n)}$ are obtained from the Stieltjes values of Eqs. (15), defined according to

$$\Gamma_s^{(n)}(\epsilon_i^{(n)}) = \tfrac{1}{2}[\Gamma_s^{(n)}(\epsilon_i^{(n)} - 0) + \Gamma_s^{(n)}(\epsilon_i^{(n)} + 0)]. \tag{17}$$

Corresponding approximations to the density $g(\epsilon)$ are obtained from the Stieltjes histogram derivative of the distribution given by Eqs. (15)–(17). The Stieltjes histogram derivative of the histogram of Eqs. (15) is obtained from the slopes of the straight-line segments connecting the successive Stieltjes values of Eq. (17) in the form

$$g_s^{(n)}(\epsilon) = \tfrac{1}{2} f_1^{(n)}/\epsilon_1^{(n)}, \quad 0 \leq \epsilon < \epsilon_1^{(n)}, \tag{18a}$$

$$g_s^{(n)}(\epsilon) = \tfrac{1}{2}(f_{i+1}^{(n)} + f_i^{(n)})/(\epsilon_{i+1}^{(n)} - \epsilon_i^{(n)}), \quad \epsilon_i^{(n)} < \epsilon < \epsilon_{i+1}^{(n)}, \tag{18b}$$

$$g_s^{(n)}(\epsilon) = 0, \quad \epsilon_n^{(n)} < \epsilon. \tag{18c}$$

Deltour has noted previously that the convergence of the Stieltjes distribution [Eq. (15)] is sufficient to insure the convergence of the associated Stieltjes density of Eq. (18) when $g(\epsilon)$ is continuous.[17] Moreover, the mean-value theorem insures that Eq. (18b) must equal $g(\epsilon)$ at some point ϵ in the interval $\epsilon_i^{(n)}$ to $\epsilon_{i+1}^{(n)}$ in this case.

Histogram approximations to the principal-value integral of Eq. (6) corresponding to the density of Eqs. (18) are obtained from

$$\begin{aligned} \mathrm{Re} R_\Phi^{(n)}(\omega_i^{(n)} \pm i0) &= P \int_0^\infty \frac{g_s^{(n)}(\epsilon)\, d\epsilon}{\omega_i^{(n)} - \epsilon} \\ &= \sum_{j=1}^{n-1} g_s^{(n)}(\omega_j^{(n)}) \ln \left| \frac{\omega_i^{(n)} - \epsilon_j^{(n)}}{\omega_i^{(n)} - \epsilon_{j+1}^{(n)}} \right|, \end{aligned} \tag{19}$$

where $\omega_i^{(n)}$ is the midpoint of the interval $\epsilon_{i-1}^{(n)}$ to $\epsilon_i^{(n)}$. It is necessary to restrict the evaluation of Eq. (6) to the points $\omega_i^{(n)}$ when the Stieltjes density of Eqs. (18) is employed in order to avoid the appearance of spurious logarithmic singularities at the points $\epsilon_i^{(n)}$. Equation (19) defines the height of a histogram approximation to Eq. (6) in the interval $\epsilon_{i-1}^{(n)}$ to $\epsilon_i^{(n)}$, which is convergent in the limit of large numbers of spectral moments, as a consequence of the convergence of the Stieltjes density.[17]

4. TCHEBYCHEFF DISTRIBUTIONS AND DENSITIES

Although the Stieltjes distributions, densities, and principal-value integrals are in many respects satisfactory,[17,24,25] it is desirable to have convergent, continuous approximations to $\Gamma(\epsilon)$ and $g(\epsilon)$ at any point ϵ. It is convenient in this connection to consider approximants to the series of Eq. (7) having a prespecified pole at ϵ (S & T, p. 47),

$$\tilde{A}_\tau^{(n)}(z) = \tilde{p}_{n-1}(z)/\tilde{q}_n(z), \tag{20}$$

where[26]

$$\tilde{p}_{n-1}(z) = p_{n-1}(z) - \tau_n(\epsilon) p_{n-2}(z), \tag{21a}$$

$$\tilde{q}_n(z) = q_n(z) - \tau_n(\epsilon) q_{n-1}(z), \tag{21b}$$

$$\tau_n(z) = q_n(\epsilon)/q_{n-1}(\epsilon). \tag{21c}$$

The $\tilde{q}_n(z)$ of Eq. (21b) are so-called quasiorthogonal polynomials of order n associated with the distribution $\Gamma(\epsilon)$, and the $\tilde{p}_{n-1}(z)$ are the related numerator polynomials. They satisfy the quasirecurrence relations

$$\tilde{p}_{n-1}(z) = [z - \tilde{\alpha}_n(\epsilon)] p_{n-2}(z) - \beta_{n-1} p_{n-3}(z), \tag{22a}$$

$$\tilde{q}_n(z) = [z - \tilde{\alpha}_n(\epsilon)] q_{n-1}(z) - \beta_{n-1} q_{n-2}(z), \tag{22b}$$

involving the orthogonal polynomials of Eqs. (10), where

$$\tilde{\alpha}_n(\epsilon) = \epsilon - [\beta_{n-1}/\tau_{n-1}(\epsilon)]. \tag{23}$$

As the parameter $\tau = \tau_n(\epsilon)$ in Eqs. (21) is varied over the entire real axis, the approximant $A_\tau^{(n)}(z)$ of Eq. (20) generates the so-called circle bounds on $R_\Phi(z)$ at the complex point z off the real axis (S & T, p. 48).[27]

The (real) roots $\epsilon_i^{(n)}(\epsilon)$ and (positive) residues $f_i^{(n)}(\epsilon)$ of Eq. (20) satisfy the moment problem

$$\mu_k = \sum_{i=1}^{n} [\epsilon_i^{(n)}(\epsilon)]^k f_i^{(n)}(\epsilon), \quad k = 0, 1, \ldots, 2n-2, \tag{24}$$

where $\epsilon \equiv \epsilon_i^{(n)}(\epsilon)$ for some value of i, since ϵ is one of the roots of $\tilde{q}_n(z)$. The roots and residues are obtained from the equations (S & T, pp. 39–41)

$$\tilde{q}_n[\epsilon_i^{(n)}(\epsilon)] = 0, \tag{25a}$$

$$\begin{aligned} f_i^{(n)}(\epsilon) &= \tilde{p}_{n-1}[\epsilon_i^{(n)}(\epsilon)]/\tilde{q}_n'[\epsilon_i^{(n)}(\epsilon)] \\ &= \left\{ \sum_{k=0}^{n-1} N_k^{-1} q_k^2[\epsilon_i^{(n)}(\epsilon)] \right\}^{-1}, \end{aligned} \tag{25b}$$

similar to the results of Eqs. (13), with N_k given by Eq. (11). As indicated, the $\epsilon_i^{(n)}(\epsilon)$ and $f_i^{(n)}(\epsilon)$ are functions of ϵ, which can be varied at will over the real axis.[28] In terms of these values, the approximant of Eq. (20) takes the form

$$\tilde{A}_\tau^{(n)}(z) = \sum_{i=1}^{n} \frac{f_i^{(n)}(\epsilon)}{z - \epsilon_i^{(n)}(\epsilon)}, \tag{26}$$

similar to Eq. (12).

Bounds on $\Gamma(\epsilon)$ are obtained from the Stieltjes histogram of Eq. (15), having a prespecified frequency point $\epsilon \equiv \epsilon_j^{(n)}(\epsilon)$ with weight $f_j^{(n)}(\epsilon)$ and $n-1$ additional points $\epsilon_i^{(n)}(\epsilon)$ and weights $f_i^{(n)}(\epsilon)$, in the form of the Tchebycheff inequalities ($m \geq n$) (S & T, p. 43)

$$\begin{aligned} \Gamma_s^{(n)}(\epsilon - 0) \leq \Gamma_s^{(m)}(\epsilon - 0) \leq \Gamma(\epsilon) \leq \Gamma_s^{(m)}(\epsilon + 0) \\ \leq \Gamma_s^{(n)}(\epsilon + 0). \end{aligned} \tag{27}$$

A convergent approximation to $\Gamma(\epsilon)$ is obtained from the Stieltjes value of Eq. (17) in the form

$$\begin{aligned} \Gamma_t^{(n)}(\epsilon) &= \tfrac{1}{2}[\Gamma_s^{(n)}(\epsilon - 0) + \Gamma_s^{(n)}(\epsilon + 0)] \\ &= \sum_{i<j} f_i^{(n)}(\epsilon) + \tfrac{1}{2} f_j^{(n)}(\epsilon), \quad \epsilon \equiv \epsilon_j^{(n)}(\epsilon), \end{aligned} \tag{28}$$

which is simply the midpoint of the Tchebycheff bounds of Eq. (27) at $\epsilon \equiv \epsilon_j^{(n)}(\epsilon)$ and which we here designate as the Tchebycheff distribution of order n. In Eq. (28), the sum is over all weights $f_i^{(n)}(\epsilon)$ corresponding to points $\epsilon_i^{(n)}(\epsilon) < \epsilon \equiv \epsilon_j^{(n)}(\epsilon)$.

An approximation to the density $g(\epsilon)$ is obtained from

the derivative of Eq. (28) in the form

$$g_t^{(n)}(\epsilon)=\frac{d\Gamma_t^{(n)}(\epsilon)}{d\epsilon}$$

$$=\sum_{i<j}\frac{df_i^{(n)}(\epsilon)}{d\epsilon}+\frac{1}{2}\frac{df_j^{(n)}(\epsilon)}{d\epsilon},\quad \epsilon\equiv\epsilon_j^{(n)}(\epsilon), \tag{29}$$

and a corresponding approximation to the principal-value integral of Eq. (6) is given by

$$\mathrm{Re}R_\Phi^{(n)}(\epsilon\pm i0)=P\int_0^\infty\frac{g_t^{(n)}(\omega)\,d\omega}{\epsilon-\omega}. \tag{30}$$

Equation (29) defines the Tchebycheff density of order n correponding to the convergent Tchebycheff distribution of Eq. (28), and Eq.(30) gives a related approximation to the associated principal-value integral.

The Tchebycheff density of Eq. (29) can be expressed entirely in terms of the polynomials appearing in Eqs. (21b) and (25b), and their derivatives, which are determined by the recurrence relations of Eq. (10b). We first note from the second form of Eq. (25b) that the residues $f_i^{(n)}(\epsilon)$ can be regarded as explicit functions of the $\epsilon_i^{(n)}(\epsilon)$. Consequently, the derivatives appearing in the Tchebycheff density of Eq. (29) are conveniently evaluated in the form

$$\frac{df_i^{(n)}(\epsilon)}{d\epsilon}=\frac{df_i^{(n)}[\epsilon_i^{(n)}(\epsilon)]}{d\epsilon_i^{(n)}(\epsilon)}\frac{d\epsilon_i^{(n)}(\epsilon)}{d\epsilon}. \tag{31}$$

The first term in the product on the right-hand side of Eq. (31) is obtained from Eq. (25b) in the form

$$\frac{df_i^{(n)}[\epsilon_i^{(n)}(\epsilon)]}{d\epsilon_i^{(n)}(\epsilon)}=-\left[2\sum_{k=0}^{n-1}N_k^{-1}q_k[\epsilon_i^{(n)}(\epsilon)]q_k'[\epsilon_i^{(n)}(\epsilon)]\right]\Bigg/\left[\sum_{k=0}^{n-1}N_k^{-1}q_k^2[\epsilon_i^{(n)}(\epsilon)]\right]^2. \tag{32}$$

To evaluate the second term in the product in Eq. (31), note that Eq. (25a) can be written as

$$\frac{q_{n-1}[\epsilon_i^{(n)}(\epsilon)]}{q_n[\epsilon_i^{(n)}(\epsilon)]}-\frac{q_{n-1}(\epsilon)}{q_n(\epsilon)}=0. \tag{33}$$

From the total differential of Eq. (33) we obtain

$$\frac{d\epsilon_i^{(n)}(\epsilon)}{d\epsilon}=\frac{\{q_{n-1}'(\epsilon)q_n[\epsilon_i^{(n)}(\epsilon)]-q_n'(\epsilon)q_{n-1}[\epsilon_i^{(n)}(\epsilon)]\}}{\{q_n(\epsilon)q_{n-1}'[\epsilon_i^{(n)}(\epsilon)]-q_{n-1}(\epsilon)q_n'[\epsilon_i^{(n)}(\epsilon)]\}}. \tag{34}$$

Making use of Eqs. (25b) and (33), Eq. (34) can be written in the alternative forms

$$\frac{d\epsilon_i^{(n)}(\epsilon)}{d\epsilon}=(q_{n-1}'(\epsilon)q_n(\epsilon)-q_n'(\epsilon)q_{n-1}(\epsilon))q_n^2(\epsilon)$$
$$\times\{(q_{n-1}'[\epsilon_i^{(n)}(\epsilon)]q_n[\epsilon_i^{(n)}(\epsilon)]$$
$$-q_n'[\epsilon_i^{(n)}(\epsilon)]q_{n-1}[\epsilon_i^{(n)}(\epsilon)])q_n^2[\epsilon_i^{(n)}(\epsilon)]\}^{-1} \tag{35a}$$
$$=\frac{f_i^{(n)}(\epsilon)q_n^2(\epsilon)}{f_j^{(n)}(\epsilon)q_n^2[\epsilon_i^{(n)}(\epsilon)]}, \tag{35b}$$

where the latter expression is obtained using

$$f_i^{(n)}(\epsilon)$$
$$=\frac{N_{n-1}}{q_{n-1}[\epsilon_i^{(n)}(\epsilon)]q_n'[\epsilon_i^{(n)}(\epsilon)]-q_n[\epsilon_i^{(n)}(\epsilon)]q_{n-1}'[\epsilon_i^{(n)}(\epsilon)]}, \tag{36}$$

which follows from manipulation of the first form of Eq. (25b) (S & T, p. 46). Equations (32)—(36) are, of course, valid in the case $i=j$ under our convention $\epsilon_j^{(n)}(\epsilon)\equiv\epsilon$. Consequently, the Tchebycheff derivative of Eq. (29) can be written in the form

$$g_t^{(n)}(\epsilon)=\sum_{i<j}\left\{\frac{f_i^{(n)}(\epsilon)^3q_n^2[\epsilon_i^{(n)}(\epsilon)]}{f_j^{(n)}(\epsilon)q_n^2(\epsilon)}\times\left(2\sum_{k=0}^{n-1}N_k^{-1}q_k[\epsilon_i^{(n)}(\epsilon)]q_k'[\epsilon_i^{(n)}(\epsilon)]\right)\right\}$$
$$+f_j^{(n)}(\epsilon)^2\sum_{k=0}^{n-1}N_k^{-1}q_k(\epsilon)q_k'(\epsilon),\quad \epsilon\equiv\epsilon_j^{(n)}(\epsilon). \tag{37}$$

Although alternative forms can be devised, Eq. (37) is sufficient for establishing, in conjunction with the development of Eqs. (31)—(36), the analytic properties of the Tchebycheff density, described in the following section.

5. PROPERTIES OF THE TCHEBYCHEFF DENSITY

In order to establish the analytic properties of the Tchebycheff density, it is first necessary to determine the number of terms that contribute to the sum in Eq. (37) as ϵ is varied over the real axis, and, consequently, to determine the number of $\epsilon_i^{(n)}(\epsilon)<\epsilon$. We note in this connection that the orthogonal polynomial of degree $n-1$, $q_{n-1}(z)$, is also a quasiorthogonal polynomial of order n having a zero weight at infinity which is linearly independent of the quasiorthogonal polynomial $\tilde{q}_n(z)$ of Eq. (21b). Consequently, the respective roots of $q_{n-1}(z)$ and $\tilde{q}_n(z)$ are distinct and mutually separated, or interlace (S & T, p. 37). Thus, the $n-1$ roots $\epsilon_i^{(n-1)}$ of $q_{n-1}(z)$ specify n intervals $-\infty-\epsilon_1^{(n-1)}$, $\epsilon_1^{(n-1)}-\epsilon_2^{(n-1)}$, ..., $\epsilon_i^{(n-1)}-\epsilon_{i+1}^{(n-1)}$, ..., $\epsilon_{n-1}^{(n-1)}-+\infty$, on the real axis in each of which there must be one and only one of the roots of $\tilde{q}_n(z)$. As ϵ is varied over the open interval $\epsilon_i^{(n-1)}<\epsilon<\epsilon_{i+1}^{(n-1)}$, the $\epsilon_i^{(n)}(\epsilon)$ traverse their respective intervals in a strictly positive sense [Eq. (35b)], and the number of terms in the sum in Eq. (37) does not change. Consequently, the Tchebycheff density is a rational function of ϵ and the $\epsilon_i^{(n)}(\epsilon)$ on the open intervals $\epsilon_i^{(n-1)}<\epsilon<\epsilon_{i+1}^{(n-1)}$, the properties of which are easily discerned. However, as $\epsilon\to\epsilon_{i+1}^{(n-1)}$ from below and passes into the next interval, the other roots of $\tilde{q}_n(z)$ pass through the upper end points (including $+\infty$) of their respective intervals, and an additional term contributes to the sum of Eq. (37). Consequently, we can anticipate that the analytic properties of the Tchebycheff density on the real axis are controlled by the behavior of Eq. (37) at the roots $\epsilon_i^{(n-1)}$ of the orthogonal polynomial $q_{n-1}(z)$.

Lemma (a): The Tchebycheff density is real, nonnegative, and substantially convergent to $g(\epsilon)$ in the limit $n\to\infty$ in the case of a determined moment problem for a continuous density $g(\epsilon)$.

Equation (37) is real by virtue of the fact that it is a real function of real arguments ϵ and $\epsilon_i^{(n)}(\epsilon)$, nonnegativity follows from the nondecreasing nature of the Tchebycheff distribution of Eq. (28), which is easily established using the Tchebycheff inequalities [Eq. (27)], and substantial convergence for a determined moment problem is assured by the associated convergence of the Tchebycheff distribution and the restriction to continuous densities.

Comment (*a*): When a denumerable number of finite-sized steps are present in the distribution $\Gamma(\epsilon)$, the Tchebycheff density $g_t^{(n)}(\epsilon)$ is convergent in the delta-function sense at the appropriate points.

Lemma (*b*): The nth-order ($n \geq 2$) Tchebycheff density is continuous on the real axis, has $2n-4$ continuous derivatives there ($1 \leq l \leq 2n-4$), and supports $2n-2$ positive-integer power moments ($0 \leq l \leq 2n-3$).

Continuity on the open interval $\epsilon_i^{(n-1)} < \epsilon < \epsilon_{i+1}^{(n-1)}$ follows by inspection of Eq. (37), the individual factors there, and in Eqs. (35) and (36) involving nonzero polynomials in ϵ and $\epsilon_i^{(n)}(\epsilon)$. To establish continuity at the points $\epsilon_i^{(n-1)}$, we examine the left- and right-hand limits of $g_t^{(n)}(\epsilon)$ there. Note that as $\epsilon \to \epsilon_i^{(n-1)}$ from below, the largest positive root of $\tilde{q}_n(\epsilon)$ must go to infinity. Consequently, we have from Eq. (37)

$$g_t^{(n)}(\epsilon_i^{(n-1)}+0) = g_t^{(n)}(\epsilon_i^{(n-1)}-0) + \lim(\epsilon_n^{(n)}(\epsilon)\to\infty)\frac{df_n^{(n)}(\epsilon)}{d\epsilon}, \tag{38}$$

where $\epsilon_n^{(n)}(\epsilon)$ is the root of $\tilde{q}_n(\epsilon)$ that goes to infinity as $\epsilon \to \epsilon_i^{(n-1)}$. From Eq. (32)

$$\frac{df_n^{(n)}[\epsilon_n^{(n)}(\epsilon)]}{d\epsilon_n^{(n)}(\epsilon)} \to \frac{2(n-1)N_{n-1}}{[\epsilon_n^{(n)}(\epsilon)]^{2n-1}}, \quad \text{as } \epsilon_n^{(n)}(\epsilon)\to\pm\infty, \tag{39}$$

and

$$\frac{d\epsilon_n^{(n)}(\epsilon)}{d\epsilon} \to \frac{q_n[\epsilon_n^{(n)}(\epsilon)]q'_{n-1}(\epsilon_i^{(n-1)})}{q'_{n-1}[\epsilon_n^{(n)}(\epsilon)]q_n(\epsilon_i^{(n-1)})},$$
$$\text{as } \epsilon\to\epsilon_i^{(n-1)}, \quad \epsilon_n^{(n)}(\epsilon)\to\pm\infty, \tag{40}$$

from Eq. (35a). Noting that $\epsilon_i^{(n-1)}$ cannot be a root of $q_n(z)$ or $q'_{n-1}(z)$ [S & T, p. 37], we have from Eqs. (39) and (40)

$$\frac{df_n^{(n)}(\epsilon)}{d\epsilon} = O\left(\frac{1}{\epsilon_n^{(n)}(\epsilon)^{2n-3}}\right)\to 0,$$
$$\text{as } \epsilon\to\epsilon_i^{(n-1)}, \quad \epsilon_n^{(n)}(\epsilon)\to\pm\infty, \tag{41}$$

so that, from Eq. (38),

$$g_t^{(n)}(\epsilon_i^{(n-1)}+0) = g_t^{(n)}(\epsilon_i^{(n-1)}-0), \tag{42}$$

establishing continuity of $g_t^{(n)}(\epsilon)$ at the points $\epsilon_i^{(n-1)}$. Similarly, continuity of the derivatives of Eq. (37) on the open interval $\epsilon_i^{(n)} < \epsilon < \epsilon_{i+1}^{(n)}$ follows by inspection, whereas, at the points $\epsilon_i^{(n)}$ we have

$$\frac{d^l g_t^{(n)}(\epsilon+0)}{d\epsilon^l} = \frac{d^l g_t^{(n)}(\epsilon-0)}{d\epsilon^l} + \lim(\epsilon_n^{(n)}(\epsilon)\to\infty)\frac{d^{l+1}f_n^{(n)}(\epsilon)}{d\epsilon^{l+1}}, \tag{43}$$

where $\epsilon_n^{(n)}(\epsilon)$ is the root of $\tilde{q}_n(\epsilon)$ that goes to infinity as $\epsilon \to \epsilon_i^{(n-1)}$. From Eqs. (32) and (35a) we have

$$\frac{d^{l+1}f_n^{(n)}(\epsilon)}{d\epsilon^{l+1}} = O\left(\frac{1}{\epsilon_n^{(n)}(\epsilon)^{2n-3-l}}\right),$$
$$\text{as } \epsilon\to\epsilon_i^{(n-1)}, \quad \epsilon_n^{(n)}(\epsilon)\to\pm\infty, \tag{44}$$

demonstrating that $2n-4$ derivatives of $g_t^{(n)}(\epsilon)$ are continuous at the points $\epsilon_i^{(n-1)}$. Finally, the number of convergent moments $g_t^{(n)}(\epsilon)$ supports is determined by its limits as $\epsilon \to \pm\infty$. We have, from Eqs. (35)–(37),

$$g_t^{(n)}(\epsilon\to\pm\infty) = \mp\frac{1}{2}\frac{df_j^{(n)}(\epsilon)}{d\epsilon} = O\left(\frac{1}{\epsilon^{2n-1}}\right), \tag{45}$$

where $f_j^{(n)}(\epsilon)$ is the weight at the point $\epsilon \equiv \epsilon_j^{(n)}(\epsilon)$. Equation (45) insures that $2n-2$ power moments of $g_t^{(n)}(\epsilon)$ converge.

Comment (*b*): The Tchebycheff density $g_t^{(n)}(\epsilon)$ does not satisfy the reduced moment problem of order n, since only $2n-2$ of its positive-integer power moments converge.[29] Equation (45) insures that the principal-value integral of Eq. (30) converges as a quadrature for $n \geq 2$. Convergence to $\mathrm{Re}R_\Phi(\epsilon \pm i0)$ in the limit $n \to \infty$ is assured by Lemma (a).

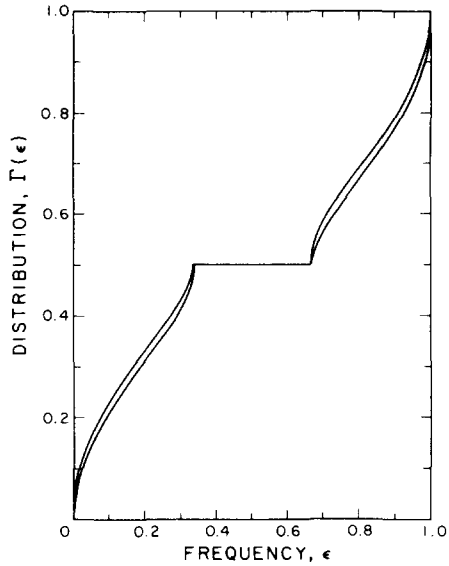


FIG. 1. Upper and lower Tchebycheff bounds on the distribution of normal-mode vibrations in a diatomic linear chain with mass ratio 2:1. Values obtained employing 100 spectral moments and the development of Sec. 4. All values are dimensionless, with $\epsilon = (\omega/\omega_{max})^2$, where ω_{max} is the largest permissible vibrational frequency.

6. ILLUSTRATIVE APPLICATION

As a simple application of the Tchebycheff development, we investigate the vibrations of a linear diatomic chain with mass ratio 2:1. The spectrum is nonzero in this case in the interval $0 \leq \omega \leq \omega_{max}$, where ω_{max} ($< \infty$) is the largest allowed vibrational frequency. It is convenient to employ the dimensionless variable $\epsilon = (\omega/\omega_{max})^2$ in the development. The moments of Eq. (4) in this case, which correspond to the traces of even powers of the dynamical matrix,[30] are computed efficiently from the algorithm of Isenberg.[31] These moments are employed in the Sack and Donovan[32] version of the Stieltjes–Tchebycheff algorithm[22] for calculating the recurrence coefficients α_i, β_i of Eqs. (8) and (10). The coefficients are employed in evaluating the polynomials of Eqs. (10b) and (21b), and in the calculation of the

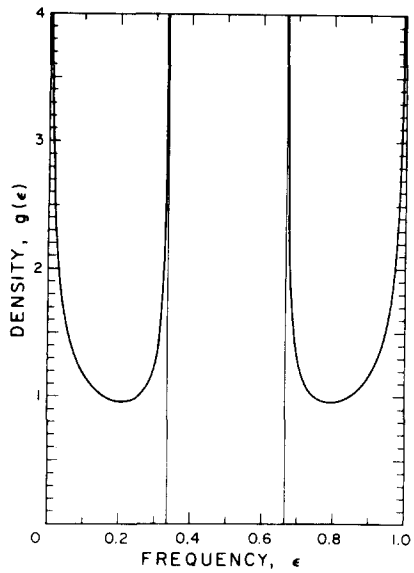


FIG. 2. Tchebycheff approximation to the density of normal-mode vibrations in a diatomic chain, obtained from the derivative of the mean values of the Tchebycheff bounds shown in Fig. 1. The results shown are indistinguishable from the correct values (Ref. 34). All values are dimensionless, with $\epsilon = (\omega/\omega_{max})^2$, as in Fig. 1.

roots $\epsilon_i^{(n)}(\epsilon)$ and residues $f_i^{(n)}(\epsilon)$ of Eqs. (20) and (25) using matrix diagonalization procedures.[33]

We have employed up to 100 spectral moments in steps of ten moments in constructing the Tchebycheff distributions and densities of Eqs. (28) and (29), respectively, for the diatomic chain. The distributions and densities are found to converge rapidly to the values shown in Figs. 1 and 2, obtained employing 100 spectral moments. The results of Figs. 1 and 2, which are seen to provide faithful images of the forbidden band gap and upper and lower van Hove singularities in both optical and acoustical branches, are indistinguishable from the corresponding correct values.[34] Moreover, the use of 50 moments provides Tchebycheff distributions and densities (not shown) in excellent quantitative agreement with Figs. 1 and 2, respectively, and as few as ten moments provide useful qualitative information (not shown) when employed in the Tchebycheff development.

7. CONCLUDING REMARKS

Convergent Tchebycheff approximations are described for nonnegative spectral densities and associated Stieltjes integrals. The approximations are obtained from derivatives of the mean (Stieltjes) values of Tchebycheff bounds on nondecreasing distributions, constructed from finite numbers of spectral moments. It is shown that the nth-order Tchebycheff density is real, nonnegative, and continuous, with $2n-4$ continuous derivatives, on the real axis, and that it supports $2n-2$ positive-integer power moments. Related approximations to the associated Stieltjes integral are obtained directly from the appropriate principal-value quadratures. Although the Tchebycheff densities are convergent in the limit of large numbers of spectral moments for determined moment problems, they are not solutions of reduced moment problems of appropriate finite order. An illustrative application in the case of normal-mode vibrations in a diatomic chain indicates that the Tchebycheff densities are suitably convergent.

ACKNOWLEDGMENTS

We are grateful to G. A. Baker, Jr., and K. B. Winterbon for helpful correspondence and comments, to the JILA Fellows for their kind hospitality and support, and to the Proctor and Gamble Corporation for financial assistance.

*Present address: Theoretical Chemistry Institute, University of Wisconsin, Madison, Wisconsin 53706.
†JILA Visiting Fellow, 1976.

[1]E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals* (Oxford U. P., London, 1948), pp. 119–51.
[2]J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, Mathematical Surveys 1 (American Mathematical Society, Providence, R.I., 1950), 2nd ed.
[3]G. A. Baker, Jr. and J. L. Gammel (Eds.), *The Padé Approximant in Theoretical Physics* (Academic, New York, 1970).
[4]P. C. Martin, Phys. Rev. **161**, 143 (1967).
[5]R. Kubo and M. Ichimura, J. Math. Phys. **13**, 1454 (1972).
[6]R. G. Gordon, J. Math. Phys. **9**, 655 (1968).
[7]A. Lonke, J. Math. Phys. **12**, 2422 (1971).
[8]J. Peretti, J. Phys. Chem. Solids **12**, 216 (1960).
[9]J. C. Wheeler and R. G. Gordon, J. Chem. Phys. **51**, 5566 (1969).
[10]P. W. Langhoff and C. T. Corcoran, Mol. Phys. **26**, 203 (1972).
[11]M. Baker, Ann. Phys. **4**, 271 (1958).
[12]R. G. Newton, *Scattering Theory of Waves and Particles* (McGraw-Hill, New York, 1966), Chap. 12.
[13]H. M. Nussenzveig, *Causality and Disperson Relations* (Academic, New York, 1972).
[14]N. I. Akhiezer, *The Classical Moment Problem* (Oliver and Boyd, London, 1965).
[15]YuV. Vorobyev, *Method of Moments in Applied Mathematics* (Gordon and Breach, New York, 1965).
[16]The problem of constructing a density from a given finite set of spectral moments has no unique solution and, consequently, is one of continuing interest. A considerable number of procedures for imaging spectral densities have been introduced in various contexts. Particularly noteworthy are the early papers of E. W. Montroll, J. Chem. Phys. **10**, 218 (1942) and **11**, 481 (1943), who employed Legendre-polynomial expansions in approximating the spectral densities of crystal lattice vibrations. The polynomial-expansion approach has been extended to the use of modified spectral moments recently by J. C. Wheeler, M. G. Prais, and C. Blumstein, Phys. Rev. B **10**, 2429 (1974), who cite additional pertinent references. Although a detailed review of the techniques devised for approximating spectral densities is beyond the scope of the present work, we comment in the sequel at appropriate points on previous work related to the methods employed in the present development.
[17]In a previous development, J. Deltour, Physica 39, 413, 424, 431 (1968), has employed histogram, or Stieltjes, derivatives of associated Stieltjes distributions in constructing approximations to the densities of lattice vibrations in solids, and has applied the method to the monatomic chain. The estimates he obtains for thermodynamic properties were previously shown to provide bounds by E. W. Montroll, Quant. Appl. Math. 5, 223 (1947), and subsequently by J. C. Wheeler and R. G. Gordon, Ref. 9. Tchebycheff distributions were employed by K. B. Winterbon, Atomic Energy of Canada Limited, Publ. #4832 (1974), in the construction of continuous approximations to densities associated with the penetration of

charged particles in matter. We are grateful to Dr. Winterbon for informing us of his work, for providing us with appropriate preprints, and for recommending the analytic development of the Tchebycheff density. In a private communication, G.A. Baker, Jr., also independently suggested to us the use of Tchebycheff distributions in approximating associated densities.

[18]P.W. Langhoff, C.T. Corcoran, J.S. Sims, F. Weinhold, and R.G. Glover, Phys. Rev. A **14**, 1042 (1976).

[19]P.W. Langhoff and C.T. Corcoran, Chem. Phys. Lett. **40**, 367 (1976).

[20]It is convenient, but not necessary, to introduce Stieltjes integrals and spectral densities in the context of the resolvent of a Hamiltonian. For a recent application of the theory of moments to the study of the discrete region of Eq. (1), see, D. Bessis and M. Villani, J. Math. Phys. **16**, 462 (1975).

[21]In the present section, and in subsequent ones, we make extensive use of well-known results from the classical theory of moments, described in detail by J.A. Shohat and J.D. Tamarkin, Ref. 2, referred to as (S & T) from now on.

[22]H.S. Wall, *Analytic Theory of Continued Fractions* (Van Nostrand, New York, 1948), Chap. XI.

[23]G.A. Baker, Jr., *Essentials of Padé Approximants* (Academic, New York, 1975).

[24]P.W. Langhoff, J.S. Sims, and C.T. Corcoran, Phys. Rev. A **10**, 829 (1974).

[25]P.W. Langhoff and C.T. Corcoran, J. Chem. Phys. **61**, 146 (1974).

[26]For notational simplicity we suppress explicit indication of the dependence of $\tilde{A}_\tau^{(n)}(z)$, $\tilde{p}_{n-1}(z)$, and $\tilde{q}_n(z)$ on ϵ, its presence being indicated by the upper tilde.

[27]The radius of the circle bounds on $R_\Phi(z)$ provided by Eq. (20) is $r_n = (2|y| \sum_{i=0}^{n-1} |N_i^{-1}\tilde{q}_i(z)|^2)^{-1}$, where $|y|$ is the distance from the real axis of the complex point z (S & T, p. 48). Consequently, as we have indicated above, when $y \to 0$, r_n diverges for any finite n. However, since the series $\sum_{i=0}^{\infty} |N_i^{-1}\tilde{q}_i(z)|^2$ diverges for a determined moment problem (S & T, p. 49), it may be possible to obtain bounds on $R_\Phi(z)$ on the real axis from a suitable limiting process in which $y \to 0$ as $n \to \infty$. We are unaware of published applications of such a limiting procedure in constructing approximations to a Stieltjes integral on the real axis from a given sequence of power moments.

[28]In contrast to the roots of Eq. (13a), which are all real and positive, $(n-1)$ of the roots of Eq. (25a) must be real and positive, while the remaining root need only be real, even though the density of Eq. (3) vanishes on the negative real axis. The weight [Eq. (25b)] associated with a negative root is found to be generally small, however, and it vanishes in the limit $n \to \infty$. The behavior of the roots of Eq. (25a) as ϵ is varied over the real axis is discussed in detail in Sec. 5.

[29]Although $g_t^{(n)}(\epsilon)$ supports $2n-2$ power moments, their numerical values do not necessarily agree with those of Eq. (4), particularly for small n. Numerical studies (Refs. 18 and 19 and Sec. 6) indicate, however, that the kth moment of Eq. (4) is accurately reproduced by $g_t^{(n)}(\epsilon)$ for $n \gg k$.

[30]A.A. Maradudin, E.W. Montroll, G.H. Weiss, and I.P. Ipatova, *Theory of Lattice Dynamics in the Harmonic Approximation*, Solid State Physics S3 (Academic, New York, 1971), 2nd ed.

[31]C. Isenberg, Phys. Rev. 132, 2427 (1963); J. Phys. C 3, L179 (1970); 4, 164 (1971).

[32]R.A. Sack and A.F. Donovan, Numer. Math. **18**, 465 (1972).

[33]In computational applications the Tchebycheff derivatives of Eq. (37) are evaluated at each of the steps of a Tchebycheff distribution of given order. Consequently, the fixed point ϵ need be varied only over one of the intervals $\epsilon_{i-1}^{(n-1)} < \epsilon < \epsilon_i^{(n-1)}$ to obtain $g_t^{(n)}(\epsilon)$ over its entire domain. Matrix diagonalization procedures similar to those described by C. Blumstein and J.C. Wheeler, Phys. Rev. B 8, 1764 (1973), are employed on the CDC 6600 computer.

[34]J. Deltour and E. Kartheuser, Physica **31**, 269 (1965). In a separate development not described here, we have employed the polynomial-moment approach of Wheeler *et al.* (Ref. 16), and an appropriate reference density, to obtain independently accurate values of the density of lattice vibrations in the diatomic chain.

Some examples of transparent potentials in the classical approximation*

Michel Cuer

Département de Physique Mathématique,† Université des Sciences et Techniques de Languedoc, 34060 Montpellier, Cedex, France
(Received 23 September 1976)

We give classes and examples of potentials which give a classical deflection function equal to zero (mod $2n\pi$) for any value of the impact parameter at a fixed energy. Thus these potentials are completely transparent for a classical collision at this energy.

1. INTRODUCTION

We consider a particle with mass m, angular momentum $M=\hbar(l+\frac{1}{2})$ in a central potential $V(r)$ at the energy E. When the particle comes from infinity, the classical deflection function is $|\pi-2\varphi|$ with

$$\varphi=\hbar(l+\tfrac{1}{2})(2m)^{-1/2}\int_{r_l}^{\infty}\{E-V(r)-\hbar^2(l+\tfrac{1}{2})^2/2mr^2\}^{-1/2}dr/r^2,$$

where the turning point r_l is the greatest value of r such that $E-V(r)-\hbar^2(l+\frac{1}{2})^2/2mr^2=0$.

Setting $\alpha=(l+\frac{1}{2})^2$, $\omega(\rho)=\rho^2(1-v(\rho))$, $v(\rho)=V(\rho/k)/E$, $k=\sqrt{2mE}/\hbar$, we can write:

$\varphi(\alpha)=\alpha^{1/2}\int_{\mu(\alpha)}^{\infty}(\omega(\rho)-\alpha)^{-1/2}d\rho/\rho$, where the function $\mu(\alpha)$ is the greatest solution of $\omega(\mu(\alpha))=\alpha$.

When $V(r)\equiv 0$, $\varphi(\alpha)=\pi/2$. In the following section, given a number $\rho_i>0$ and a positive integer n, we construct functions $\omega(\rho)$ equal to ρ^2 for $\rho\geq\rho_i$ with one maximum α_m in the interval $\rho\in[0,\rho_i]$ such that

$$\alpha_m>\rho_i^2 \quad \text{(see Fig. 1b)}, \tag{1}$$

(Cond. 1) $\varphi(\alpha)=\pi/2+n\pi$ for $0<\alpha\leq\rho_i^2$.

Because of Condition 1 the potentials $V(r)=E\{1-\omega(kr)/k^2r^2\}$ are transparent at the energy E in the classical approximation. In such a potential, the trajectories of particles with angular momentum $M<\rho_i/\hbar$ consist of n revolutions around the center (Fig. 1a).

Before going into the matter further, let us examine the interest of this result. The semiclassical approximation is frequently used in molecular and atomic physics (Buck[1]). The current assumption is that scattering experiments can give the interaction potential. But at fixed energy important kinds of ambiguities (Sabatier[2,3]) and transparent potentials (Sabatier,[4] Chadan[5]) can appear. We could think that this fact is only a quantum property linked with the interpolation of phase shifts. The following examples show that ambiguities in the inverse problem at fixed energy still exist in the classical limit. This simple result is connected with the possibility of nonmonotone functions $\omega(\rho)$.

2. A CLASS OF TRANSPARENT POTENTIALS IN CLASSICAL MECHANICS

The condition (1) is

$$\alpha^{1/2}\int_{\mu(\alpha)}^{\infty}[\omega(\rho)-\alpha]^{-1/2}d\rho/\rho=\pi/2+n\pi \quad \text{for } 0<\alpha\leq\rho_i^2.$$

Since $\pi/2=\alpha^{1/2}\int_{\alpha^{1/2}}^{\infty}(\rho^2-\alpha)^{-1/2}d\rho/\rho=\alpha^{1/2}\times\int_{\alpha}^{\infty}(\beta-\alpha)^{-1/2}d\log\sqrt{\beta}$, this can be written as

$$\alpha^{1/2}\int_{\mu(\alpha)}^{\rho_i}[\omega(\rho)-\alpha]^{-1/2}d\rho/\rho=\alpha^{1/2}\int_{\alpha}^{\rho_i^2}(\beta-\alpha)^{-1/2}d\log\sqrt{\beta}+n\pi \quad \text{for } 0<\alpha\leq\rho_i^2$$

[we have assumed $\omega(\rho)=\rho^2$ for $\rho\geq\rho_i$]. Because $\omega(\rho)$ is monotone in the intervals $[0,\rho_m]$ and $[\rho_m,\rho_i]$, we can introduce $\mu_1(\beta)$, the inverse function of $\omega(\rho)$ for $0\leq\rho\leq\rho_m$ and $\mu_2(\beta)$, the inverse function of $\omega(\rho)$ for $\rho_m\leq\rho\leq\rho_i$ (Fig. 1b). These two functions must be respectively increasing and decreasing. Setting $\mu_2(\beta)=\rho_i$ for $\beta\leq\rho_i^2$ condition (1) is equivalent to

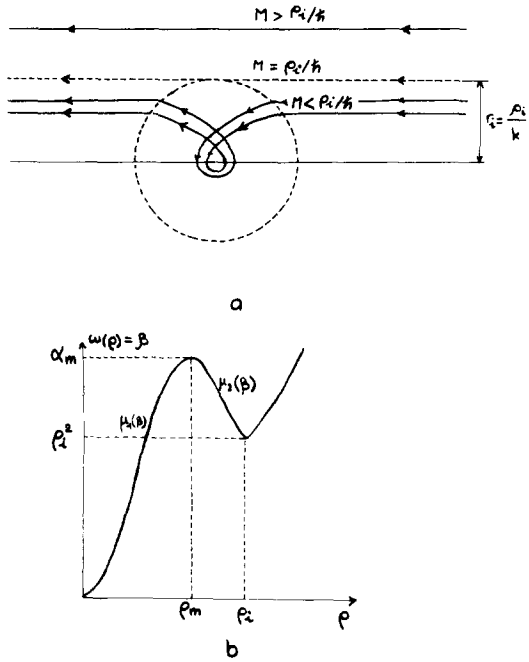


FIG. 1. (a) Behavior of trajectories of particles in a transparent potential; $r_i=\rho_i/k$ is the radius of the potential; M is the angular momentum of the particle; for $M>\rho_i/\hbar$ the trajectory is a straight line; for $M<\rho_i/\hbar$ there are exactly n revolutions around the center ($n=1$ in the figure). (b) The functions $\omega(\rho)$, $\mu_1(\beta)$, $\mu_2(\beta)$ and the parameters ρ_i, ρ_m, α_m: $\omega(\rho)$ is increasing for $0\leq\rho\leq\rho_m$ or $\rho\geq\rho_i$ and decreasing for $\rho_m\leq\rho\leq\rho_i$.

 Copyright © 1977 American Institute of Physics

(Cond. 2) $$\int_\alpha^{\alpha_m}(\beta-\alpha)^{-1/2}\,d\log\frac{\mu_1(\beta)}{\mu_2(\beta)}$$

$$=\int_\alpha^{\rho_i^2}(\beta-\alpha)^{-1/2}\,d\log\sqrt{\beta}+n\pi/\sqrt{\alpha}\quad\text{for } 0<\alpha\leq\rho_i^2. \tag{2}$$

Since $\mu_1(\alpha_m)=\mu_2(\alpha_m)$, the Abel transform

$$\theta'(\alpha)=\int_\alpha^{\alpha_m}(\beta-\alpha)^{-1/2}\,d\log[\mu_1(\beta)/\mu_2(\beta)]$$

can be inverted, giving

$$\log[\mu_1(\beta)/\mu_2(\beta)]=-(1/\pi)\int_\beta^{\alpha_m}(\alpha-\beta)^{-1/2}\theta'(\alpha)\,d\alpha.$$

So condition (2) is equivalent to

$$\log[\mu_1(\beta)/\mu_2(\beta)]$$

$$=-(1/\pi)\int_\beta^{\rho_i^2}(\alpha-\beta)^{-1/2}\left\{\int_\alpha^{\rho_i^2}(\gamma-\alpha)^{-1/2}\,d\log\sqrt{\gamma}+n\pi/\sqrt{\alpha}\right\}d\alpha-1/\pi\int_{\rho_i^2}^{\alpha_m}(\alpha-\beta)^{-1/2}\times\left\{\int_\alpha^{\alpha_m}(\gamma-\alpha)^{-1/2}\,d\log[\mu_1(\gamma)/\mu_2(\gamma)]\right\}d\alpha$$

$$\text{for } 0\leq\beta\leq\rho_i^2 \tag{3}$$

and

$$\int_{\rho_i^2}^{\alpha_m}(\beta-\rho_i^2)^{-1/2}\,d\log[\mu_1(\beta)/\mu_2(\beta)]=n\pi/\rho_i. \tag{4}$$

We now show that if $\mu_2(\beta)=\rho_i$ for $\beta\leq\rho_i^2$, if the function $\beta\to\log[\mu_1(\beta)/\mu_2(\beta)]$ is increasing for $\rho_i^2\leq\beta\leq\alpha_m$, if $\mu_1(\alpha_m)=\mu_2(\alpha_m)$, and if (4) is satisfied, then the function $\mu_1(\beta)$, defined by (3) for $0\leq\beta\leq\rho_i^2$, is increasing. So we will be able to construct a function $\omega(\rho)$ and a transparent potential.

Let us invert the integrations in (3) and remark that

$$\int\alpha^{-1/2}(\alpha-\beta)^{-1/2}\,d\alpha=2\,\text{arg cosh}\sqrt{\alpha/\beta},$$

$$\int(\alpha-\beta)^{-1/2}(\gamma-\alpha)^{-1/2}\,d\alpha=2\arcsin[(\alpha-\beta)^{1/2}/(\gamma-\beta)^{1/2}].$$

It follows

$$\log[\mu_1(\beta)/\rho_i]=\log(\sqrt{\beta}/\rho_i)-2n\,\text{arg cosh}(\rho_i/\sqrt{\beta})-(1/\pi)\int_{\rho_i^2}^{\alpha_m}\{\pi-2\arcsin[(\rho_i^2-\beta)^{1/2}/(\gamma-\beta)^{1/2}]\}\times d\log[\mu_1(\gamma)/\mu_2(\gamma)]\quad\text{for } 0\leq\beta\leq\rho_i^2.$$

Since $\pi/2-\arcsin[(\rho_i^2-\beta)^{1/2}/(\gamma-\beta)^{1/2}]$
$=\arcsin[(\gamma-\rho_i^2)^{1/2}/(\gamma-\beta)^{1/2}]$, this can be written as

$$\log\mu_1(\beta)=\log\sqrt{\beta}-2n\,\text{arg cosh}(\rho_i/\sqrt{\beta})-(2/\pi)\int_{\rho_i^2}^{\alpha_m}\arcsin[(\gamma-\rho_i^2)^{1/2}/(\gamma-\beta)^{1/2}]\times d\log[\mu_1(\gamma)/\mu_2(\gamma)]. \tag{5}$$

So

$$\frac{\mu_1'(\beta)}{\mu_1(\beta)}=\frac{1}{2\beta}+\frac{n\rho_i}{\beta(\rho_i^2-\beta)^{1/2}}-\frac{1}{\pi}\int_{\rho_i^2}^{\alpha_m}\frac{(\gamma-\rho_i^2)^{1/2}}{(\gamma-\beta)(\rho_i^2-\beta)^{1/2}}\times d\log\frac{\mu_1(\gamma)}{\mu_2(\gamma)}\quad\text{for } 0\leq\beta\leq\rho_i^2$$

and since

$$\frac{1}{\gamma-\beta}=\frac{1}{\gamma-\rho_i^2}-\frac{\rho_i^2-\beta}{(\gamma-\beta)(\gamma-\rho_i^2)},$$

$$\frac{\mu_1'(\beta)}{\mu_1(\beta)}=\frac{1}{2\beta}+\frac{n\rho_i}{\beta(\rho_i^2-\beta)^{1/2}}-\frac{1}{\pi}\int_{\rho_i^2}^{\alpha_m}\left(\frac{1}{(\gamma-\rho_i^2)^{1/2}(\rho_i^2-\beta)^{1/2}}-\frac{(\rho_i^2-\beta)^{1/2}}{(\gamma-\beta)(\gamma-\rho_i^2)^{1/2}}\right)d\log\frac{\mu_1(\gamma)}{\mu_2(\gamma)}\quad\text{for } 0\leq\beta\leq\rho_i^2.$$

If we substitute (4) in this expression, we obtain

$$\frac{\mu_1'(\beta)}{\mu_1(\beta)}=\frac{1}{2\beta}+\frac{n(\rho_i^2-\beta)^{1/2}}{\beta\rho_i}+\frac{(\rho_i^2-\beta)^{1/2}}{\pi}\int_{\rho_i^2}^{\alpha_m}\frac{1}{(\gamma-\beta)(\gamma-\rho_i^2)^{1/2}}\times d\log\frac{\mu_1(\gamma)}{\mu_2(\gamma)}\quad\text{for } 0\leq\beta\leq\rho_i^2.$$

Thus, if the function $\beta\to\log[\mu_1(\beta)/\mu_2(\beta)]$ is increasing for $\rho_i^2\leq\beta\leq\alpha_m$ and satisfies (4), the function $\beta\to\log\mu_1(\beta)$ defined by (5) for $0\leq\beta\leq\rho_i^2$, is increasing.

3. SOME EXPLICIT EXAMPLES

Let us examine the potentials when

$$\frac{\mu_1'(\beta)}{\mu_1(\beta)}-\frac{\mu_2'(\beta)}{\mu_2(\beta)}=\frac{n}{\rho_i}(\alpha_m-\beta)^{-1/2}\quad\text{for } \rho_i^2\leq\beta\leq\alpha_m.$$

The condition (4) is satisfied and

$$\mu_1(\beta)/\mu_2(\beta)=\exp[-(2n/\rho_i)(\alpha_m-\beta)^{1/2}]\quad\text{for } \rho_i^2\leq\beta\leq\alpha_m. \tag{6}$$

Besides,

$$\frac{(\rho_i^2-\beta)^{1/2}}{\pi}\int_{\rho_i^2}^{\alpha_m}\frac{1}{(\gamma-\beta)(\gamma-\rho_i^2)^{1/2}}\,d\log\frac{\mu_1(\gamma)}{\mu_2(\gamma)}=\frac{n}{\rho_i(\alpha_m-\beta)^{1/2}}.$$

Thus we obtain

$$\frac{\mu_1'(\beta)}{\mu_1(\beta)}=\frac{1}{2\beta}+\frac{n(\rho_i^2-\beta)^{1/2}}{\rho_i\beta}+\frac{n}{\rho_i(\alpha_m-\beta)^{1/2}}$$

and

$$\log\frac{\mu_1(\beta)}{\mu_1(\rho_i^2)}=\tfrac{1}{2}\log\frac{\beta}{\rho_i^2}+n\log\left(\frac{\rho_i-(\rho_i^2-\beta)^{1/2}}{\rho_i+(\rho_i^2-\beta)^{1/2}}\right)+2\frac{n}{\rho_i}[(\rho_i^2-\beta)^{1/2}-(\alpha_m-\beta)^{1/2}+(\alpha_m-\rho_i^2)^{1/2}]\quad\text{for } 0\leq\beta\leq\rho_i^2. \tag{7}$$

If we choose $\mu_1'(\beta)/\mu_1(\beta)=(n/2\rho_i)(\alpha_m-\beta)^{-1/2}$ for $\rho_i^2\leq\beta\leq\alpha_m$, the transparent potential is determined by (6), (7), $\mu_2(\rho_i^2)=\rho_i$, and

$$\mu_1(\beta)=\rho_i\exp\{-(n/\rho_i)[(\alpha_m-\beta)^{1/2}+(\alpha_m-\rho_i^2)^{1/2}]\}\quad\text{for } \rho_i^2\leq\beta\leq\alpha_m. \tag{8}$$

The expressions (6)—(8) can be simplified if we define

$$x_1(y)=\mu_1(\beta)/\rho_i,\quad x_2(y)=\mu_2(\beta)/\rho_i$$

with $y=\beta/\rho_i^2$ and $y_m=\alpha_m/\rho_i^2$. Then

$$x_1(y)=\sqrt{y}\left(\frac{1-(1-y)^{1/2}}{1+(1-y)^{1/2}}\right)^n\exp\{-2n[(y_m-y)^{1/2}-(1-y)^{1/2}]\}\quad\text{for } 0\leq y\leq 1, \tag{9}$$

$$x_1(y)=\exp\{-n[(y_m-y)^{1/2}+(y_m-1)^{1/2}]\}\quad\text{for } 1\leq y\leq y_m, \tag{10}$$

$$x_2(y)=\exp\{n[(y_m-y)^{1/2}-(y_m-1)^{1/2}]\}\quad\text{for } 1\leq y\leq y_m. \tag{11}$$

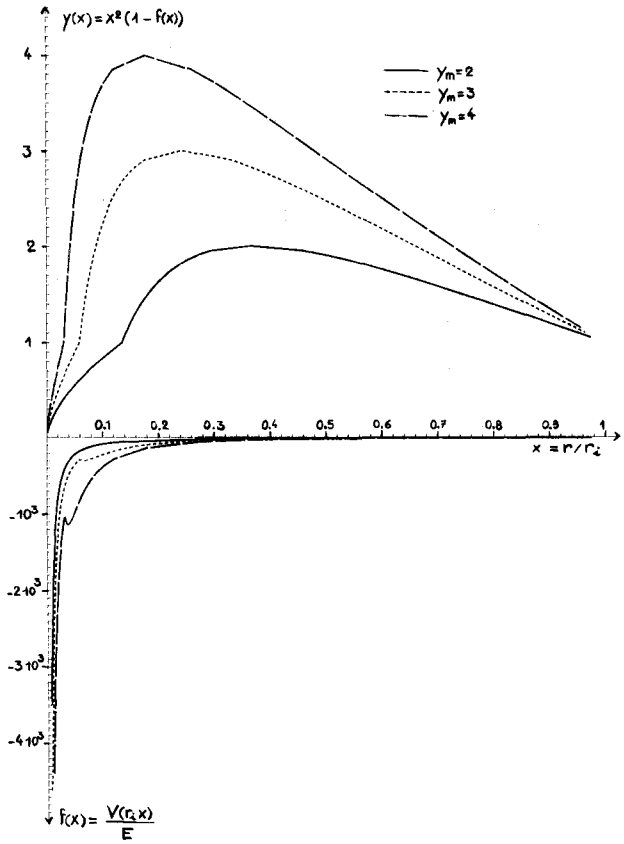


FIG. 2. Some transparent potentials for $n=1$: $V(r)=0$ for $r \geq r_i$ (r_i arbitrary); E is the energy; $y_m = \max_{0 \leq x \leq 1} y(x)$ with $y(x) = x^2[1 - V(r_i x)/E]$

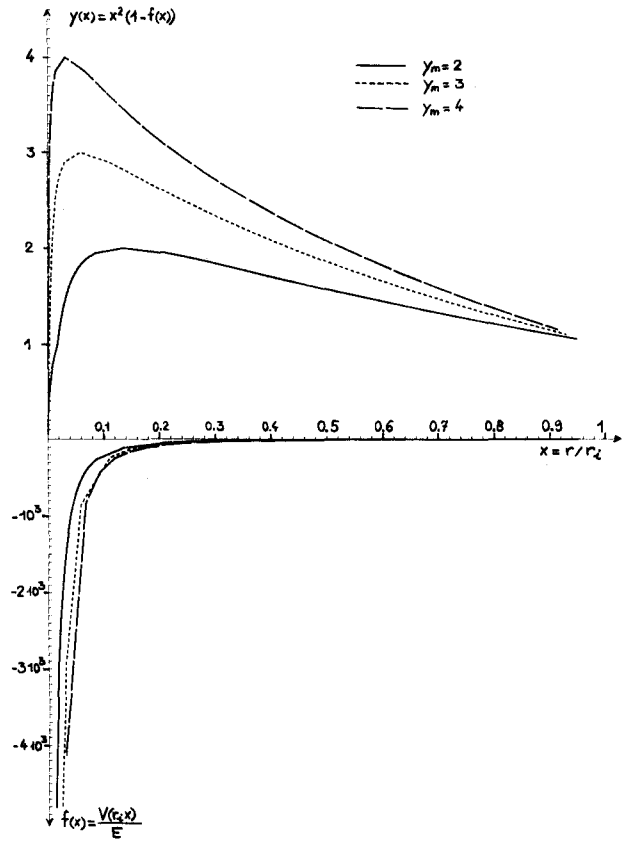


FIG. 3. Some transparent potentials for $n=2$.

For each positive integer n and each number $y_m > 1$, (9)—(11) define a function $x \to y_{n,y_m}(x)$ such that the inverse function of $y_{n,y_m}(x)$ is $x_1(y)$ for $0 \leq x \leq x_1(y_m)$ and $x_2(y)$ for $x_1(y_m) \leq x \leq 1$.

If $f_{n,y_m}(x) = 1 - y_{n,y_m}(x)/x^2$, the transparent potential is

$$V_{n,y_m,r_i}(r) = E f_{n,y_m}(r/r_i), \quad \text{where } r_i = \rho_i/k.$$

Figures 2 and 3 give the result for $n=1$ and 2 and $y_m = 2, 3, 4$.

Let us remark that the formula (9) shows that

$$V_{n,y_m,r_i}(r) = O(r^{-2+1/n+1/2}) \text{ when } r \to 0.$$

These transparent potentials are not singular.

Note added in proof: The previous analysis can also be made at *fixed angular momentum*, because the angle $\varphi(E) = \hbar(l+\frac{1}{2})(2m)^{-1/2} \int_{r(E)}^{\infty} \{E - V(r) - \hbar^2(l+\frac{1}{2})^2/2mr^2\}^{-1/2} dr/r^2$ can be considered as a function of E at fixed l [$r(E)$ being the greatest solution of $V(r) + \hbar^2(l+\frac{1}{2})^2/2mr^2 = E$). So given a number $r_i > 0$ and a positive integer n, we shall see that it is possible to construct functions $W(r) = V(r) + \hbar^2(l+\frac{1}{2})^2/2mr^2$ equal to $\hbar^2(l+\frac{1}{2})^2/2mr^2$ for $r \geq r_i$, with one minimum $E_m = W(r_m) < E_i = \hbar^2(l+\frac{1}{2})^2/2mr_i^2$ in the interval $]0, r_i[$ and such that $\varphi(E) = \pi/2 + n\pi$ for $E \geq E_i$ (*Condition* 1′). Because $W(r)$ is monotone in the intervals $]0, r_m[$ and $[r_m, r_i]$ we can introduce $r_1(\epsilon)$, the inverse function of $W(r)$ for $0 < r \leq r_m$, and $r_2(\epsilon)$, the inverse function of $W(r)$ for $r_m \leq r \leq r_i$. These two functions must be respectively decreasing and increasing. Setting $r_2(\epsilon) = r_i$ for $\epsilon \geq E_i$, Condition (1′) is equivalent to

$$\hbar(2+\tfrac{1}{2})(2m)^{-1/2} \int_{E_m}^{E} (E-\epsilon)^{-1/2} d\left(\frac{1}{r_1(\epsilon)} - \frac{1}{r_2(\epsilon)}\right)$$

$$= \int_{E_i}^{E} (E-\epsilon)^{-1/2} d\epsilon^{1/2} + n\pi$$

for $E \geq E_i$ (*Condition* 2′). Omitting details similar to the ones of the previous analysis, we can show that if $r_2(E_i) = r_i$, $r_2(E_m) = r_1(E_m)$, if the function $\epsilon \to 1/r_1(\epsilon) - 1/r_2(\epsilon)$ is increasing for $E_m \leq \epsilon \leq E_i$ and if

$$\hbar(l+\tfrac{1}{2})(2m)^{-1/2} \int_{E_m}^{E_i} (E_i - \epsilon)^{-1/2} d\left(\frac{1}{r_1(\epsilon)} - \frac{1}{r_2(\epsilon)}\right) = n\pi,$$

then a potential which satisfies Condition 1′ is defined, for $r \leq r_1(E_i)$, by

$$\frac{1}{r_1(E)} - \frac{1}{r_i}$$

$$= \frac{(2m)^{1/2}}{\hbar(l+\frac{1}{2})} \{\sqrt{E} - \sqrt{E_i} + 2n(E-E_i)^{1/2}\}$$

$$+ \frac{2}{\pi} \int_{E_m}^{E_i} \arcsin \frac{(E_i - \epsilon)^{1/2}}{(E-\epsilon)^{1/2}} \, d\left(\frac{1}{r_1(\epsilon)} - \frac{1}{r_2(\epsilon)}\right).$$

Explicit examples are easily constructed by choosing

$$-\frac{r_1'(E)}{r_1^2(E)} = \frac{r_2'(E)}{r_2^2(E)} = \frac{(2m)^{1/2}}{\hbar(l+\frac{1}{2})} \frac{n}{(E-E_m)^{1/2}}$$

for $E_m \leq E \leq E_i$. It is possible to show that these transparent potentials are singular and satisfy

$$V(r) = O\left[\left(\frac{1}{(1+2n)^2} - 1\right)\frac{1}{r^2}\right] \text{ when } r \to 0.$$

Another remark must be made about the physical meaning of these results. The quantity

$$\Delta\varphi(E)=\hbar(l+\tfrac{1}{2})(2m)^{-1/2}\int_{E_m}^{E}(E-\epsilon)^{-1/2}\,d\left(\frac{1}{r_1(\epsilon)}-\frac{1}{r_2(\epsilon)}\right)$$

$$=\hbar(l+\tfrac{1}{2})(2m)^{-1/2}\int_{r_1(E)}^{r_2(E)}\left(E-V(r)-\frac{\hbar^2(l+\frac{1}{2})^2}{2mr^2}\right)^{-1/2}\frac{dr}{r^2}$$

which appears in the first member of Condition (2′) (and also 2), is, for $E>E_i$, the deflection angle of the particle when it goes from $r_i=r_2(E)$ to the turning point $r_1(E)$, and for $E<E_i$, it is the angular period of a finite motion in the potential $V(r)$ [between $r_1(E)$ and $r_2(E)$]. The previous results show that if $\varphi(E)$ is known for $E\geq 0$ ("scattering states") and if $\Delta\varphi(E)$ is known for $E_m<E<E_i$ ("bound states") then we can compute $r_1(E)$ for $E\geq E_i$ and $1/r_1(E)-1/r_2(E)$ for $E_m\leq E\leq E_i$. To construct the potential it is necessary to know another characteristic of "bound states" for $E_m\leq E\leq E_i$, and the time period $\Delta T(E)=(m/2)^{1/2}\int_{r_1(E)}^{r_2(E)}[E-V(r)-\hbar^2(l+\frac{1}{2})^2/2mr^2]^{-1/2}\,dr$ [which gives $r_2(E)-r_1(E)$ for $E_m\leq E\leq E_i$] is convenient. So to obtain $V(r)$ it is necessary to know one *function $\varphi(E)$ of "scattering states" and two functions $\Delta\varphi(E)$ and $\Delta T(E)$ of "bound states."*

The author thinks that this analogy could be a reminiscence of the Marchenko[6] result in quantum mechanics (see also Wheeler[7]). Furthermore, the energies E_n of bound states for $E_m\leq E\leq E_i$ are quantized and the JWKB method gives the approximate condition

$$\frac{(2m)^{1/2}}{h}\int_{r_1(E_n)}^{r_2(E_n)}\left(E_n-V(r)-\frac{h^2(l+\frac{1}{2})^2}{2mr^2}\right)^{1/2}dr=(n+\tfrac{1}{2})\pi.$$

So the *number of bound states of energy $\leq E$ and of angular momentum $h(l+\frac{1}{2})$* is

$$n(E,l)\simeq\frac{(2m)^{1/2}}{\hbar\pi}\int_{r_1(E)}^{r_2(E)}\left(E-V(r)-\frac{\hbar^2(l+\frac{1}{2})^2}{2mr^2}\right)^{1/2}dr$$

and the functions $\Delta\varphi(E)\simeq(\partial/\partial l)\{\pi n(E,l)\}$ and $\Delta T(E)\simeq\hbar(\partial/\partial E)\{\pi n(E,l)\}$ are two characteristics of the discrete spectrum of the potential $V(r)$.

ACKNOWLEDGMENTS

I would like to thank Professor P.C. Sabatier, who suggested this problem.

*This work has been done as a part of the program of the "Recherche coopérative sur Programme $n°$ 264, Etude interdisciplinaire des problèmes inverses."

†Physique mathematique et théorique, Equipe de recherche associée au CNRS $n°$ 154.

[1]U. Buck, Rev. Mod. Phys. **46**, 369—89 (1974).
[2]P.C. Sabatier, J. Math. Phys. **13**, 675—99 (1972).
[3]P.C. Sabatier, Phys. Rev. A **8** 589—601 (1973).
[4]P.C. Sabatier, J. Math. Phys. **7**, 8, 1515—1521 (1966).
[5]K. Chadan, Nuovo Cimento A **47**, 510—25 (1967).
[6]V.A. Marchenko, Dokl. Akad. Nauk SSSR **104**, 695—8 (1955); Math. Rev. **17**, 7, 40 (1956).
[7]J.A. Wheeler, in *Studies in Mathematical Physics, Essays in Honor of Valentine Bargman*, edited by E.H. Lieb, S. Simon, and A.S. Wightman (Princeton, U.P., Princeton, N.J., 1976), pp. 351—422.

Free electromagnetic fields on a compact Lie group manifold

Leighton L. Henry

Department of Mathematics, University of the West Indies, Kingston 7, Jamaica, West Indies
(Received 30 June 1976)

Maxwell's equations for free fields are studied on the underlying C^∞ manifold of a compact Lie group. The formulation is in terms of exterior differential forms as given by Wheeler. It is found that on a compact connected Lie group there are no free electromagnetic fields. The results obtained are essentially a physical interpretation of the well-known theorem that the second Betti number of a compact semisimple Lie group is zero.

1. INTRODUCTION

The advents of the theories of relativity and quantum mechanics have led recently to far-reaching attempts to develop global formulations of physical theories. The process of passing from a local to a global formulation is not at the moment well defined. Experience has, however, shown that those aspects of a local theory that can be formulated in terms of exterior differential forms can be readily globalized, by applying the theorems of Stokes and deRham. Ohkuro[1] and Wheeler[2] have shown how the equations of Maxwell may be written as equations in differential form. In the case of Ref. 2 the equations on a Minkowski space are generalized to equations on a C^∞ differentiable manifold.

Let M be a compact C^∞ manifold with local coordinates (x^μ) and let $f=f_{(\mu\nu)}dx^\mu\wedge dx^\nu$ be a 2-form on M. Here, in Minkowski space, $f_{\mu\nu}$ is the usual electromagnetic field tensor. Wheeler[2] defines Maxwell's equations in free M space as

$$df=0,\quad \delta f=0. \tag{1.1}$$

Thus f is an element of the second cohomology group of M. [Here d and δ are respectively the exterior derivative and codifferential operators.]

The motivation for Eqs. (1.1) is that if M is the usual Minkowski space, these equations reduce to the normal Maxwell's free space equations. The 2-form f is therefore Hodge harmonic and since M is compact f is also Kodaira harmonic, and satisfies

$$\Delta f=0=(\delta d+d\delta)f. \tag{1.2}$$

If the condition that f is co-closed (i.e., $\delta f=0$) is not imposed, certain consequences result. Since M is compact and f is closed, then by Poincaré's lemma: there exists locally a 1-form a such that

$$da=f.$$

Physically, this 1-form corresponds to a vector potential for the electromagnetic field.

Wheeler[2] introduces electric and magnetic charges as follows: He defines

$$\begin{aligned}&\int_{C_2} f=4\pi \text{ (magnetic charges on } C_2),\\ &\int_{C2} {}^*f=4\pi \text{ (electric charge on } C_2).\end{aligned} \tag{1.3}$$

Here C_2 is a 2-chain in the manifold M, and * is the Hodge duality map. These definitions are physically motivated, since if C_2 is any 2-cycle in M (i.e., $\partial C_2=0$) then by Stokes' theorem,

$$\int_{C_2} f=\int_{C_2} da=\int_{\partial C_2} a=0. \tag{1.4}$$

The physical interpretation of Eq. (1.4) is that there are no free magnetic charges on M.

Suppose now, $C_2^{(1)}$ and $C_2^{(2)}$ are any homologous 2-cycles on M (i.e., there exists a chain C such that $C_2^{(1)}-C_2^{(2)}=\partial C$). Then by Stokes' theorem, we have

$$\int_{\partial C} f=\int_C df=0=\int_{C_2^{(1)}} f-\int_{C_2^{(2)}} f. \tag{1.5}$$

Hence, the flux out of any closed surface is independent of the size of the surface (Gauss' law).

Wheeler[2] discusses the existence of charged solutions of Maxwell's equations on M, using deRham's first and second theorems.

deRham's first theorem[3]: Let B_p be the pth Betti number of M.

If C_p are B_p-independent p-cycles on a compact manifold M and (q_p) are any B_p-real numbers, then there exists a p-form α which is closed and differentiable throughout M and which satisfies

$$\int_{C_p} \alpha=Q_p.$$

This theorem provides the interpretation that there exists a solution of $df=0$ or $d^*f=0$ which has arbitrary assigned magnetic or electric charges.

deRham's second theorem[3]: A closed form having zero periods is an exact form.

This means that the requirement that there are no free magnetic charges guarantees the existence of a vector potential. We therefore get the stronger result that the existence of charges is a function of the topology of M.

2. FREE ELECTROMAGNETIC FIELDS ON A LIE GROUP MANIFOLD

We now proceed to study the free field case where M is the underlying space of a compact orientable connected Lie group.

Let G be a compact connected semisimple Lie group with Lie algebra g, which has as underlying vector space

Copyright © 1977 American Institute of Physics

the tangent space at the identity in G. Further, let $\{X_\alpha\}$ be a basis for g with $\{\omega^\alpha\}$ the corresponding basis for the Cartan–Maurer forms of g. Then $\{\omega^\alpha\}$ satisfies the Cartan–Maurer relations[4]

$$d\omega^\alpha = \tfrac{1}{2} C^\alpha_{\beta\gamma}\omega^\beta \wedge \omega^\gamma , \tag{2.1}$$

where $C^\alpha_{\beta\gamma}$ are the structure constants of g. The Maxwell 2-form is now defined by

$$F = F_{\alpha\beta}\omega^\alpha \wedge \omega^\beta . \tag{2.2}$$

If we now impose the requirement that F is closed, we have

$$\begin{aligned} dF = 0 = (\Gamma_{\alpha\beta})_\gamma \omega^\gamma \wedge \omega^\alpha \wedge \omega^\beta \\ + F_{\alpha\beta} d\omega^\alpha \wedge \omega^\beta - F_{\alpha\beta}\omega^\alpha \wedge d\omega^\beta , \end{aligned} \tag{2.3}$$

where

$$dF_{\alpha\beta} = (\Gamma_{\alpha\beta})_\gamma \omega^\gamma .$$

Using the Cartan–Maurer relations, and changing dummy indices, we get

$$[(\Gamma_{\alpha\beta})_\gamma - \tfrac{1}{2} F_{\alpha\rho} C^\rho_{\gamma\beta} + \tfrac{1}{2} F_{\delta\beta} C^\delta_{\gamma\alpha}]\omega^\gamma \wedge \omega^\alpha \wedge \omega^\beta . \tag{2.4}$$

The independence of the set $(\omega^\gamma \wedge \omega^{\alpha\wedge\beta})$ leads to half of Maxwell's equations on G, namely

$$\begin{aligned} &[(\Gamma_{\alpha\beta})_\gamma + (\Gamma_{\gamma\alpha})_\beta + (\Gamma_{\beta\gamma})_\alpha] \\ &\quad - \tfrac{1}{2}[F_{\alpha\rho} C^\rho_{\gamma\beta} + F_{\gamma\rho} C^\rho_{\beta\alpha} + F_{\beta\rho} C^\rho_{\alpha\gamma}] \\ &\quad + \tfrac{1}{2}[F_{\delta\beta} C^\delta_{\gamma\alpha} + F_{\delta\alpha} C^\delta_{\beta\gamma} + F_{\delta\gamma} C^\delta_{\alpha\gamma}] = 0. \end{aligned} \tag{2.5}$$

The second half of the Maxwell's equations is obtained by the requirement that F is co-closed. We will not, however, proceed to determine these equations. Since F is closed and co-closed, it is harmonic on G. Now it is known[5,6] that the harmonic forms on a compact or orientable Riemann manifold are g-invariant differential forms, where g is the Lie algebra of the infinitesimal motions on G. It is also well known that a Riemann structure can be put on G using the structure constants to define a metric tensor via

$$g_{\alpha\beta} = C^\rho_{\alpha\tau} C^\tau_{\rho\beta} . \tag{2.6}$$

Since F is harmonic and therefore g-invariant, taking the Lie derivative with respect to the generator X_α, we get

$$0 = L_{X\alpha} F = (F_{\tau\beta} C^\tau_{\alpha\gamma} + F_{\alpha\tau} C^\tau_{\beta\gamma})\omega^\beta \wedge \omega^\gamma \tag{2.7}$$

which then implies

$$F_{\tau\beta} C^\tau_{\alpha\gamma} + F_{\alpha\tau} C^\tau_{\beta\gamma} = 0. \tag{2.8}$$

We now proceed to solve Eq. (2.8) using a technique of Goldberg.[3] Permuting the indices α, β, and γ cyclically and adding, we get

$$F_{\tau\beta} C^\tau_{\alpha\gamma} + F_{\tau\alpha} C^\tau_{\gamma\beta} + F_{\tau\gamma} C^\tau_{\beta\alpha} = 0. \tag{2.9}$$

Subtracting (2.8) from (2.9) and using the antisymmetry of $F_{\tau\alpha}$ and the structure constants, we get

$$F_{\tau\gamma} C^\tau_{\beta\alpha} = 0. \tag{2.10}$$

The independence of the structure constants [or use of Eq. (2.6)] shows that

$$F_{\alpha\beta} \equiv 0.$$

We therefore conclude that there are no free electromagnetic fields on a compact connected Lie group manifold.

3. COMMENTS

Here the result obtained is merely a physical interpretation of the classical result that the second Betti number of a compact semisimple Lie group is identically zero.[3] Starting with this latter result, one could readily deduce the result that $F_{\alpha\beta} \equiv 0$ since for a compact orientable Riemann manifold the number of linearly independent p-harmonic forms is equal to the pth Betti number of the manifold. The second Betti number being zero for a compact semisimple Lie group, it follows that there are no harmonic 2-forms on G.

[1] S. Ohkuro, J. Math. Phys. **11**, 2005 (1970).
[2] J. A. Wheeler *et al.*, Ann. Phys. (N.Y.) **2**, 525 (1957).
[3] S. Goldberg, *Curvature and Homology* (Academic, New York, (1962).
[4] P. M. Cohn, *Lie Groups* (Cambridge U. P., Cambridge, (1957).
[5] S. Goldberg, Bull. Am. Math. Soc. **66**, 54 (1960).
[6] K. Yano, Ann. Math. **55**, 38 (1952).

A class of solutions of the Dirac equation in the Kerr–Newman space

S. Einstein and R. Finkelstein

Department of Physics, University of California, Los Angeles, California 90024
(Received 1 November 1976)

In a region of spacetime that may be described by the Kerr–Schild metric, the gravitational field equations define a field of O(3) matrices. By examining the spin representations of these rotations it is first shown how the gravitational field equations define a spinor field, and it is then shown how this spinor field is related to special solutions of the massless Dirac equation in the Kerr–Newman space. These special solutions have arbitrary angular momentum about the axis of rotation and in the classical limit correspond to orbits that coincide with the principal null congruences.

1. INTRODUCTION

The Dirac equation holding in the Kerr field of a black hole, and expressed in Boyer—Lindquist coordinates, was separated in the massless case by Teukolsky[1] and in the massive case by Chandrasekhar,[2] but no explicit solution in terms of known functions has been obtained, as far as we know. In this paper we find a special class of massless solutions that are characterized by arbitrary angular momentum about the axis of rotation and by classical limits that correspond to very simple orbits.

Here we work with the Kerr—Schild metric, instead of the Boyer—Lindquist coordinates, and follow the approach introduced by Schiffer *et al.*[3] Their method is based on the observation that in this problem the Einstein equations define a field of rotation matrices. In Refs. 3 and 4 one works with the three-dimensional representation of the rotation group, while here we work with the spin representation.

The introduction to this paper is slightly extended to facilitate the comparison of the spinor and vector representations and also to include in Sec. 3 a simplification and generalization of the discussion of the Kerr—Newman case that was given in Ref. 4. The main part of this paper is devoted to obtaining a special class of spinor fields by two different procedures. In the first method, Secs. 4—6, one arrives at a spinor field via the gravitational equations and without any reference to the Dirac equation. In the second approach, beginning in Sec. 7, we solve the Dirac equation holding in the Kerr—Newman geometry and subject to an additional constraint that is natural in this geometry. The spinor fields obtained in these different ways differ only by a position dependent normalization. The eikonal limit of these wave-functions describes orbits that coincide with the principal null congruences.

Let us first summarize the results of Refs. 3 and 4. We assume a metric tensor of the Kerr—Schild form:

$$g_{\mu\nu}=\eta_{\mu\nu}-2ml_\mu l_\nu, \tag{1.1}$$

where $\eta_{\mu\nu}$ is the Minkowski metric and l_μ describes the null geodesics of massless particles that fall into the rotating attractive mass. The l_μ field describes spiral paths that represent the dragging effect of the rotating mass; this field completely determines the geometry and may itself be obtained from a complex harmonic potential γ or its reciprocal ω, which is a complex radial coordinate:

$$\gamma=\alpha+i\beta, \tag{1.2}$$

$$\omega=\rho+i\sigma=\gamma^{-1}. \tag{1.3}$$

Here the real and imaginary parts of ω play the roles of radial and angular coordinates according to the relations

$$\rho^2=r^2-a^2\sin^2\theta, \tag{1.4a}$$

$$\sigma=-a\cos\theta, \tag{1.4b}$$

where (r,θ) become the Schwarzschild polar coordinates when the specific angular momentum a vanishes.

The fundamental vector field l_μ may be expressed entirely in terms of the single complex variable γ by the following equations:

$$l_\mu=(l_0,l_0\lambda_k), \tag{1.5a}$$

$$\lambda_k\lambda_k=1, \tag{1.5b}$$

$$l_0^2=(1-Q^2/2m\rho)\,\alpha(\rho,\sigma), \tag{1.6a}$$

$$\lambda=(\nabla\omega+\nabla\omega^*-i\nabla\omega\times\nabla\omega^*)/(1+\nabla\omega\nabla\omega^*). \tag{1.6b}$$

Here Q and m are the charge and mass of the source. The specific angular momentum a enters only into the relations (1.4) that connect (ρ,σ) with (r,θ).

The electric and magnetic fields are also entirely determined by γ as follows[4]:

$$\mathbf{E}+i\mathbf{H}=-Q\nabla\gamma. \tag{1.7}$$

These simple results are perhaps most easily approached by observing that the Einstein field equations define a field of three-dimensional rotations, provided that the geometry may be described by the metric (1.1). In this paper we shall pursue the same ideas but shall work with the spin instead of the three-dimensional representation of the rotation group.

2. THREE-DIMENSIONAL REPRESENTATION OF FIELD EQUATIONS

In the Kerr—Schild metric (1.1) the field equations everywhere outside of the rotating source imply

$$\Box(l_\alpha l_\beta)+\partial_\alpha j_\beta+\partial_\beta j_\alpha=0, \tag{2.1}$$

Copyright © 1977 American Institute of Physics

where

$$j_\alpha = (A+B)l_\alpha \tag{2.2}$$

and the invariants A and B are defined as follows:

$$l^\sigma \partial_\sigma l_\mu = -Al_\mu, \tag{2.3}$$

$$\partial_\alpha l^\alpha = -B. \tag{2.4}$$

Since l_α is null, it follows from (2.1) that j_α is conserved:

$$\partial_\alpha j^\alpha = 0. \tag{2.5}$$

The complete field is characterized by these two scalar functions A and B.

We are here interested in time-independent fields. Since the problem then becomes three-dimensional, let us now go over to the three-dimensional notation (1.5).

The structure of the problem is determined mainly by the jk equations, that follow from (2.1), namely,

$$\partial_n \lambda_j \partial_n \lambda_k = p[\partial_j \lambda_k + \partial_k \lambda_j], \tag{2.6}$$

where p is determined by A and B. Let

$$M_{ij} = \partial_j \lambda_i \tag{2.7}$$

and

$$\mathcal{R}_{ij} = 1 - (1/p)M_{ij}. \tag{2.8}$$

Then (2.6) becomes

$$\mathcal{R}\mathcal{R}^T = 1 \tag{2.9}$$

while

$$\mathcal{R}\boldsymbol{\lambda} = \boldsymbol{\lambda}. \tag{2.10}$$

Therefore, $\mathcal{R}$ is a rotation matrix with axis $\boldsymbol{\lambda}$. When $\partial_j \lambda_i$ assumes all values permitted by (2.6), $\mathcal{R}$ runs over all rotations. There is a $\boldsymbol{\lambda}$ and a ψ such that every matrix of the three-dimensional representation of the rotation group may be written in the following form:

$$\mathcal{R} = \exp(\psi\Lambda), \tag{2.11}$$

where

$$\Lambda_{ij} = \epsilon_{ijk}\lambda_k. \tag{2.12}$$

If one expands the exponential one encounters

$$P = -\Lambda^2 \tag{2.13a}$$

with components

$$P_{ij} = \delta_{ij} - \lambda_i \lambda_j. \tag{2.13b}$$

In terms of P and Λ one finds

$$\mathcal{R} = 1 - P + P\cos\psi + \Lambda\sin\psi \tag{2.11'}$$

and therefore

$$M = \alpha P + \beta\Lambda \tag{2.14a}$$

or

$$\partial_i \lambda_j = \alpha P_{ji} + \beta\Lambda_{ji}, \tag{2.14b}$$

where α and β are simple functions of p and ψ. The transformation of (2.6) into (2.14) has linearized these differential equations in $\partial_i \lambda_j$. It is now useful to rewrite (2.14) as follows[4]:

$$M = \mathrm{Re}\gamma/\mathcal{H}, \tag{2.15}$$

where γ is the harmonic function introduced in (1.2), and

$$\mathcal{H} = P - i\Lambda. \tag{2.16}$$

Notice that the Hermitian matrix $\mathcal{H}$ has the following properties:

$$\mathcal{H}\boldsymbol{\lambda} = 0, \tag{2.17a}$$

$$\mathcal{H}(-\boldsymbol{\lambda}) = \mathcal{H}^*(\boldsymbol{\lambda}), \tag{2.17b}$$

$$\mathcal{H}(\boldsymbol{\lambda})\mathcal{H}(-\boldsymbol{\lambda}) = 0, \tag{2.18a}$$

$$\mathcal{H}^2 = 2\mathcal{H}. \tag{2.18b}$$

The integrability conditions holding for the set (2.14) may be used to invert these equations so that one obtains the following differential equations for γ in terms of $\boldsymbol{\lambda}$ [4]:

$$\lambda_s \partial_s \gamma + \gamma^2 = 0, \tag{2.19}$$

$$\mathcal{H}_{sk}\partial_k \gamma = 0, \tag{2.20}$$

$$\nabla^2 \gamma = 0. \tag{2.21}$$

If Laplace's equation is solved for the complex γ one may find λ_s according to (1.6b). Note that the primary equation (2.20) and all subsequent equations derived from it are invariant under the transformation

$$\boldsymbol{\lambda} \to -\boldsymbol{\lambda}, \tag{2.22a}$$

$$\gamma \to -\gamma^* \tag{2.22b}$$

in virtue of (2.17b).

Equation (2.19) may also be written in the form

$$\lambda_s \partial_s \omega = 1. \tag{2.23}$$

Finally we need the connection between the coordinates (x,y,z), in which (1.1) holds, and the complex distance ω; and also between (x,y,z) and a system of polar coordinates (ρ,θ,ϕ):

$$\omega^2 = x^2 + y^2 + (z - ia)^2, \tag{2.24}$$

$$x + iy = (\rho + ia)\sin\theta\exp(i\phi), \tag{2.25a}$$

$$z = \rho\cos\theta, \tag{2.25b}$$

where the axis of rotation is in the z direction. From these relations (1.4) follows.

In polar coordinates the fundamental vector field is by (1.6b)

$$\boldsymbol{\lambda} = (\sin\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\theta). \tag{2.26}$$

Note also the following relations:

$$\lambda_s \partial_s \rho = 1, \tag{2.27a}$$

$$\lambda_s \partial_s \sigma = 0 \tag{2.27b}$$

that follow from (2.23). In addition one has

$$\lambda_s \partial_s \phi = 0 \tag{2.27c}$$

and therefore the important relation

$$\lambda_s \partial_s = \frac{\partial}{\partial\rho}. \tag{2.28}$$

3. THE TRACELESS ENERGY-MOMENTUM TENSOR

The preceding section outlines the derivation of $\boldsymbol{\lambda}$ followed in Refs. 3 and 4 for the case that the energy—

momentum tensor vanishes outside of the rotating mass, We now consider the more general case of an energy-momentum tensor that does not vanish but is still restricted by the conditions that it has no trace, that the Kerr–Schild metric still holds, and that the vector field λ is unchanged. This new energy–momentum tensor is then able to influence the geometry only through l_0. We shall now give a new derivation of l_0 that is valid for the charged, rotating source, as well as for any other traceless energy–momentum tensor subject to the limitations just mentioned.

We now have

$$g^{\mu\lambda}R_{\mu\lambda}=g^{\mu\lambda}\theta_{\mu\lambda}=0. \tag{3.1}$$

By (2.1) the divergence of j_μ vanishes. If the source is traceless, the contracted form of (2.1) remains valid and we still have (2.5), and in the static case

$$\partial^k j_k=0. \tag{3.2}$$

We look for a solution that is as close as possible to the one already found for the case of vanishing energy–momentum tensor. Therefore, we assume the Kerr–Schild form again, where l_μ is still given by (1.5a) and the λ_k are the same as before; but we now put

$$l_0^2=\psi(\rho,\sigma)\alpha, \tag{3.3}$$

where $\psi(\rho,\sigma)$ is a new function to be determined by (3.2). Let us now compute

$$j_k=(A+B)\,l_0\lambda_k. \tag{3.4}$$

By (2.14), (2.19), and (3.3) one finds

$$j_k=\lambda_k(\gamma\gamma^*)\lambda_s\partial_s(\psi\rho). \tag{3.5}$$

The divergence of the preceding current vanishes by (3.2):

$$\partial_k j_k=(\gamma\gamma^*)(\lambda_i\partial_i)^2(\psi\rho)=0 \tag{3.6}$$

or

$$\frac{\partial^2}{\partial\rho^2}(\psi\rho)=0, \tag{3.7}$$

where we have made use of (2.28). The solution of (3.7) is

$$\psi=C(\sigma)+D(\sigma)/\rho. \tag{3.8}$$

Then also, by (3.5),

$$j_k=C(\sigma)\gamma\gamma^*\lambda_k. \tag{3.9}$$

In the example of the charged source we determine these integration constants C and D by comparing with the Reissner–Nordstrom solution at large distances. Then

$$\psi=1-Q^2/2m\rho. \tag{3.10}$$

If Q vanishes, we obtain $l_0^2=\alpha$ for the uncharged source. In the following we need not specialize to the charged spinning source, but we do assume that the Kerr–Schild null vector is given by (1.5a), (3.3), and (3.8). As already remarked λ_k is still given by (1.6b). The preceding argument is a simplification and generalization of the discussion in Ref. 4.

4. TWO-DIMENSIONAL REPRESENTATION OF ROTATIONS

Just as (2.11) has been written for the three-dimensional representation of rotations, it is also possible to introduce the spin representation of the same equation. The subsequent discussion may be carried through the same as before, but the computations are now a little simpler. We shall show for example how (2.19)–(2.21) may be obtained with the aid of the spin representation.

Instead of (2.11) we now have

$$\overset{(2)}{R}=\exp\left(\tfrac{1}{2}\overset{(2)}{\Lambda}\psi\right), \tag{4.1}$$

where

$$\overset{(2)}{\Lambda}=i\sigma\lambda. \tag{4.2}$$

Then (2.13) becomes

$$\overset{(2)}{P}=-\overset{(2)}{\Lambda}{}^2=+1 \tag{4.3}$$

and by (2.11a), or by expanding (4.1),

$$\overset{(2)}{R}=\cos(\psi/2)+\overset{(2)}{\Lambda}\sin(\psi/2). \tag{4.4}$$

Finally

$$\overset{(2)}{M}=\overset{(2)}{P}-i\overset{(2)}{\Lambda}=-1+\sigma\lambda. \tag{4.5}$$

The two-dimensional Hermitian matrix $\overset{(2)}{M}$ satisfies relations analogous to (2.17)–(2.18). Instead of (2.17) one has

$$\overset{(2)}{M}(\lambda)V(\lambda)=0, \tag{4.6}$$

where $V(\lambda)$ is a two-component spinor and in place of (2.17b) one has

$$C\overset{(2)}{M}(\lambda)^*C^{-1}=\overset{(2)}{M}(-\lambda), \tag{4.7}$$

where

$$C\sigma^*C^{-1}=-\sigma. \tag{4.8}$$

With the usual Pauli representation,

$$C=\sigma_y. \tag{4.9}$$

Let us parametrize λ according to (2.26). Then

$$\sigma\lambda=\begin{pmatrix}\cos\theta & \sin\theta\exp(-i\phi)\\ \sin\theta\exp(i\phi) & -\cos\theta\end{pmatrix} \tag{4.10}$$

and according to (4.6)

$$\sigma\lambda\cdot V(\lambda)=V(\lambda) \tag{4.11}$$

and

$$V(\lambda)=F(\rho,\sigma,\phi)\begin{pmatrix}\cos(\theta/2)\exp(-i\phi/2)\\ \sin(\theta/2)\exp(i\phi/2)\end{pmatrix} \tag{4.12}$$

while

$$\overset{(2)}{M}(-\lambda)V(-\lambda)=0, \tag{4.13}$$

where

$$V(-\lambda)=CV(\lambda)^*=\sigma_y V(\lambda)^* \tag{4.14}$$

by (4.7), or

$$V(-\lambda)=F(\rho,\sigma,\phi)^*\begin{pmatrix}-i\sin(\theta/2)\exp(-i\phi/2)\\ i\cos(\theta/2)\exp(i\phi/2)\end{pmatrix}. \tag{4.14'}$$

In addition

$$\pm\lambda=V^\dagger(\pm\lambda)\boldsymbol{\sigma}V(\pm\lambda), \tag{4.15}$$

$$1=V^\dagger(\pm\lambda)V(\pm\lambda) \tag{4.16}$$

if

$$|F(\rho,\sigma,\phi)|^2=1. \tag{4.17}$$

We shall now regard the spinor $V(x)$ as the basic field and shall describe the Kerr—Newman geometry in terms of $V(x)$ instead of the scalar field $\gamma(x)$ and the vector field $\lambda(x)$ that were employed in Refs. 3 and 4.

The $V(x)$ defined in this section has the same relation to $\mathcal{H}^{(2)}$ that λ or $\nabla\gamma$ has to $\mathcal{H}$.

5. DIFFERENTIAL EQUATION FOR SPINOR

Let us now find the differential equation for V when λ is constructed in terms of V according to (4.15). Let

$$\partial_k V=(a\sigma_k+b\lambda_k+\mu_k)V, \tag{5.1a}$$

$$\partial_k V^*=V^*(a\sigma_k+b\lambda_k+\mu_k^*), \tag{5.1b}$$

where a and b are assumed to be real and μ_k is a matrix to be determined. We require by (4.16)

$$\partial_k(V^\dagger V)=0. \tag{5.2}$$

By (5.1), (5.2), and (4.15)

$$(a+b)\lambda_k+\tfrac{1}{2}V^*(\mu_k^*+\mu_k)V=0.$$

Choose

$$a+b=0. \tag{5.3a}$$

Then

$$\mu_k^*=-\mu_k \tag{5.3b}$$

and

$$\partial_k V=[a(\sigma_k-\lambda_k)+\mu_k]V. \tag{5.4}$$

We now impose (2.14) and find

$$\partial_k(V^*\sigma_l V)=2a[\delta_{kl}-\lambda_k\lambda_l]+V^*(\sigma_l,\mu_k)V. \tag{5.5}$$

Therefore, choose

$$a=\tfrac{1}{2}\alpha, \tag{5.6a}$$

$$\mu_k=-\tfrac{1}{2}i\beta\sigma_k+\nu_k, \tag{5.6b}$$

where now ν_k remains to be determined. Then

$$\partial_k\lambda_l=\alpha P_{lk}+\beta\Lambda_{lk}+V^\dagger(\sigma_l,\nu_k)V. \tag{5.7}$$

By comparing (2.14) with (5.7) one concludes that

$$(\sigma_l,\nu_k)=0 \tag{5.8}$$

or ν_k is simply a (pure imaginary) c-number. Then

$$\partial_k V=[\tfrac{1}{2}\alpha(\sigma_k-\lambda_k)-\tfrac{1}{2}i\beta\sigma_k+\nu_k]V$$

or alternatively

$$\partial_k V=[\tfrac{1}{2}\gamma^*(\sigma_k-\lambda_k)+i\zeta_k]V, \tag{5.9a}$$

where ζ_k is also a real c-number.[5] By (4.14)

$$\partial_k V(-\lambda)=[\tfrac{1}{2}(-\gamma)(\sigma_k+\lambda_k)-i\zeta_k]V(-\lambda). \tag{5.9b}$$

Again in going from $V(\lambda)$ to $V(-\lambda)$ one replaces γ^* by $-\gamma$ and λ by $-\lambda$ as in (2.22) and in addition ζ_k becomes $-\zeta_k$. Notice also that

$$\sigma(\nabla-i\boldsymbol{\zeta})V(\lambda)=\gamma^*V(\lambda), \tag{5.10a}$$

$$\boldsymbol{\sigma}(\nabla+i\boldsymbol{\zeta})V(-\lambda)=-\gamma V(-\lambda). \tag{5.10b}$$

Expressed entirely in terms of the spinor field (5.9a) is

$$\partial_k V=[\tfrac{1}{2}\gamma^*(\sigma_k-V^\dagger\sigma_k V)+V^\dagger\partial_k V]V. \tag{5.11}$$

We have now introduced two spinor fields $V(+\lambda)$ and $V(-\lambda)$ from which one can reconstruct the vector fields λ and $-\lambda$. The remaining analysis may now be expressed more simply in terms of the new spinor fields rather than the old vector fields.

6. INTEGRABILITY

We have already tacitly assumed that Eqs. (5.9) are integrable since we are requiring that the $V(\pm\lambda)$ are simply given by (4.12) and (4.14). On the other hand Eqs. (5.9a) are of the form

$$\partial_k V=D_k V, \tag{6.1}$$

$$D_k=\tfrac{1}{2}\gamma^*(\sigma_k-\lambda_k)+i\zeta_k, \tag{6.2}$$

and these equations are not integrable unless

$$\partial_k D_l V=\partial_l D_k V. \tag{6.3}$$

Then

$$V^*D_{kl}V=0 \tag{6.3a}$$

and

$$V^*\sigma_j D_{kl}V=0, \tag{6.3b}$$

where

$$D_{kl}=\partial_k D_l-\partial_l D_k+(D_l,D_k). \tag{6.4}$$

By (6.3a)

$$-\tfrac{1}{2}\gamma^*(\partial_k\lambda_l-\partial_l\lambda_k)+i(\partial_k\zeta_l-\partial_l\zeta_k)+\tfrac{1}{2}i(\gamma^*)^2\epsilon_{lks}\lambda_s=0. \tag{6.5}$$

The real and imaginary parts of this equation are

$$\partial_k\lambda_l-\partial_l\lambda_k=2\beta\epsilon_{lks}\lambda_s \tag{6.6a}$$

and

$$\partial_k\zeta_l-\partial_l\zeta_k=\tfrac{1}{2}\gamma\gamma^*\epsilon_{kls}\lambda_s. \tag{6.6b}$$

Equation (6.6a) is just the antisymmetric part of (2.14) and provides no new information, but (6.6b) shows that ζ cannot be dropped from (5.9).

In computing (6.3b) we find

$$D_{kl}=\tfrac{1}{2}[(\sigma_l-\lambda_l)\partial_k\gamma^*-(\sigma_k-\lambda_k)\partial_l\gamma^*]+\tfrac{1}{2}i(\gamma^*)^2\epsilon_{lks}(\sigma_s-\lambda_s), \tag{6.7}$$

$$V^*\sigma_j D_{jl}V=\tfrac{1}{2}[-(\partial_l+\lambda_l D)\gamma^*+i\epsilon_{jls}(\partial_j\gamma^*)\lambda_s]-(\gamma^*)^2\lambda_l. \tag{6.8}$$

Here

$$D=\lambda_s\partial_s.$$

The integrability conditions then imply, among other relations,

$$\lambda_l V^+ \sigma_j D_{jl} V = 0. \tag{6.9}$$

By (6.8) and (6.9) one now finds

$$D\gamma^* + (\gamma^*)^2 = 0. \tag{6.9'}$$

By (6.9'), (6.3b), and (6.8) one finds

$$\mathcal{H}_{ls}\, \partial_s \gamma = 0. \tag{6.10}$$

Therefore, the fundamental differential equations (2.19) and (2.20) satisfied by λ and by γ now appear directly as integrability conditions on the spinor differential equations. Laplace's equation for γ is not an additional equation but follows from (6.9') and (6.10). In this way the complete algebra may be simply obtained from the spinor formulation of the field equations. It is now natural to investigate the relation between the spinors introduced in this way and the spinors that satisfy Dirac's equation in the same geometry.

7. DIRAC EQUATION

We may write the Dirac equation in the following form:

$$\gamma^\mu(\partial_\mu - \Gamma_\mu)\psi = im\psi \tag{7.1}$$

where the γ_μ are position dependent and satisfy

$$\gamma_\mu\gamma_\lambda + \gamma_\lambda\gamma_\mu = 2g_{\mu\lambda}. \tag{7.2}$$

Here Γ_μ is the spin connection. It is determined implicitly by the vanishing of the covariant derivative of γ_μ as follows:

$$\partial_\lambda\gamma_\mu - \Gamma^\sigma_{\mu\lambda}\gamma_\sigma + (\gamma_\mu, \Gamma_\lambda) = 0. \tag{7.3}$$

The second term represents the ordinary Christoffel contribution to the covariant derivative while the commutator is the contribution of the spin connection.

To realize the Kerr–Schild metric, we assume

$$\gamma_\mu = \eta_\mu + \sqrt{2m}\, l_\mu \eta_5, \tag{7.4}$$

where

$$(\eta_\mu, \eta_\sigma)_+ = 2\eta_{\mu\sigma}, \quad (\eta_\mu, \eta_5)_+ = 0, \quad \eta_5^2 = -1, \tag{7.5}$$

and $\eta_{\mu\sigma}$ is still the Minkowski metric. The η_μ in turn may be expressed in terms of constant matrices η_μ:

$$\eta_\mu = S^{-1}\eta_\mu S, \tag{7.6}$$

where S induces a local Lorentz rotation:

$$S^{-1}\eta_\mu S = l^\sigma_\mu \eta_\sigma = \eta_\mu, \tag{7.7a}$$

$$S^{-1}\eta_5 S = \eta_5 = \eta_5. \tag{7.7b}$$

Then

$$\gamma_\mu = S^{-1}\gamma_\mu S \tag{7.8}$$

where

$$\gamma_\mu = \eta_\mu + \sqrt{2m}\, l_\mu \eta_5. \tag{7.9}$$

In terms of these new γ matrices (7.3) becomes

$$\partial_\lambda\gamma_\mu - \Gamma^\sigma_{\mu\lambda}\gamma_\sigma - (\Gamma_\lambda, \gamma_\mu) = 0, \tag{7.10}$$

where

$$\Gamma_\lambda = S\Gamma_\lambda S^{-1} - S\partial_\lambda S^{-1}. \tag{7.11}$$

Equation (7.1) is invariant under (7.8) and (7.11) if ψ is simultaneously transformed according to

$$\psi = S\psi. \tag{7.12}$$

Therefore, we may first find solutions ψ of (7.1) when γ_μ has the simple position dependence given by (7.9). Then the general solution is given by

$$\boldsymbol{\psi} = S^{-1}\psi. \tag{7.12'}$$

It is first necessary, however, to obtain Γ_λ. To do this, make the ansatz

$$\Gamma_\lambda = \tfrac{1}{2}C^{\alpha\beta}_\lambda(\eta_\alpha, \eta_\beta) + d^\alpha_\lambda \eta_\alpha \eta_5. \tag{7.13}$$

If one now expands (7.10), one finds terms in η_5 as well as η_μ. The coefficients of these terms must vanish separately and therefore:

$$\sqrt{2m}\, l_{\mu\,|\lambda} = -2d^\alpha_\lambda \eta_{\alpha\mu}, \tag{7.14}$$

$$\Gamma^\sigma_{\mu\lambda} + 4C^{\sigma\beta}_\lambda \eta_{\beta\mu} - 2\sqrt{2m}\, d^\sigma_\lambda l_\mu = 0. \tag{7.15}$$

Here (7.14) and (7.15) correspond to η_5 and η_σ separately. We then find

$$d_{\lambda\mu} = -\tfrac{1}{2}\sqrt{2m}(2mAl_\mu l_\lambda + \partial_\lambda l_\mu), \tag{7.16}$$

$$\tfrac{1}{2}C^{\alpha\beta}_\lambda(\eta_\alpha, \eta_\beta) = -\tfrac{1}{2}m[(A+B)l_\lambda + (\eta^\mu\partial_\mu)(l_\lambda\tau) + (\partial_\lambda\tau)\tau], \tag{7.17}$$

where

$$\tau = \eta^\mu l_\mu, \tag{7.18}$$

$$\tau^2 = 0. \tag{7.19}$$

Therefore,

$$\gamma^\lambda\Gamma_\lambda = \tfrac{1}{2}m(3A + B - \eta^\mu\partial_\mu\tau)\tau - \tfrac{1}{2}\sqrt{2m}\,\eta_5\eta^\lambda\partial_\lambda\tau \tag{7.20}$$

and

$$(\eta^\mu + \sqrt{2m}\, l^\mu\eta_5)\,\partial_\mu\psi = [im + \tfrac{1}{2}m(3A+B)\tau - \tfrac{1}{2}m(\eta^\mu\partial_\mu\tau)\tau - \tfrac{1}{2}\sqrt{2m}\,\eta_5\eta^\mu\partial_\mu\tau]\psi. \tag{7.21}$$

This is the general form of the Dirac equation when the Kerr–Schild metric is realized by (7.4). The η^μ are the constant Dirac matrices corresponding to $S = 1$. The general ψ is $S^{-1}\psi$, where S is a local spin transformation induced by a Lorentz rotation.

8. SPECIAL SOLUTIONS OF THE DIRAC EQUATION

Consider the class of solutions for which

$$\tau\psi = 0. \tag{8.1}$$

It will be shown that the wave functions satisfying this condition describe currents that flow along the null geodesics defined by the λ field. Then

$$\eta^\mu l_\mu \psi = 0 \tag{8.1'}$$

or

$$\eta_0 l_0 \psi = \eta_k l_k \psi. \tag{8.1''}$$

Let us choose the representation

$$\eta_k = \begin{pmatrix} 0 & -\sigma_k \\ \sigma_k & 0 \end{pmatrix}, \quad \eta_0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \eta_5 = -i\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{8.2}$$

Then (8.1b) becomes, if we denote these special solutions by χ,

$$\begin{pmatrix} \sigma_k\lambda_k & 0 \\ 0 & -\sigma_k\lambda_k \end{pmatrix}\begin{pmatrix} \chi_+ \\ \chi_- \end{pmatrix} = \begin{pmatrix} \chi_+ \\ \chi_- \end{pmatrix}$$

or

$$\sigma_k\lambda_k\chi_\pm = \pm\chi_\pm. \tag{8.3}$$

The $\chi_\pm$ are two-component spinors with opposite helicities, and they are related by

$$\chi_- = H\sigma_y\chi_+^* \tag{8.3'}$$

as in (4.14) except for a factor (H) that may be position dependent. Then (7.21) becomes

$$(\eta^\mu + \sqrt{2m}\, l^\mu\eta_5)\,\partial_\mu\chi = [-\tfrac{1}{2}\sqrt{2m}\,\eta_5\eta^\mu\partial_\mu\tau + i\mathbf{m}]\chi. \tag{8.4}$$

But the operator appearing on the right side of (8.4) may be further simplified as follows:

$$\begin{aligned}\eta^\mu\,\partial_\mu\tau &= \eta^\mu\eta^\sigma\,\partial_\mu l_\sigma \\ &= [\tfrac{1}{2}(\eta^\mu,\eta^\sigma)_+ + \tfrac{1}{2}(\eta^\mu,\eta^\sigma)_-]\,\partial_\mu l_\sigma \\ &= -B + \eta^\mu\eta^\sigma\tfrac{1}{2}(\partial_\mu l_\sigma - \partial_\sigma l_\mu) \\ &= -B + l_0\eta^j\eta^k\tfrac{1}{2}(\partial_j\lambda_k - \partial_k\lambda_j) \\ &\quad - \tfrac{1}{2}\eta^\mu\eta^\sigma(l_\mu\partial_\sigma - l_\sigma\partial_\mu)\ln l_0.\end{aligned}$$

The last term is

$$\begin{aligned}\tfrac{1}{2}\eta^\mu\eta^\sigma(l_\mu\partial_\sigma - l_\sigma\partial_\mu)\ln l_0 &= \tfrac{1}{2}(\eta^\mu,\eta^\sigma)l_\mu\partial_\sigma\ln l_0 \\ &= (\eta^{\mu\sigma} - \eta^\sigma\eta^\mu)(l_\mu\partial_\sigma)\ln l_0 \\ &= (l^\sigma\partial_\sigma - \eta^\sigma\tau\partial_\sigma)\ln l_0 \\ &= -A - \eta^\sigma\tau\partial_\sigma(\ln l_0).\end{aligned} \tag{8.5}$$

Therefore,

$$\eta^\mu\,\partial_\mu\tau\cdot\chi = [A - B + l_0\eta^j\eta^k\tfrac{1}{2}(\partial_j\lambda_k - \partial_k\lambda_j)]\chi$$

or

$$\eta^\mu\partial_\mu\tau\circ\chi = -2l_0[\alpha + i\beta\Sigma^k\lambda_k]\chi, \tag{8.6}$$

where

$$\Sigma^k = \begin{pmatrix} \sigma^k & 0 \\ 0 & \sigma^k \end{pmatrix} = \begin{pmatrix} -\sigma_k & 0 \\ 0 & -\sigma_k \end{pmatrix}. \tag{8.6'}$$

The reduced Dirac equation is then

$$(\eta^\mu + \sqrt{2m}l^\mu\eta_5)\,\partial_\mu\chi = [i\mathbf{m} + \sqrt{2m}\,l_0\eta_5(\alpha + i\beta\Sigma^k\lambda_k)]\chi. \tag{8.7}$$

Multiplying by η^0 and writing in terms of two component spinors, one finds

$$(\partial_0 - \sigma_k\,\partial_k)\chi_+ = [i\mathbf{m} + i\sqrt{2m}l_0(-\lambda^\mu\partial_\mu + \gamma)]\chi_-, \tag{8.8+}$$

$$(\partial_0 + \sigma_k\partial_k)\chi_- = [i\mathbf{m} + i\sqrt{2m}\,l_0(\lambda^\mu\partial_\mu - \gamma^*)]\chi_+, \tag{8.8-}$$

where $\chi_\pm$ must, of course satisfy (8.3) as well. The notation $\lambda^\mu = l^\mu/l_0$ has also been introduced in (8.8). If ϕ is an arbitrary spinor, then solutions of (8.3) may be represented in the following form:

$$\chi_\pm = \rho_\pm\phi, \tag{8.9}$$

where $\rho_\pm$ are projection operators:

$$\rho_\pm = \tfrac{1}{2}(\mathbf{1} \pm \sigma_k\lambda_k) = -\tfrac{1}{2}\overset{(2)}{\not{\Pi}}(\mp\lambda), \tag{8.10}$$

$$\rho_\pm^2 = \rho_\pm, \quad \rho_\pm\rho_\mp = 0. \tag{8.11}$$

Notice also that $\rho_\pm$ depends only on the angular coordinates, for

$$\rho_\pm = \tfrac{1}{2}\begin{pmatrix} 1 \pm \lambda_3 & \pm\lambda_- \\ \pm\lambda_+ & 1 \mp \lambda_3 \end{pmatrix}, \tag{8.12}$$

where $\lambda_\pm = \lambda_1 \pm i\lambda_2$. Since λ depends only on angular coordinates, so does $\rho_\pm$, and therefore $\rho_\pm$ commutes with $\lambda^\mu\partial_\mu = \partial_t - \partial_\rho$:

$$(\rho_\pm, \lambda^\mu\partial_\mu) = (\rho_\pm, \partial_t - \partial_\rho) = 0. \tag{8.13}$$

It follows from (8.8) that

$$\rho_\pm(\partial_0 \mp \sigma_k\partial_k)\chi_\pm = 0. \tag{8.14}$$

But

$$\begin{aligned}\rho_\pm(\mp\sigma_k\partial_k)\chi_\pm &= \mp\tfrac{1}{2}(1 \pm \sigma_m\lambda_m)(\sigma_k\partial_k)\chi_\pm \\ &= \mp\tfrac{1}{2}[(\sigma_k\partial_k)(\pm\sigma_m\lambda_m) \pm (\sigma_m\lambda_m)(\sigma_k\partial_k)]\chi_\pm \\ &= -[\lambda_m\partial_m + (\alpha - i\beta\lambda_m\sigma_m)]\chi_\pm \\ &= -[\lambda_m\partial_m + (\alpha \mp i\beta)]\chi_\pm.\end{aligned} \tag{8.15}$$

Then (8.14) becomes

$$\partial_0\chi_\pm - \lambda_m\partial_m\chi_\pm = (\alpha \mp i\beta)\chi_\pm$$

or

$$\lambda^\mu\partial_\mu\chi_\pm = (\alpha \mp i\beta)\chi_\pm. \tag{8.16}$$

It then follows from (8.8) that

$$(\partial_0 \mp \sigma_k\partial_k)\chi_\pm = i\mathbf{m}\chi_\mp, \tag{8.17}$$

and therefore $\chi_\pm$ satisfy the Klein—Gordon equation:

$$(\partial_0^2 - \partial_k^2)\chi_\pm = -\mathbf{m}^2\chi_\pm. \tag{8.18}$$

The required solutions must satisfy (8.3), (8.16), and (8.17) simultaneously. Equation (8.16) may be expressed as follows:

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial\rho}\right)\chi_+ = \left(\frac{1}{\omega^*}\right)\chi_+, \tag{8.16+}$$

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial\rho}\right)\chi_- = \left(\frac{1}{\omega}\right)\chi_-. \tag{8.16-}$$

These are easily solved in the form

$$\chi_+ = \left(\frac{1}{\omega^*}\right)\exp[ik_+(t + \omega^*)]\,C_+(\sigma,\phi), \tag{8.19+}$$

$$\chi_- = \left(\frac{1}{\omega}\right)\exp[ik_-(t + \omega)]\,C_-(\sigma,\phi), \tag{8.19-}$$

where we have replaced ρ by ω in the exponential and the angular factors $C_\pm(\sigma,\phi)$ remain to be determined by (8.17). By substituting (8.19) into (8.18) one finds that the mass $\mathbf{m}$ must vanish. This result evidently stems from the condition (8.1). Therefore, the mass $\mathbf{m}$ should be dropped from all equations which depend on (8.1).

9. EXTREME RELATIVISTIC LIMIT

Let us consider the case of neutrinos or of charged particles for which $E/\mathbf{m} \gg 1$. Then (8.17) applies since $\mathbf{m}$ must vanish in that equation anyway, and

$$(\partial_0 \mp \sigma_k\partial_k)\chi_\pm = 0. \tag{9.1}$$

The χ must satisfy (9.1), (8.16), and (8.3). To con-

form to the condition (8.3), we may put

$$\chi_+ = \frac{1}{\omega^*}\exp[ik_+(t+\omega^*)]F_+(\sigma,\phi)\begin{pmatrix}1\\ L_+\end{pmatrix}, \tag{9.2}$$

where

$$L_+ = \frac{\lambda_+}{1+\lambda_3} = \frac{\sin\theta\exp(i\phi)}{1+\cos\theta} = \left(\frac{1+\sigma/a}{1-\sigma/a}\right)^{1/2}\exp(i\phi) \tag{9.3}$$

and where $F_+(\sigma,\phi)$ is a scalar function. One may now reduce (9.1) to the following form:

$$O_\pm\chi_\pm = 0, \tag{9.4±}$$

where

$$O_\pm = \overset{(3)}{M}_{j3}(\pm\lambda)\partial_j \tag{9.4a}$$

$$= \partial_3 - \lambda_3\lambda_s\partial_s \pm i(\lambda_2\partial_1 - \lambda_1\partial_2). \tag{9.4b}$$

If (9.4b) is written in (ρ,σ,ϕ) coordinates, then it becomes

$$O_+ = O_-^* = -i\gamma[\partial_\phi + a(1-\sigma^2/a^2)(\partial_\rho - i\partial_\sigma)]. \tag{9.4c}$$

Substitute (8.19+) into (9.4+). Then

$$[\partial_\phi + a(1-\sigma^2/a^2)(\partial_\rho - i\partial_\sigma)](1/\omega^*) \times\exp[ik_+(t+\omega^*)C_+(\sigma,\phi) = 0. \tag{9.5}$$

Observe that

$$\partial_\omega\omega^* = (\partial_\rho - i\partial_\sigma)\omega^* = 0, \tag{9.6a}$$

$$\partial_\phi\omega^* = 0, \tag{9.6b}$$

so that

$$O_+C_+(\sigma,\phi) = 0, \tag{9.7}$$

where by (9.2)

$$C_+(\sigma,\phi) = F_+(\sigma,\phi)\begin{pmatrix}1\\ L_+\end{pmatrix}. \tag{9.8}$$

Observe also that

$$O_+\begin{pmatrix}1\\ L_+\end{pmatrix} = 0. \tag{9.9}$$

Therefore,

$$O_+F_+(\sigma,\phi) = 0 \tag{9.10}$$

or

$$[\partial_\phi + a(1-\sigma^2/a^2)(\partial_\rho - i\partial_\sigma)]F_+(\sigma,\phi) = 0. \tag{9.10'}$$

The solution of (9.10) is

$$F_+(\sigma,\phi) = \exp(im_+\phi)\left(\frac{1+\sigma/a}{1-\sigma/a}\right)^{m_+/2}. \tag{9.11}$$

Therefore, by (9.2) the required solutions are

$$\chi_+(\rho,\sigma,\phi) = \left(\frac{\exp[ik_+(t+\omega^*)]}{\omega^*}\right)\left(\frac{1+\sigma/a}{1-\sigma/a}\right)^{m_+/2} \times\frac{\exp[i(m_+ + \frac{1}{2})\phi]}{(1-\sigma/a)^{1/2}}\begin{pmatrix}\exp(-i\phi/2)(1-\sigma/a)^{1/2}\\ \exp(i\phi/2)(1+\sigma/a)^{1/2}\end{pmatrix}. \tag{9.12+}$$

In the same way we obtain

$$\chi_-(\rho,\sigma,\phi) = \left(\frac{\exp[ik_-(t+\omega)]}{\omega}\right)\left(\frac{1+\sigma/a}{1-\sigma/a}\right)^{m_-/2} \times\frac{\exp[-i(m_- + \frac{1}{2})\phi]}{(1-\sigma/a)^{1/2}}\begin{pmatrix}i\exp(-i\phi/2)(1+\sigma/a)^{1/2}\\ -i\exp(i\phi/2)(1-\sigma/a)^{1/2}\end{pmatrix}. \tag{9.12-}$$

These final wavefunctions are very simple in terms of the coordinate system (ρ,σ,ϕ) that has been used in this paper. They are also special cases of the spinors introduced in (4.12) and (4.14), but they differ from the $V(\pm\lambda)$ appearing in Sec. 5 by a position dependent normalization which was there chosen so that the vector bilinear covariant would agree with λ. As a result, however, $V^\dagger\sigma V$ was not conserved, because λ is itself not conserved. On the other hand χ satisfies the Dirac equation; the corresponding Dirac bilinear covariant is simply the Dirac current, and therefore it is conserved, as we shall now verify.

10. CURRENTS

The covariant Dirac equation implies the covariant conservation of current in the usual way. We have

$$\gamma^\mu(\partial_\mu - \Gamma_\mu)\psi = \mathrm{i}m\psi, \tag{10.1a}$$

$$\partial_\mu\bar\psi\gamma^\mu + \bar\psi\Gamma_\mu\gamma^\mu = -im\bar\psi, \tag{10.1b}$$

and therefore

$$\partial_\mu(\bar\psi\gamma^\mu\psi) - \bar\psi\partial_\mu\gamma^\mu\psi + \bar\psi(\Gamma_\mu,\gamma^\mu)\psi = 0. \tag{10.2}$$

But

$$\partial_\mu\gamma^\mu + \Gamma^\mu_{\sigma\mu}\gamma^\sigma + (\gamma^\mu,\Gamma_\mu) = 0. \tag{10.3}$$

In a general metric it follows from the preceding two equations that the current is covariantly conserved:

$$\partial_\mu(\bar\psi\gamma^\mu\psi) + \Gamma^\mu_{\sigma\mu}(\bar\psi\gamma^\sigma\psi) = 0. \tag{10.4}$$

For the Kerr–Schild metric, however, (10.4) becomes the ordinary conservation law,

$$\partial_\mu(\bar\psi\gamma^\mu\psi) = 0. \tag{10.5}$$

Here

$$\bar\psi\gamma^\mu\psi = \psi^\dagger\eta^0\gamma^\mu\psi. \tag{10.6}$$

Notice that η^0 rather than γ^0 appears in the Dirac adjoint spinor; and that (10.1b) is consistent with this definition.

The currents associated with the particular solutions that we have found are

$$S_\mu(\pm) = (\chi(\pm)^*\chi(\pm), \pm\chi(\pm)^*\boldsymbol{\sigma}\chi(\pm)), \tag{10.7}$$

or

$$S_0(\pm) = \gamma\gamma^*\exp(\pm 2k_\pm\sigma)\left(\frac{1+\sigma/a}{1-\sigma/a}\right)^{m_\pm}\left(\frac{2}{1-\sigma/a}\right), \tag{10.7a}$$

$$\mathbf{S}(\pm) = \gamma\gamma^*\exp(\pm 2k_\pm\sigma)\left(\frac{1+\sigma/a}{1-\sigma/a}\right)^{m_\pm}\left(\frac{2}{1-\sigma/a}\right)(\lambda). \tag{10.7b}$$

Since S^0 is time independent, the conservation law is

simply

$$\partial_k S^k(\pm) = 0, \tag{10.8}$$

where the above two helicity currents may be written in the following way:

$$S^k(\pm) = \gamma\gamma^* F_\pm(\sigma)\lambda^k. \tag{10.9}$$

Then the conservation law may be checked:

$$\partial_k S^k(\pm) = \gamma\gamma^* F_\pm(\sigma)\,\partial_k\lambda^k + \gamma\gamma^*\,\lambda^k\partial_k F_\pm(\sigma) + F_\pm(\sigma)\,\lambda^k\partial_k(\gamma\gamma^*).$$

But

$$\lambda^k\partial_k F_\pm(\sigma) = \frac{\partial}{\partial\rho}F_\pm(\sigma) = 0 \tag{10.10}$$

and

$$\partial_k S^k(\pm) = F_\pm(\sigma)[\gamma\gamma^*\,\partial_k\lambda^k + \lambda^k\,\partial_k(\gamma\gamma^*)] = 0 \tag{10.11}$$

by (2.19).

Notice that the conserved Dirac current S^α agrees with the conserved currents j^α introduced in (2.2) and computed in (3.9) if the angular factor, $C(\sigma)$, in (3.9) is properly chosen.

The wavefunctions obtained in this paper describe massless spinor particles that fall into a rotating, and possibly charged, source. The current associated with these wavefunctions flows inward along the direction of the λ lines. The corresponding outgoing wavefunctions are, of course, the time reversed spinors. By maximizing S_0 with respect to σ in the limit of large $m\hbar$ one obtains the classical limit for these quantal motions. One finds

$$L_z = AE\sin^2\theta, \tag{10.12}$$

where E and L_z are the energy and the angular momentum about the axis of rotation. These classical orbits lie on cones, with constant polar angle θ, determined by the constants L_z and E. These spirals are the principal null congruences.

*Work supported in part by the National Science Foundation.

[1] S. A. Teukolsky, Astrophys. J. **185**, 635 (1973).
[2] S. Chandrasekhar, preprint, Institute for Advanced Study, Princeton (1976).
[3] M. M. Schiffer, R. J. Adler, J. Mark, and C. Sheffield, J. Math. Phys. **14**, 52 (1973).
[4] R. Finkelstein, J. Math. Phys. **16**, 1271 (1975). The algebraic relations summarized in the first two sections of this paper are derived in Refs. 3 and 4 and here given without proof.

[5] By multiplying (5.9) by $V^*(\pm\lambda)$ an explicit form can be obtained for ζ_k, namely, $\zeta(\pm\lambda) = \mp iV^*(\pm\lambda)\,\partial_k V(\pm\lambda)$.

Characterization of the Szekeres inhomogeneous cosmologies as algebraically special spacetimes

J. Wainwright

Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada
(Received 27 September 1976)

The Szekeres inhomogeneous cosmological models are invariantly characterized as a subclass of the algebraically special type {22} solutions of the Einstein field equations for irrotational dust, and their relationship to the locally rotationally symmetric dust solutions is clarified.

1. INTRODUCTION

Recently, Szekeres[1] derived a class of exact solutions of the Einstein field equations with irrotational dust as source, which he interpreted as inhomogeneous cosmological models. He imposed no *a priori* symmetry assumptions, but instead demanded that the metric and velocity u of the dust have the form

$$ds^2 = -dt^2 + e^{2a}\,dr^2 + e^{2b}\,d\xi\,d\bar{\xi}, \tag{1.1}$$

$$u = \frac{\partial}{\partial t}, \tag{1.2}$$

where a, b are arbitrary functions of the coordinates, and $\bar{\xi}$ is the complex conjugate of the complex coordinate ξ.

In this note, we show, using the Newman–Penrose formalism, that the Szekeres solutions are algebraically special and of type {22}, and that from this point of view, they generalize in a natural way the locally rotationally symmetric[2] (LRS) dust solutions, which were classified by Ellis.[3] This leads to an invariant characterization of the Szekeres solutions as a subclass of the type {22} solutions of the Einstein field equations for irrotational dust.

We assume knowledge of the Newman–Penrose formalism and of the associated geometrical concepts: the shear, and expansion of a congruence of null geodesics and the geometrical version of the Petrov classification in terms of principal null directions.[4] The Ricci and Bianchi identities in Newman–Penrose form are labeled $I_1 - I_{18}$ and $II_1 - II_{11}$ respectively. Most of the calculations were performed using a computer program written in the language CAMAL by Campbell. Given the components of a null tetrad, the program first calculates the spin coefficients, and then, using the Newman–Penrose form of the Ricci identities, calculates the complex tetrad components of the Weyl and Ricci tensors.

We begin in Sec. 2 by describing some properties of the LRS perfect fluid solutions,[5] some of which have been given previously.[6] The main result is proved in Sec. 3.

2. THE LRS PERFECT FLUID SOLUTIONS

The line element and fluid velocity of the LRS perfect fluid solutions can be written in the form

$$ds^2 = X^2(dx^0 - p\,dx^3)^2 - Y^2(dx^1 - q\,dx^3)^2 - Z^2[(dx^2)^2 + r^2(dx^3)^2], \tag{2.1}$$

$$u = X^{-1}\frac{\partial}{\partial x^0}, \tag{2.2}$$

where p, q, r are functions of x^2, and one of the following specializations occurs:

Class I: $Y = 1$, $X = X(x^1)$, $Z = Z(x^1)$, $q = 0$, $p' \neq 0$,

Class II: $X = X(x^0, x^1)$, $Y = Y(x^0, x^1)$, $Z = Z(x^0, x^1)$, $p = q = 0$,

Class III: $X = 1$, $Y = Y(x^0)$, $Z = Z(x^0)$, $p = 0$, $q' \neq 0$.

For a pressure-free fluid (dust), we have $X = 1$. The specific form of the functions p, q, r is given by Ellis[3] (see p. 1181). All that we require here is that $(dx^2)^2 + r^2(dx^3)^2$ is the metric of a 2-space of constant curvature.

A null tetrad for the metric (2.1) is defined[6] by

$$k_a\,dx^a = 2^{-1/2}[X\,dx^0 - Y\,dx^1 - (pX - qY)\,dx^3], \tag{2.3a}$$

$$n_a\,dx^a = 2^{-1/2}[X\,dx^0 + Y\,dx^1 - (pX + qY)\,dx^3], \tag{2.3b}$$

$$m_a\,dx^a = 2^{-1/2}Z(dx^2 + i r\,dx^3). \tag{2.3c}$$

It follows from Eqs. (2.1)–(2.3) that the fluid velocity lies in the 2-space generated by k and n,

$$u_{[a}k_b n_{c]} = 0. \tag{2.4}$$

The CAMAL program referred to in the Introduction shows that the Newman–Penrose quantities for this tetrad are restricted as follows:

$$\kappa = \sigma = \nu = \lambda = \tau = \pi = \alpha + \bar{\beta} = 0,$$

$$\Phi_{02} = \Psi_0 = \Psi_4 = 0, \tag{2.5}$$

$$\Psi_1 + \Phi_{01} = \Psi_3 + \Phi_{21} = 0.$$

On account of (2.4), the Einstein field equations with perfect fluid source imply[6] that

$$\Phi_{01} = \Phi_{02} = \Phi_{12} = 0,$$

and hence that

$$\Psi_0 = \Psi_1 = 0 = \Psi_3 = \Psi_4.$$

I.e., the null vectors k^a, n^a are repeated principal null directions of the Weyl tensor. The LRS solutions thus satisfy the following properties.

Property (A): The Weyl tensor is algebraically special of type {22}, and the fluid velocity lies in the 2-space spanned by the repeated principal null directions.

Remark: The 2-spaces referred to in property (A) are in fact surface forming, as follows from (2.5) and (A1), though we shall not need this fact here.

The second property which we wish to emphasize relates to the shear[5] of the fluid.

Copyright © 1977 American Institute of Physics

Property (*B*): Any vector which is orthogonal to both of the repeated principal null directions is an eigenvector of the shear tensor σ_{ab}.

This can equivalently be written

$$m_{[a}\sigma_{b]c}m^c = 0, \tag{2.6}$$

in which form its validity is clear, on account of (2.5) and (A6).

Remark: The vorticity tensor, and acceleration vector satisfy properties analogous to (2.6), namely

$$m_{[a}\omega_{b]c}m^c = 0, \quad \dot{u}_a m^a = 0,$$

[see (A7) and (A8)] though we shall not need this fact here.

Finally, we have from (2.5) [$\kappa = \nu = 0$],

Property (*C*): The repeated principal null directions are tangent to (null) geodesics.

Properties (A)—(C) hold for each class of LRS solutions. The three classes can be invariantly distinguished by considering the rotation of the fluid (described by the vorticity tensor ω_{ab}) and the twist of the principal null directions (described by the spin coefficients $\rho - \bar{\rho}$ and $\mu - \bar{\mu}$). From Stewart and Ellis,[5] and Appendix B, one obtains:

Class I: $\omega_{ab} \neq 0$; $\rho - \bar{\rho} \neq 0$, $\mu - \bar{\mu} \neq 0$;

Class II: $\omega_{ab} = 0$; $\rho - \bar{\rho} = 0$, $\mu - \bar{\mu} = 0$;

Class III: $\omega_{ab} = 0$; $\rho - \bar{\rho} \neq 0$. $\mu - \bar{\mu} \neq 0$.

This classification implies, on account of (A2) and (A3), that the class II solutions, and only these, satisfy

$$k_{[a;b}k_b n_{c]} = 0 = n_{[a;b}n_c k_{d]}. \tag{2.7}$$

This leads to Property (D), which we stress *is only satisfied by class II solutions*.

Property (*D*): The 2-spaces generated by the repeated principal null directions admit orthogonal 2-surfaces.[7]

3. THE SZEKERES SOLUTIONS

On comparison of (1.1), (1.2), and (2.1), (2.2) (with $X = 1$, $\xi = x^2 + ix^3$), it is clear that the Szekeres class of solutions contains the LRS class II *dust* solutions.[8] We will show that the Szekeres solutions satisfy Properties (A) and (B) of the LRS solutions, but in general not (C), i.e., *the Szekeres solutions are algebraically special of type* {22} *but the repeated principal null directions are in general not tangent to geodesics*. However, the Szekeres solutions also satisfy Property (D), and are in fact characterized by Properties (A), (B), and (D).

Theorem: The Szekeres solutions with nonzero Weyl tensor comprise all solutions of the Einstein field equations

$$G_{ij} = \kappa\rho u_i u_j, \quad \rho > 0, \tag{3.1}$$

with u_i hypersurface-orthogonal (i.e., with an irrotational dust source), which satisfy Properties (A), (B), and (D). The subclass which in addition satisfies Property (C) comprises the LRS class II dust solutions.

The proof of the theorem depends on the following lemmas.

Lemma 1: Consider the Szekeres line element (1.1) and associated contravariant null tetrad, defined by

$$k = 2^{-1/2}\left(\frac{\partial}{\partial t} + \exp(-a)\frac{\partial}{\partial r}\right),$$

$$n = 2^{-1/2}\left(\frac{\partial}{\partial t} - \exp(-a)\frac{\partial}{\partial r}\right), \tag{3.2}$$

$$m = 2^{1/2}\exp(-b)\frac{\partial}{\partial \xi}.$$

The Newman—Penrose quantities associated with this tetrad satisfy

$$\lambda = \sigma = 0,$$

$$\kappa + \tau = \tau + \bar{\pi} = \nu + \pi = \alpha + \bar{\beta} = 0,$$

$$\rho = \bar{\rho}, \quad \mu = \bar{\mu}, \quad \epsilon = \bar{\epsilon}, \quad \gamma = \bar{\gamma}, \quad \epsilon + \gamma = 0, \tag{3.3}$$

$$\bar{\psi}_0 = \psi_4 = \phi_{20}, \quad \psi_2 = \bar{\psi}_2,$$

$$\bar{\psi}_1 - \phi_{21} = \psi_3 - \phi_{10} = \dot{b}_\xi \exp(-b).$$

Proof: These results were established using the CAMAL program mentioned in the Introduction.

Corollary: The Szekeres solutions satisfy Properties (A), (B), and (D).

Proof: It follows from (1.1) and (3.2) that (2.4) holds. Hence the Einstein equations (3.1) imply, as before, that $\phi_{01} = \phi_{02} = \phi_{12} = 0$. In addition Szekeres proves that $\dot{b}_\xi = 0$. Hence it follows from (3.3) that $\psi_0 = \psi_1 = \psi_3 = \psi_4 = 0$, which establishes Property (A). Properties (B) and (D) follow immediately from (3.3) using (A6), (A2), and (A3).

Lemma 2: Suppose that the Einstein field equations for irrotational dust are satisfied and that Properties (A), (B), and (D) are valid. Then the complex null vector m, which is defined up to scale by

$$m^a k_a = 0 = m^a n_a,$$

and the real spacelike vector v which is defined up to scale by

$$v_{[a}k_b n_{c]} = 0, \quad u^a v_a = 0,$$

satisfy[9]

$$m_{[a;b}m_{c]} = 0 = v_{[a;b}v_{c]}. \tag{3.4}$$

Proof: On account of Property (A), we can adjust the scale of k and n so that

$$u^a = 2^{-1/2}(k^a + n^a).$$

Then v can be scaled so that

$$v^a = 2^{-1/2}(k^a - n^a),$$

and m can be chosen so that $\{k, n, m, \bar{m}\}$ form a null tetrad. Condition (A), in conjunction with (3.1), then implies that

$$\psi_0 = \psi_1 = \psi_3 = \psi_4 = 0,$$

$$\phi_{01} = \phi_{02} = \phi_{12} = 0, \quad \phi_{00} = \phi_{22} = 2\phi_{11}.$$

On using these conditions, it is found that the Bianchi identity II_2 reduces to

$$0 = 3\sigma\psi_2 + (\sigma - \bar{\lambda})\phi_{00}.$$

On account of (A6), Condition (B) implies $\sigma=\bar{\lambda}$, so that since $C_{abcd}\neq 0$, we obtain

$$\sigma=\lambda=0. \tag{3.5}$$

Secondly, Condition (B) and the requirement of non-rotating dust implies, on account of (A6) and (A7), that

$$\alpha+\bar{\beta}=0, \quad \bar{\tau}+\pi=\bar{\kappa}+\nu. \tag{3.6}$$

It thus follows from the Bianchi identities II_1 and II_8, after elimination of $\delta\phi_{00}$, that

$$\kappa\psi_2+\bar{\nu}\bar{\psi}_2=0 \tag{3.7}$$

In addition, Condition (D) implies [see (A2) and (A3)] that

$$\rho=\bar{\rho}, \quad \mu=\bar{\mu}. \tag{3.8}$$

On making use of Eqs. (3.5)—(3.8), the Ricci identity I_{12} simplifies to

$$\delta\alpha+\bar{\delta}\bar{\alpha}=\mu\rho+4\alpha\bar{\alpha}-\psi_2+\Lambda+\phi_{11},$$

from which we conclude that

$$\psi_2=\bar{\psi}_2\neq 0.$$

Equations (3.7) and (3.6) thus yield

$$\bar{\tau}+\pi=\bar{\kappa}+\nu=0. \tag{3.9}$$

Equations (3.5), (3.6), (3.8), and (3.9), in conjunction with (A4) and (A5), now yield the desired result.

Proof of the Theorem: The corollary to Lemma 1 proves that the Szekeres solutions satisfy (A), (B), and (D).

Conversely, suppose that (A), (B), and (D) hold. Since m_a and v_a satisfy (3.4), by Lemma 2, and u_a is hypersurface-orthogonal by assumption, it follows that u_a, v_a, and m_a are locally proportional to gradients, i.e., there exist real functions X, Y, t, r and complex functions Z, ξ, such that

$$u_a\,dx^a=X\,dt, \quad v_a\,dx^a=Y\,dr, \quad m_a\,dx^a=2^{-1/2}Z\,d\xi.$$

Using a phase change $m^a\to\exp(iS)m^a$, we can make Z real. If we now use t, r, ξ, $\bar{\xi}$ as local coordinates, the line element assumes the form

$$ds^2=X^2\,dt^2-Y^2dr^2-Z^2\,d\xi\,d\bar{\xi}.$$

In addition, since u^a is tangent to geodesics (the flow lines of the dust), it follows that $X=X(t)$, so that we can set $X=1$ by redefining the t coordinate. Then $u=\partial/\partial t$, and we have arrived at Szekeres' ansatz, namely Eqs. (1.1) and (1.2).

The final statement in the theorem is proved by noting (see Appendix B) that

$$\kappa=\nu=0\Leftrightarrow a_\xi=0,$$

i.e., k^a and n^a are tangent to geodesics if and only if $a_\xi=0$. It then follows, using the results of Szekeres' paper, that the metric has the form

$$ds^2=-dt^2+\exp[2a(r,t)]\,dr^2+A(r,t)^2(1+K\xi\bar{\xi})^{-2}\,d\xi\,d\bar{\xi},$$

with $K=\text{const}$. Since $(1+K\xi\bar{\xi})^{-2}\,d\xi\,d\bar{\xi}$ is the metric of a 2-space of constant curvature, it follows that the solutions in this case comprise the LRS class II dust solutions.

4. CONCLUSION

We have characterized the Szekeres solutions as a subclass of the type $\{22\}$ solutions of the Einstein field equations for irrotational dust, and at the same time have clarified the way in which the Szekeres solutions generalize the LRS dust solutions. This approach to the Szekeres solutions also suggests possible generalizations, and the existence of related classes of solutions. For example, one could assume (A), (B), and (D), but permit the source to be an irrotational fluid (i.e., nonzero pressure). Indeed particular solutions of this type, with Property (C) not valid, were derived several years ago by Stephani,[10] from a completely different point of view. A second possibility would be to attempt to generalize the LRS I and III solutions in the same way that Szekeres' solutions generalize the LRS II solutions, i.e., by not demanding that the repeated principal null directions be tangent to geodesics. Indeed, it may be possibly to classify all solutions of the Einstein field equations with a perfect fluid source, subject only to property (A),[11] and this could well yield further examples of inhomogeneous cosmologies.

APPENDIX A

Given a null tetrad $\{k, n, m, \bar{m}\}$ it is straightforward to verify that

$$k^a{}_{;b}n^b-n^a{}_{;b}k^b=(\gamma+\bar{\gamma})k^a+(\epsilon+\bar{\epsilon})n^a-(\bar{\tau}+\pi)m^a-(\tau+\bar{\pi})\bar{m}^a, \tag{A1}$$

$$k_{[a;b}k_cn_{d]}=-(\rho-\bar{\rho})k_{[a}n_bm_c\bar{m}_{d]}, \tag{A2}$$

$$n_{[a;b}n_ck_{d]}=(\mu-\bar{\mu})k_{[a}n_bm_c\bar{m}_{d]}, \tag{A3}$$

$$m_{[a;b}m_{c]}=(\tau+\bar{\pi})k_{[a}n_bm_{c]}+\bar{\lambda}k_{[a}m_b\bar{m}_{c]}-\sigma n_{[a}m_b\bar{m}_{c]}, \tag{A4}$$

$$v_{[a;b}v_{c]}=-\tfrac{1}{2}[2(\alpha+\bar{\beta})-(\bar{\tau}+\pi)-(\bar{\kappa}+\nu)]k_{[a}n_bm_{c]}+(*)-\tfrac{1}{2}(\rho-\bar{\rho}-\mu+\bar{\mu})k_{[a}m_b\bar{m}_{c]}+(*), \tag{A5}$$

where

$$v^a=2^{-1/2}(k^a-n^a),$$

and $(*)$ denotes the complex conjugate of the preceding term.

Consider the unit timelike vector field defined by

$$u^a=2^{-1/2}(k^a+n^a).$$

The shear tensor σ_{ab}, vorticity tensor ω_{ab}, acceleration vector $\dot{u}_a$, and expansion θ of this vector field are given by[12]

$$\sigma_{ab}=A[v_av_b-m_{(a}\bar{m}_{b)}]+Bv_{(a}m_{b)}+(*)+Cm_am_b+(*), \tag{A6}$$

where

$$A=-\tfrac{1}{3}2^{-1/2}[(\rho+\bar{\rho})-(\mu+\bar{\mu})+2(\epsilon+\bar{\epsilon})-2(\gamma+\bar{\gamma})],$$

$$B=-\tfrac{1}{2}[(\bar{\tau}+\pi)+2(\alpha+\bar{\beta})-(\bar{\kappa}+\nu)],$$

$$C=2^{-1/2}(\bar{\sigma}-\lambda),$$

and

$$\omega_{ab}=Uv_{[a}m_{b]}+(*)+Vm_{[a}\bar{m}_{b]}, \tag{A7}$$

where

$$U=\tfrac{1}{2}[(\bar{\tau}+\pi)-2(\alpha+\bar{\beta})-(\bar{\kappa}+\nu)],$$

$$V=2^{-1/2}[-(\rho-\bar{\rho})-(\mu-\bar{\mu})],$$

and

$$\dot{u}_a=2^{-1/2}[(\epsilon+\bar{\epsilon})+(\gamma+\bar{\gamma})]v_a+\tfrac{1}{2}(-\bar{\tau}+\pi-\bar{\kappa}+\nu)m_a+(*), \tag{A8}$$

$$\theta=2^{-1/2}[(\epsilon+\bar{\epsilon})-(\gamma+\bar{\gamma})-(\rho+\bar{\rho})+(\mu+\bar{\mu})]. \tag{A9}$$

APPENDIX B

We list here the nonzero spin coefficients for the LRS perfect fluid solutions, and for the Szekeres metric (1.1).

LRS class I:

$$\rho=\mu=-2^{-3/2}(ip'r^{-1}XZ^{-2}+2Z'Z^{-1}),$$

$$\epsilon=\gamma=-2^{-5/2}(ip'r^{-1}XZ^{-2}-2X'X^{-1}),$$

$$\alpha=-\beta=2^{-3/2}r'r^{-1}Z^{-1}.$$

LRS class II:

$$\rho+\mu=-2^{1/2}Z_1Z^{-1}Y^{-1},$$

$$\rho-\mu=-2^{1/2}Z_0Z^{-1}X^{-1},$$

$$\epsilon+\gamma=2^{-1/2}X_1X^{-1}Y^{-1},$$

$$\epsilon-\gamma=2^{-1/2}Y_0X^{-1}Y^{-1},$$

$$\alpha=-\beta=2^{-3/2}r'r^{-1}Z^{-1},$$

where subscripts 0, 1 denote differentiation with respect to x^0 and x^1.

LRS class III:

$$\rho=-\mu=2^{-3/2}(iq'r^{-1}YZ^{-2}-2Z'Z^{-1}),$$

$$\epsilon=-\gamma=2^{-5/2}(iq'r^{-1}YZ^{-2}+2Y'Y^{-1}),$$

$$\alpha=-\beta=2^{-3/2}r'r^{-1}Z^{-1}.$$

Szekeres class:

$$\kappa=-\tau=\bar{\pi}=-\bar{\nu}=2^{-1/2}a_t e^{-b},$$

$$\rho+\mu=-2^{1/2}b'e^{-a},$$

$$\rho-\mu=-2^{1/2}\dot{b},$$

$$\epsilon=-\gamma=2^{-3/2}\dot{a},$$

$$\bar{\alpha}=-\beta=-2^{-1/2}b_t e^{-b}.$$

The only nonzero complex tetrad component of the Weyl tensor is given, after simplification using the field equations, by

$$\psi_2=[\ddot{b}-\ddot{a}+(\dot{b}-\dot{a})(\dot{b}+\dot{a})]/3.$$

Note added in proof: The computer program used in this paper is described in detail in S.J. Campbell and J. Wainwright, "Symbolic Computation and the Newman–Penrose Formalism in General Relativity," preprint, University of Waterloo, Dec. 1976. The CAMAL language is described in the "CAMAL User's Manual," written by J. Fitch, The Computer Laboratory, University of Cambridge, Cambridge, England.

ACKNOWLEDGMENTS

It is a pleasure to thank S. J. Campbell for writing the computer program used in this work. Financial support from a National Research Council of Canada research grant and from the Faculty of Mathematics at the University of Waterloo is gratefully acknowledged.

[1] P. Szekeres, Commun. Math. Phys. 41, 55 (1975). We point out a number of minor errors in the form of the solutions given on page 61: In class (i), $u(r)$ and $w(r)$ should satisfy $u(r)+w(r)=0$; in class (ii), $u(r)$ and $w(r)$ should satisfy $u(r)-w(r)=0$; in class (iiia), the density should be $\kappa\rho=(4/3)\sigma/t^{4/3}(\mu+\phi\sigma)$. The existence of these errors has also been pointed out by W.B. Bonnor and N. Tomimura, Mon. Not. R. Astron. Soc. 175, 85 (1976). The coordinates used in this paper are analogous to the coordinates used in Ref. 3.

[2] This class of solutions is characterized by the fact that they admit a (local) group of isometries which acts multiply transitively on the group orbits.

[3] G.F.R. Ellis, J. Math. Phys. 8, 1171 (1967).

[4] E.T. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962). For more details and a complete listing of the Newman–Penrose equations see F.A.E. Pirani, in *Brandeis Summer Institute in Theoretical Physics Lectures on General Relativity* (Prentice-Hall, Englewood Cliffs, N.J., 1964).

[5] J.M. Stewart and G.F.R. Ellis, J. Math. Phys. 9, 1072 (1968).

[6] J. Wainwright, Commun. Math. Phys. 17, 42 (1970).

[7] H. Cartan, *Differential Forms* (Kershaw, London, 1971), 97. Note that Eq.'s (2.7 reads $dk\wedge k\wedge n=0\,dn\wedge n\wedge k$ in differential form notation.

[8] Szekeres pointed out that his solutions contain the Tolman–Bondi solutions [H. Bondi, Mon. Not. R. Astron. Soc. 410, 107 (1947)], the Kantowski–Sachs solutions [R. Kantowski and R.K. Sachs, J. Math. Phys. 7, 443 (1966)], and a class of solutions discussed by D. Eardley, E. Liang, and R.R. Sachs, J. Math. Phys. 13, 99 (1972), which are known to be LRS Class II solutions.

[9] The first of equations (3.4) is satisfied by all LRS perfect fluid solutions, the second only by Classes I and II.

[10] H. Stephani, Commun. Math. Phys. 9, 53 (1968).

[11] Not all type {22} solutions of the Einstein equations with a perfect fluid source satisfy Property (A). An example of one which does not is a solution given by H. Wahlquist, Phys. Rev. 172, 1291 (1968).

[12] M. Olesen, "Algebraically Special Fluid Space–Times with Shearing Rays," Ph.D. thesis, University of Waterloo, 1972 (unpublished).

Acoustic emission and the plate Green's function

W. J. Pardee

Science Center, Rockwell International, 1049 Camino dos Rios, Thousand Oaks, California 91360
(Received 27 October 1976)

The stress at an arbitrary point in an elastic medium due to an acoustic emission is written in a stress Green's function formalism. This formalism separates the mathematical description of acoustic radiation into separate factors describing the source and the medium response, which, it is hoped, will facilitate identification of source characteristics from remote observations. The description of the medium response is shown to be equivalent to a well-posed elliptic boundary value problem for a 36-component Green's function. This (frequency analyzed) Green's function is calculated for the infinite isotropic elastic medium and for a similar plate. The latter example includes a complete description of dispersion, mode conversion, and the normal modes of the slab, although the latter need not be explicitly identified.

1. INTRODUCTION

The term acoustic emission (AE) usually refers to the spontaneous generation of elastic waves concomitant to the mechanical deformation of an elastic solid. We are going to consider only short duration (burst) emissions. These are an important, albeit imperfect, warning of failure in certain critical structures.[1-3] Similar emissions accompany other processes of considerable materials science interest, such as corrosion pitting and many phase transformations. The fundamental experimental problem, analogous to the inverse problem of scattering theory, is to identify the AE source from the remote observation of its elastic radiation. This experimental problem is complicated by the rich variety of phenomena which accompany the propagation of such a pulse in a finite medium. Notable among these are different excitations of the normal modes of the system, the dispersion resulting from this, and the mode conversion accompanying each reflection from a boundary. To facilitate the experimental effort to identify sources, it is helpful to introduce a Green's function formalism[4-6] which gives the output stress tensor resulting from a general acoustic emission source. Since experimental results are often Fourier analyzed in time,[6,7] and because it offers considerable mathematical advantages, the Green's function defined and calculated here is always expressed in frequency, not time. This Green's function includes the boundary conditions and gives the complete linear elastic response of the medium.

The formalism developed here does permit description of a source which is nonlinear and of nonzero extent. Although the description of the source then requires separate solution of a generally difficult problem, within certain limitations, the Green's function is still an important asset. The limitations are that the nonlinear region be small compared to the distance between boundaries, and preferably small compared to a wavelength and, second, that the changes in the source itself due to reflections from distant boundaries be negligible. The first condition is common in practical AE experiments, where typical source regions are of order 20 μm, small compared to the experimentally observed wavelengths (~1 mm) and certainly small compared to sample dimensions. The second condition is usually a reasonable approximation because of the decrease in amplitude of the reflected signal, and can in some circumstances be made exact by a suitable time gate. If these two conditions are met, The Green's function is quite useful even for the nonlinear problem, because in that event the source problem can be solved neglecting the boundaries of the medium, and the Green's function contains all of the information about wave propagation in the medium. Thus, under these restrictions the medium response (boundary value) problem is decoupled from the source description. The remaining difficulty is that imposed by nature, not by the formalism. If the source of interest is reasonably described as a point body force, the Green's function can be used in the conventional way, but this Green's function is equally useful with the more realistic (though still restricted) source described above. If the source is small compared to a wavelength, the stress at an arbitrary point of the medium is the matrix product of the Green's function (a matrix depending on frequency, the observation point, and the source point) by a six-component vector describing the source. If the source is extended, the Green's function must be integrated over the region where the source is nonzero. This paper devoted to the careful statement of the formalism, and the calculation of the Green's function for two important examples.

This formalism and its potential applications have been briefly discussed in three previous phenomenologically oriented works[4-6] on this problem. This paper, however, is the first major report on a continuing project on the detailed quantitative description of elastic radiation. In Sec. 2 the formalism is derived in detail and demonstrated to give rise to a well posed elliptic boundary value problem. The first part of Sec. 2 is devoted to a derivation of the important equation and a heuristic argument for the boundary conditions to be imposed. Although this paper deals in depth only with the medium response, a discussion of alternative descriptions of the sources of spontaneous elastic radiation is included to explain the motivation for this unconventional formalism and to suggest how it will be applied. The latter part of Sec. 2 contains a mathematical justification for these boundary conditions. This has the pleasant consequence of producing a substantial computational economy, as well as a deeper understanding. In Sec. 3 the stress Green's function is obtained for the infinite medium by introducing an operational calculus. This

 Copyright © 1977 American Institute of Physics

result serves as the particular integral for the calculation of the Green's function of the plate in Sec. 4. This is the principal result of the paper. The plate should be relevant to a large class of practical problems, since the structures monitored (reactor pressure vessels, pipes) frequently have two parallel surfaces whose separation is small compared to the distance to any other boundary. The appendices contain certain detailed steps not needed to follow the argument, but needed in the numerical application.

2. THE STRESS GREEN'S FUNCTION FORMALISM

As explained in the introduction, acoustic emission is the radiation of elastic waves. This is, in some respects, a more complicated problem than the more familiar dynamic elasticity problems of transmission, reflection, and diffraction. The primary difference, of course, is the complexity of the process producing the elastic radiation. The theoretical problem has two distinct aspects, the description of the source of the radiation, and the description of the effect of the medium response on the radiation. The stress Green's function formalism developed here provides a mathematical separation of these two problems. It permits a detailed solution of the usually nonlinear problem near the source, but retains the simplicity of linear elasticity elsewhere. In addition, the stress Green's function formalism greatly simplifies the exact treatment of the boundary conditions, at the cost of a more complicated differential equation.

The acoustic emissions considered in this paper are bursts of elastic radiation which accompany, for example, the mechanical deformation of an elastic solid. There is usually concomitant continuous low level elastic wave generation. The latter type of emission, though probably just the small amplitude, low frequency limit of the burst emissions, poses substantially different problems in detection and interpretation and will not be considered here.

These elastic bursts are then, by definition, localized in time. Since we are interested in the early stages of material failure, not total disintegration of the solid, we restrict ourselves to sources which are localized in space, though, of course, the resulting elastic wave is not localized. A source might be, for example, a sudden change of surface tractions accompanying microcrack initiation. We will refer to any source which can be represented as a change in surface tractions or displacements as a surface source.

Sources internal to the solid are equally important. The physical origin of this elastic radiation is a sudden change of the internal stress. This may occur when mechanical deformation produces a stress level which locally exceeds the yield strength of the material or fractures an inclusion, or when a chemical reaction produces a precipitate in the lattice, and under many other circumstances. Whatever the detailed physical origin, the resulting elastic wave is excited by the change in internal stresses, not by some externally arising body force. Of course, it may be mathematically convenient to try to simulate this change in internal stresses by a combination of body forces. This is frequently done in seismology, and can be quite useful,[8] so the possibility of a body force is retained in the development below. However, the body force description precludes treatment of the nonlinear elastic and plastic phenomena which are intrinsic to the source process and essential to fundamental understanding. An appealing third possibility is the description of the radiation of an initial value problem with nonzero initial displacement or strain. One frequently has a better intuition for the displacements and strains than for the accompanying stresses which produce the radiation. The sudden removal of a dislocation pinning point, for example, may be naturally described as a nonzero initial strain, instead of a sudden change in the constitutive relation. It will be shown below that this description can be incorporated simply in the stress Green's function formalism, and it may be useful, but the simplicity is specious. It conceals rather than avoids the complications due to plasticity and nonlinear elasticity. This is not to say that the initial value description cannot be done correctly within the formalism given below, but doing it correctly requires solution of the same nonlinear problem. These comments are not intended to discourage use of either of these phenomenological models, but merely to encourage consideration of the fundamental origins of acoustic emission.

The physical situation, then, is one which may involve large stresses and a strain rate dependence *locally*, but which involves small total amounts of energy. Therefore, except very near the source, the emitted radiation pulse, which decreases in amplitude as it expands, will be of such small amplitude that it can be described adequately by linear elasticity.[9] Let the instantaneous position of a particle with unstrained position $\mathbf{a}$ be $\mathbf{x}(t,\mathbf{a})$. Denote by $\mathbf{u}$ the difference between strained and unstrained coordinates, $\mathbf{u}=\mathbf{x}-\mathbf{a}$. Except when it is necessary to distinguish between derivatives with respect to different (material) coordinates $\mathbf{a}$, $\mathbf{a}'$, or when attention is to be focused on differentiation operator itself, the derivative with respect to the material coordinate a_i will be denoted by a comma; that is, for example,

$$\frac{\partial f_m}{\partial a_i}=f_{m,i}. \tag{1}$$

The Lagrangian strains η_{pq} are related to the material derivatives of $\mathbf{x}$ and $\mathbf{u}$ by

$$\eta_{pq}=\tfrac{1}{2}(x_{j,p}x_{j,q}-\delta_{pq})=\tfrac{1}{2}(u_{p,q}+u_{q,p}+u_{j,p}u_{j,q}). \tag{2}$$

Although there will be no attempt in this paper to study the strain η_{pq} in the nonlinear region, to avoid ambiguity let us use a separate symbol for the small strain approximation to η_{pq},

$$e_{pq}=\tfrac{1}{2}(u_{p,q}+u_{q,p}). \tag{3}$$

The general equation of motion of $\mathbf{u}$ is[10,11]

$$\rho(t,\mathbf{x}(t,\mathbf{a}))\ddot{u}_k=\frac{\partial T_{km}}{\partial x_m}(t,\mathbf{x}(t,\mathbf{a}))+f_k(t,\mathbf{x}(t,\mathbf{a})), \tag{4}$$

where T_{km} is the full stress tensor, f_k is a possible body force, and the dots denote total time derivatives. The explicit time dependence of T is intrinsic to most AE sources. Some involve the conversion of other forms of

energy into elastic energy; for example, a solid–solid chemical reaction can produce a precipitate and concomitant lattice stress. Secondly, yield and fracture processes represent a sudden change in the constitutive relation, and this produces an explicit time dependence in **T**. In addition, the dissipative terms in **T** can be very important at the source. The amplitude and frequency spectrum of the radiation are strongly influenced by the speed with which the change in stress takes place, and this is partially determined by the strain rate dependence of **T**.

This equation is inconvenient for two reasons; it is written in terms of the strained mass density ρ, and the derivative is with respect to the strained position **x**. These deficiencies can be formally eliminated by introducing the first Piola–Kirchoff stress tensor $P_{km}(t,\mathbf{a})$ which includes the volume dilation factor

$$\rho_0/\rho = \det(x_{i,j}) = \epsilon_{ijk}\epsilon_{qrs}x_{q,i}x_{r,j}x_{s,k}/6 \tag{5}$$

and the chain rule factor

$$\frac{\partial a_m}{\partial x_j} = \tfrac{1}{2}\epsilon_{mkp}\epsilon_{jqr}x_{q,k}x_{r,p}. \tag{6}$$

That is,

$$P_{km} = \det(x_{i,j})T_{kj}\frac{\partial a_m}{\partial x_j}. \tag{7}$$

This enables us to write Eq. 4 in the form[10,11]

$$\rho_0\ddot{u}_k = P_{km,m} + F_k, \tag{8}$$

where $\mathbf{F} = \det(x_{i,j})\mathbf{f}$. Since the stresses can be expected to be linearly elastic except very near the source, it is useful to define $\mathbf{J} = \mathbf{P} - \sigma$, where σ is the linear stress defined by

$$\sigma_{ij} = C_{ijkl}e_{kl} \tag{9}$$

with C_{ijkl} the elastic stiffness tensor. The Eq. (8) becomes

$$\rho_0\ddot{u}_k - \sigma_{km,m} = J_{km,m} + F_k. \tag{10}$$

By differentiation we obtain an equation for the linear strain e_{pq},

$$\rho_0\ddot{e}_{pq} - \tfrac{1}{2}(\sigma_{pm,mq} + \sigma_{qm,mp}) = s_{pq} \tag{11}$$

with

$$s_{pq} = \tfrac{1}{2}(J_{pm,mq} + J_{qm,mp}) + \tfrac{1}{2}(F_{p,q} + F_{q,p}). \tag{12}$$

No small strain approximation has been made in obtaining Eq. (11). The assumption that the strains are linear outside a small region implies that the source s_{pq} vanishes except in that small region. The left side of Eq. (11) is then a wave equation describing the propagation of the radiation between source and observer. Note that, although P_{km} is not symmetric, the source s_{pq} is.

The program for a full description can now be outlined. This paper is devoted to the calculation of a Green's function for the Helmholtz-like version of Eq. (11),

$$\rho_0\omega^2\hat{e}_{pq} + \tfrac{1}{2}(\hat{\sigma}_{pm,mq} + \hat{\sigma}_{qm,mq}) = -\hat{s}_{pq}, \tag{13}$$

where the Fourier transform $\hat{g}(\omega)$ of a function $g(t)$ is

$$\hat{g}(\omega) = \int_{-\infty}^{\infty} dt \exp(i\omega t)g(t). \tag{14}$$

The Green's function contains the linearly elastic properties of the medium and the boundary conditions. The complete solution is obtained by specifying the Fourier transform in time of the source s_{pq}. In principle, this is obtained by solving a nonlinear elasticity problem, for, paradoxically, an infinite medium, for the boundaries can be neglected. Because experimental measurements are seldom sensitive to the detailed behavior over very short times (< 0.1 μsec) or small distances (< 100 μm), a fairly crude treatment of this problem will frequently be sufficient. We will not, henceforth, use any time dependent functions, so it will not be necessary to retain the cumbersome "$\hat{}$" to distinguish between a function in the time domain and its Fourier transform. Then Eq. (13) becomes

$$D_{pqab}(\omega,\nabla)e_{ab} = -s_{pq}, \tag{15}$$

where the differential operator **D** is defined by

$$D_{pqab}(\omega,\nabla) = \rho\omega^2\delta_{pa}\delta_{qm} + \tfrac{1}{2}(C_{pmab}\nabla_q + C_{qmab}\nabla_p)\nabla_m. \tag{16}$$

The subscript "0" has been dropped from the unstrained mass density. Let us formally define a fourth-rank tensor Green's function $G_{cdmn}(\omega,\mathbf{r},\mathbf{r}')$ as the solution to

$$D_{mnab}(\omega,\nabla')G_{cdmn}(\omega,\mathbf{r},\mathbf{r}') = -\tfrac{1}{2}(\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc})\delta^3(\mathbf{r}-\mathbf{r}'), \tag{17}$$

subject to boundary conditions to be specified later. Its precise definition and a demonstration of its existence will be given later. For the present, note simply that **G** is the kernel of a fourth-rank tensor integral operator on a space of second-rank tensors, and, considering both tensor and spatial indices on this operator, Eq. (17) is of the form

$$\mathbf{DG} = -\mathbf{1}, \tag{18}$$

where a superior "~" denotes matrix transpose.

In order to obtain an identity for e_{pq} in terms of **G**, **s**, and certain surface integrals, it is necessary to multiply Eq. (15) by **G** and perform suitable integrations. For notational economy, define a tensor $F_{ab}(\mathbf{r}')$ which is identically zero by

$$F_{ab}(\omega,\mathbf{r}') = D_{abcd}(\omega,\nabla')e_{cd}(\omega,\mathbf{r}') + s_{ab}(\omega,\mathbf{r}'). \tag{19}$$

After multiplying $F_{ab}(\omega,\mathbf{r}')$ by $G_{mnab}(\omega,\mathbf{r},\mathbf{r}')$ and twice writing $gf_{,k} = (gf)_{,k} - g_{,k}f$, we integrate over $\mathbf{r}'$ and apply Gauss' theorem to the two divergences introduced. That is, we have identically, $= \nabla_b$

$$\begin{aligned} &G_{mnab}(C_{ajcd}\nabla'_b + C_{bjcd}\nabla'_a)\nabla'_j e_{cd} \\ &\quad = \nabla'_b(G_{mnab}C_{ajcd}\nabla'_j e_{cd}) + \nabla'_a(G_{mnab}C_{bjcd}\nabla'_j e_{cd}) \\ &\quad\cdot\ -\nabla'_j[(\nabla'_b G_{mnab})C_{ajcd}e_{cd} + (\nabla'_a G_{mnab})C_{bjcd}e_{cd}] \\ &\quad\quad + [(C_{ajcd}\nabla'_b + C_{bjcd}\nabla'_a)\nabla'_j G_{mnab}]e_{cd}. \end{aligned} \tag{20}$$

The integral of $\mathbf{r}'$ over the volume of the body then gives, with the aid of Eq. (17) and Gauss' theorem,

$$\begin{aligned} 0 = &-e_{mn}(\omega,\mathbf{r}) + \int d^3r\, G_{mnab}(\omega,\mathbf{r},\mathbf{r}')s_{ab}(\omega,\mathbf{r}') \\ &-\tfrac{1}{2}\int dA'\,\{(\nabla'_b G_{mnab}C_{ajcd} + \nabla'_a G_{mnab}C_{bjcd})e_{cd}n_j(\mathbf{r}') \\ &-G_{mnab}(n_a(\mathbf{r}')C_{bjcd} + n_b(\mathbf{r}')C_{ajcd})e_{cd,j}\}. \end{aligned} \tag{21}$$

It is clear from Eq. (17) that we can choose

$$G_{mnab} = G_{nmab} = G_{mnba}. \tag{22}$$

Equation (21) is more concisely written in terms of the linearly elastic surface stress,

$$\sigma_{aj} = C_{ajcd} e_{cd}. \tag{23}$$

If the total stress at the surface is linearly elastic we have the further simplification

$$C_{ajcd} e_{cd,j}(\omega, \mathbf{r}') = -\rho\omega^2 u_a(\omega, \mathbf{r}'). \tag{24}$$

We use the notation $t_i(\omega, \mathbf{r}')$ for the surface tractions,

$$t_i(\omega, \mathbf{r}') = \sigma_{ij}(\omega, \mathbf{r}') n_j(\mathbf{r}'). \tag{25}$$

Then Eq. (21) simplifies to

$$\begin{aligned} e_{mn}(\omega, \mathbf{r}) = & \int d^3\mathbf{r}'\, G_{mnab}(\omega, \mathbf{r}, \mathbf{r}') s_{ab}(\omega, \mathbf{r}') \\ & - \int dA' \{ t_a(\omega, \mathbf{r}') \nabla'_b G_{mnab}(\omega, \mathbf{r}, \mathbf{r}') \\ & + \rho\omega^2 u_a(\omega, \mathbf{r}') n_b(\mathbf{r}') G_{mnab}(\omega, \mathbf{r}, \mathbf{r}') \}. \end{aligned} \tag{26}$$

Equation (26) expresses e_{mn}, which we will refer to as the linear strain, and which is equal to the invariant strain η_{mn} in most of the body, in terms of an internal inelastic source s_{ab}, the surface tractions t_a, and the surface displacements u_a. This provides a heuristic motivation for the boundary conditions needed to complete the definition of the Green's function. Since we usually want to specify the surface tractions, not the surface displacements, we will usually choose

$$G_{mnab}(\omega, \mathbf{r}, \mathbf{r}') n_b(\mathbf{r}') = 0 \tag{27}$$

when $\mathbf{r}'$ is on the surface. Because no internal stress can produce a traction on a free surface, we have the additional restriction

$$n_p(\mathbf{r}) C_{pqmn} G_{mnab}(\omega, \mathbf{r}, \mathbf{r}') = 0, \tag{28}$$

where $\mathbf{r}'$ is arbitrary and $\mathbf{r}$ is on the surface. We will see below that these are not independent. To meet this condition [Eq. (26)] it is convenient to define a different Green's function

$$H_{pqab} = C_{pqmn} G_{mnab} \tag{29}$$

which satisfies

$$n_p(\mathbf{r}) H_{pqab}(\omega, \mathbf{r}, \mathbf{r}') = 0 \tag{30}$$

when $\mathbf{r}$ is on the surface with $\mathbf{r}'$ arbitrary, and

$$H_{pqab}(\omega, \mathbf{r}, \mathbf{r}') n_b(\mathbf{r}') = 0 \tag{31}$$

where $\mathbf{r}'$ is on the surface for arbitrary $\mathbf{r}$. The simplicity of the boundary conditions proves to be an enormous asset.

Since the elastic compliance tensor S_{pqrs} satisfies

$$S_{pqrs} C_{rsmn} = \tfrac{1}{2}(\delta_{pm}\delta_{qn} + \delta_{pn}\delta_{qm}), \tag{32}$$

we can obtain the differential equation satisfied by H by multiplying Eq. (17) by $S_{abpq} C_{rscd}$ and summing over repeated indices. The highly symmetric result is

$$D^s_{mnpq}(\omega, \nabla') H_{rsmn}(\omega, \mathbf{r}, \mathbf{r}') = -\tfrac{1}{2}(\delta_{ps}\delta_{qr} + \delta_{pr}\delta_{qs})\delta^3(\mathbf{r} - \mathbf{r}') \tag{33}$$

with

$$\begin{aligned} D^s_{mnpq}(\omega, \nabla) &= D_{mnab}(\omega, \nabla) S_{abpq} \\ &= \rho\omega^2 S_{mnpq} + \tfrac{1}{2}(\delta_{mp}\nabla_n\nabla_q + \delta_{nq}\nabla_m\nabla_p \\ &\quad + \delta_{mq}\nabla_n\nabla_p + \delta_{np}\nabla_m\nabla_q). \end{aligned} \tag{34}$$

$\mathbf{D}^s$ is an operator which maps a physical stress tensor describing an acoustic emission into its internal source. Such a source cannot create a traction on a free surface, or, more generally, change the surface tractions if the experimenter chooses to hold them constant, so the natural domain of $\mathbf{D}^s$ is the space L of symmetric twice continously differentiable second-rank tensors σ_{ij} defined within the body and producing no surface tractions,

$$\sigma_{ij}(\mathbf{r}) n_j(\mathbf{r}) = 0, \tag{35}$$

for $\mathbf{r}$ on surface of the body. From this point we are going to consider only stress-specified (Dirichlet) boundary conditions. A clamped boundary requires Neumann boundary conditions on $\mathbf{H}$ and has a different natural domain for $\mathbf{D}$. If s_{mn} and σ_{ik} are members of L we have the identity (using $*$ for complex conjugation)

$$\begin{aligned} s^*_{mn}(\mathbf{r}) D^s_{mnik}(\omega, \nabla)\sigma_{ik}(\mathbf{r}) = & (D^s_{mnik}(\omega, \nabla) s^*_{mn}(\mathbf{r}))\sigma_{ik}(\mathbf{r}) \\ & + \nabla_n(s^*_{mn}\sigma_{mk,k})(\mathbf{r}). \\ & - \nabla_k(s^*_{mn,n}\sigma_{mk})(\mathbf{r}). \end{aligned} \tag{36}$$

Then, using the symmetry of D^s_{mnik} under $ik \leftrightarrow mn$, we obtain

$$\begin{aligned} & \int d^3r\, s^*_{mn}(\mathbf{r}) D^s_{mnik}(\omega, \nabla)\sigma_{ik}(\mathbf{r}) \\ & = \int d^3r\, (D_{ikmn}(\omega, \nabla) s^*_{mn}(\mathbf{r}))\sigma_{ik}(\mathbf{r}). \end{aligned} \tag{37}$$

We define an inner product on L by

$$(\mathbf{s}, \sigma) = \int d^3r\, s^*_{mn}(\mathbf{r})\sigma_{mn}(\mathbf{r}) \tag{38}$$

and define a metric from the inner product in the usual way. We then extend L to be a Hilbert space L_H by adding to it all limits of Cauchy sequences.[12] The second order elastic constants C_{pqrs} are defined by derivatives of a real free energy with respect to real invariant strains and are, therefore, real numbers. The compliance tensor $\mathbf{S}$ is, in consequence, a real self-adjoint operator on L_H. $\mathbf{D}^s$ is therefore a self-adjoint, although unbounded, operator on L_H. Note that the linear elliptic operator $\mathbf{K}$ defined by

$$K_{abcd}(\nabla) = \rho\omega^2 S_{abcd} - D^s_{abcd}(\omega, \nabla) \tag{39}$$

is also self-adjoint and is positive definite on L_H, so that its spectrum is confined to the positive real axis. Thus for $\mathrm{Im}(z) \neq 0$, z cannot be an element of the spectrum of $\mathbf{D}^s(z, \nabla)$ and $\mathbf{D}^s(z, \nabla)$ has a unique inverse $\mathbf{D}^s(z)^{-1}$ on L_H. This inverse is a bounded self-adjoint operator on L_H.

Consider the integral operator $\mathbf{H}$ on L_H whose kernel $H_{abcd}(z, \mathbf{r}, \mathbf{r}')$ satisfies Eq. (33) with $\omega = z$, $\mathrm{Im}z > 0$. This $\mathbf{H}$ satisfies

$$\mathbf{D}^s(z, \nabla)\mathbf{H} = -\mathbf{1} \tag{40}$$

and $-\mathbf{H}$ is a candidate for $D^s(z)^{-1}$. Because $D^s(z)^{-1}$ is self-adjoint, the kernel must satisfy

$$H_{abcd}(z, \mathbf{r}, \mathbf{r}') = H_{cdab}(z, \mathbf{r}', \mathbf{r})^*. \tag{41}$$

Because it is inverse to $\mathbf{D}^s(z,\nabla)$, it must also satisfy

$$\mathbf{H}\mathbf{D}^s(z,\nabla) = -\mathbf{1}. \tag{42}$$

In fact, Eq. (42) is the primary application of H, to calculate the dynamic stress distribution from a specifed source. We can show that Eq. (42) follows from Eq. (40), Eq. (41), and the boundary conditions Eq. (30) which require that H not generate a surface traction from a physical source. Let σ be an element of L. Then, since [Eq. (41)] H is self-adjoint,

$$(\sigma;\mathbf{H}\mathbf{D}^s\sigma) = (\mathbf{H}\sigma,\mathbf{D}^s\sigma) \tag{43}$$

and because neither σ nor $\mathbf{H}\sigma$ have surface tractions it follows from Eq. (37) that

$$(\sigma,\mathbf{H}\mathbf{D}^s\sigma) = (\mathbf{D}^s\mathbf{H}\sigma,\sigma) = -(\sigma,\sigma), \tag{44}$$

where the last step used Eq. (40). Thus for all σ in L it is true that

$$(\sigma,(\mathbf{H}\mathbf{D}^s+1)\sigma) = 0. \tag{45}$$

Equation (45) remains true for limits in L_H. Now the validity of Eq. (45) for all vectors σ in the complex Hibert space L_H is equivalent to the operator equality

$$\mathbf{H}\mathbf{D}^s + 1 = 0. \tag{46}$$

Thus the differential equation, self-adjointness, and the boundary conditions Eq. (30) are necessary and sufficient for

$$\mathbf{D}^s\mathbf{H} = \mathbf{H}\mathbf{D}^s = -1. \tag{47}$$

This important consequence, which could not have been anticipated on physical grounds, is that the boundary conditions Eq. (31) are not independent, but a consequence of Eq. (30) and self-adjointness, Eq. (41).

Since the spectrum of K on the positive real axis is not void, we must be careful in letting z approach the real axis. The physically interesting solution is obtained by approaching the real positive ω axis from above, $\mathrm{Im}(z) \to 0^+$, since with our convention for the time dependence it is this limit which results in outgoing waves. The limit is an operator limit

$$(\mathbf{H}^*\mathbf{s})_{ij}(\omega,\mathbf{r}) = \lim_{\mathrm{Im}z\to 0^+}\lim_{\mathrm{Re}z\to\omega}\int d^3r'\, H_{ijmn}(\mathrm{z},\mathbf{r},\mathbf{r}')s_{mn}(\omega,\mathbf{r}'). \tag{48}$$

The operators $\mathbf{D}^s$ and H are thus much better behaved than D (which is not self-adjoint) and G. The increased symmetry is, as we showed, a great simplication in satisfying the boundary conditions. If we let $\sigma_{mn}(\omega,\mathbf{r})$ denote the linear stress, Eq. (26) becomes

$$\begin{aligned}\sigma_{mn}(\omega,\mathbf{r}) = {} & \int d^3r'\, H_{mnab}(\omega,\mathbf{r},\mathbf{r}')s_{ab}(\omega,\mathbf{r}) \\ & - \int dA'\{\nabla'_b H_{mnab}(\omega,\mathbf{r},\mathbf{r}')t_a(\omega,\mathbf{r}') \\ & + \rho\omega^2 H_{mnab}(\omega,\mathbf{r},\mathbf{r}')n_b(\mathbf{r}')u_a(\omega,\mathbf{r}')\}.\end{aligned} \tag{49}$$

With our boundary conditions the second term in the surface integral vanishes. We are frequently interested in the surface displacements, which we obtain from Eq. (10),

$$u_n(\omega,\mathbf{r}) = -\sigma_{mn,n}(\omega,\mathbf{r})/(\rho\omega^2), \tag{50}$$

since the body force, if it is used at all, is localized within the body.

It is usually desirable to remove the derivatives from the source term s_{ab}, transforming the volume integral to

$$\begin{aligned}\int d^3r' & [(\nabla'_b\nabla'_k H_{mnab}(\omega,\mathbf{r},\mathbf{r}'))J_{ak}(\omega,\mathbf{r}') \\ & - \nabla'_b H_{mnab}(\omega,\mathbf{r},\mathbf{r}')F_a(\omega,\mathbf{r}')].\end{aligned} \tag{51}$$

One sees that the usual displacement field Green's function giving the displacement from a delta function body force is

$$g_{ma}(\omega,\mathbf{r},\mathbf{r}') = \nabla_n\nabla'_b H_{mnab}(\omega,\mathbf{r},\mathbf{r}')/(\rho_0\omega^2). \tag{52}$$

It is easy to obtain the radiation resulting from an initial nonzero strain distribution $e_{ab}(0,\mathbf{r})$ if one is willing to treat even the source region linearly. One uses the Laplace transform instead of the Fourier transform,

$$\int_0^\infty dt\, e^{-pt}e_{ab}(t,\mathbf{r}) = p^2\hat{e}_{ab}(p,\mathbf{r}) - pe_{ab}(0,\mathbf{r}) - \dot{e}_{ab}(0,\mathbf{r}). \tag{53}$$

The Green's function for this problem differs from H only by $\omega \to ip$, and we have seen that H is analytic in the upper-half ω plane (the right-half p plane), and the source has the additional term

$$\Delta s_{ab}(p,\mathbf{r}) = -pe_{ab}(0,\mathbf{r}) - \dot{e}_{ab}(0,\mathbf{r}). \tag{54}$$

After Fourier transforming the inverse Laplace transform, one obtains Eq. (49), but with an additional source term $\Delta s_{ab}(-i\omega,r)$.

3. ISOTROPIC INFINITE SOLID GREEN'S FUNCTION

In this section we solve Eq. (33) for the "stress transfer function," or Green's function, $H_{abcd}(\omega,\mathbf{r},\mathbf{r}')$ of an isotropic elastic medium without boundaries. This is the Green's function for the infinite isotropic medium, but its greatest importance to us is as a particular integral of Eq. (33), to be used in Sec. 4 to construct the Green's function for a solid of finite thickness in the "z" direction and infinite extent in the xy plane (a plate). It is possible to obtain the 81 components of this fourth-rank tensor integral operator by differentiation of the usual displacement field Green's function,[13] but it is more illuminating, and probably easier, to solve Eq. (33) directly. A considerable simplification is obtained by introducing a suitable operational calculus. We are concerned primarily with differential and integral operators on the space L_H, and these are fourth-rank tensors. They are conveniently expressed, however, by construction from second-rank tensor operators. For any two second-rank tensors A ond B we define the fourth-rank tensor $\mathbf{A}\otimes\mathbf{B}$ by

$$(\mathbf{A}\otimes\mathbf{B})_{abcd} = A_{ac}B_{bd} \tag{55}$$

and the dyad AB by

$$(\mathbf{AB})_{abcd} = A_{ab}B_{cd}. \tag{56}$$

We denote the operator product of two fourth-rank tensors M and N by $\mathbf{M}\times\mathbf{N}$ and obtain its components from

$$(\mathbf{M}\times\mathbf{N})_{abcd} = M_{abij}N_{ijcd}. \tag{57}$$

Similarly, for any two second-rank tensor A and B, $\mathbf{A}\times\mathbf{B}$ is a second-rank tensor with Cartesian components

$$(\mathbf{A}\times\mathbf{B})_{ij} = A_{ik}B_{kj}. \tag{58}$$

These definitions make multiplication of the fourth-rank tensors of interest much simpler. For example, we have

$$(\mathbf{A}\otimes\mathbf{B})\times(\mathbf{C}\otimes\mathbf{D})=(\mathbf{A}\times\mathbf{C})\otimes(\mathbf{B}\times\mathbf{D}) \tag{59}$$

and

$$(\mathbf{AB})\times(\mathbf{CD})=\mathbf{A}\,\mathrm{tr}(\tilde{\mathbf{B}}\mathbf{C})\mathbf{D}, \tag{60}$$

where

$$\tilde{B}_{ij}=B_{ji}. \tag{61}$$

The remaining multiplication rules for tensor and dyad products are given in Appendix A. Their advantage lies in the construction of projection operators. If P is a second-rank tensor projection, than $\mathbf{P}\times\mathbf{P}$ is a fourth-rank tensor projection,

$$(\mathbf{P}\otimes\mathbf{P})\times(\mathbf{P}\otimes\mathbf{P})=(\mathbf{P}\times\mathbf{P})\otimes(\mathbf{P}\times\mathbf{P})=\mathbf{P}\otimes\mathbf{P}. \tag{62}$$

The basic elements of our operational calculus are the second-rank tensor projection operators[14] P and Q defined by

$$P_{ij}=\nabla_i\nabla_j/\nabla^2 \tag{63}$$

and

$$Q_{ij}=\delta_{ij}-P_{ij}. \tag{64}$$

The elastic stiffness tensor C for an isotropic body is conveniently expressed in terms of the compressional (L) and shear (T) projections

$$L_{abcd}=\tfrac{1}{3}\delta_{ab}\delta_{cd} \tag{65}$$

and

$$T_{abcd}=\tfrac{1}{2}(\delta_{ac}\delta_{bd}+\delta_{ad}\delta_{bc}-\tfrac{2}{3}\delta_{ab}\delta_{cd}) \tag{66}$$

by

$$\mathbf{C}=3K\mathbf{L}+2\mu\mathbf{T}, \tag{67}$$

where K is the bulk modulus and μ the shear modulus. The compliance tensor S is then

$$\mathbf{S}=\frac{\mathbf{L}}{3K}+\frac{\mathbf{T}}{2\mu}. \tag{68}$$

By restricting our domain temporarily to the space L of *symmetric* second-rank twice differentiable tensors vanishing at infinity, we can simplify the solution of Eq. (33) by not requiring that our operators be symmetric under $xy\to yx$, etc. The operators we permit still map symmetric tensors into symmetric tensors, so the nonsymmetric component never contributes and can be eliminated at the end of the calculation, permitting us then to use a domain of six-component tensors instead of nine-component symmetric tensors.

Let U denote the second-rank tensor identity,

$$U_{ab}=\delta_{ab}, \tag{69}$$

and I the fourth-rank tensor identity,

$$\mathbf{I}=\mathbf{U}\otimes\mathbf{U}. \tag{70}$$

Note that the longitudinal projection L is given by

$$\mathbf{L}=\mathbf{UU}/3 \tag{71}$$

and the shear projection T by

$$\mathbf{T}=\mathbf{I}-\mathbf{L}. \tag{72}$$

The differential operator $\mathbf{D}^s$ which we are going to invert [Eq. (34)] is

$$D^s(\omega,\nabla)=\rho\omega^2\mathbf{S}+\tfrac{1}{2}\nabla^2(\mathbf{P}\otimes\mathbf{U}+\mathbf{U}\otimes\mathbf{P}). \tag{73}$$

After writing

$$\mathbf{U}=\mathbf{P}+\mathbf{Q} \tag{74}$$

and expanding the fourth-rank operators in Eq. (73) in terms of products of P and Q we have the easily invertible (Appendix A) expression

$$\mathbf{D}^s=\left(\frac{\rho\omega^2}{E}+\nabla^2\right)\mathbf{P}\otimes\mathbf{P}+\frac{1}{2}\left(\frac{\rho\omega^2}{\mu}+\nabla^2\right)(\mathbf{P}\otimes\mathbf{Q}+\mathbf{Q}\otimes\mathbf{P})$$
$$+\frac{\rho\omega^2}{2\mu}\mathbf{Q}\otimes\mathbf{Q}-\frac{\rho\omega^2}{E}(\mathbf{PQ}+\mathbf{QP}+\mathbf{QQ}), \tag{75}$$

where

$$E=9K\mu/(3K+\mu) \tag{76}$$

is Young's modulus and σ is Poisson's ratio,

$$\sigma=\tfrac{1}{2}(3K-2\mu)/(3K+\mu). \tag{77}$$

The operator inverse to Eq. (75) is $\mathbf{H}^\circ$,

$$\mathbf{H}^\circ=-[(1-\sigma)\mathbf{P}+\sigma\mathbf{Q}][(1-\sigma)\mathbf{P}+\sigma\mathbf{Q}]/[(1-\sigma)^2(\nabla^2+(\omega/c_t)^2)]$$
$$-2(\mathbf{P}\otimes\mathbf{Q}+\mathbf{Q}\otimes\mathbf{P})/(\nabla^2+(\omega/c_t)^2)$$
$$-2(c_t/\omega)^2[\mathbf{Q}\otimes\mathbf{Q}+\sigma\mathbf{QQ}/(1-\sigma)]. \tag{78}$$

It is useful to have an integral representation for the operator $\mathbf{H}^\circ$, so that the result of $\mathbf{H}^\circ$ operating on the second-rank tensor $s_{cd}(\omega,\mathbf{r}')$ is the second-rank tensor $\mathbf{H}^\circ\cdot\mathbf{s}(\omega,\mathbf{r})$ with a-b component

$$(\mathbf{H}^\circ\circ\mathbf{s})_{ab}(\omega,\mathbf{r})=\int d^3r'\,\mathbf{H}^\circ_{abcd}(\omega,\mathbf{r},\mathbf{r}')s_{cd}(\omega,\mathbf{r}'). \tag{79}$$

The Green's function is just the kernel of this integral operator. We can extract it by defining a tensor delta function source. Let c' and d' be fixed indices and $\mathbf{r}''$ some fixed point; then

$$f_{cd}(\mathbf{r}')=I_{cdc'd'}\delta^3(\mathbf{r}'-\mathbf{r}'') \tag{80}$$

is a second-rank tensor field and, from Eq. (79),

$$H^\circ_{abc'd'}(\omega,\mathbf{r},\mathbf{r}'')=(\mathbf{H}^\circ\cdot\mathbf{f})_{ab}(\omega,\mathbf{r}). \tag{81}$$

The representation is obtained by using the Fourier representation for f,

$$f_{cd}(\mathbf{r}')=I_{cdc'd'}\int\frac{d^3p}{(2\pi)^3}\exp[i\mathbf{p}\circ(\mathbf{r}'-\mathbf{r}'')]. \tag{82}$$

When this representation is inserted in Eq. (81) and the order of integrations is interchanged we have the formal operator given in Eq. (78) operating on $\exp(i\mathbf{p}\circ\mathbf{r}')$. This is simple, since, for example,

$$P_{ij}\exp(i\mathbf{p}\circ\mathbf{r}')=\frac{p_ip_j}{p^2}\exp(i\mathbf{p}\circ\mathbf{r}'). \tag{83}$$

We now symmetrize by replacing

$$H^\circ_{abcd}(\omega,\mathbf{r},\mathbf{r}')\to\tfrac{1}{2}(H^\circ_{abcd}(\omega,\mathbf{r},\mathbf{r}')+H^\circ_{abdc}(\omega,\mathbf{r},\mathbf{r}')). \tag{84}$$

This modifies only the $\mathbf{Q}\otimes\mathbf{Q}$ term and, as noted earlier, is unnecessary when operating on symmetric second-rank tensors, but permits us to consider only the six independent components of these tensors, reducing $\mathbf{H}^\circ$ from 9×9 to 6×6, a reduction which will be of great use in Sec. 4. Our result for the Green's function of the

infinite isotropic medium is then the kernel of $\mathbf{H}^\circ$,[15]

$$H^\circ_{abcd}(\omega,\mathbf{r},\mathbf{r}')=\int\frac{d^3p}{(2\pi)^3}M_{abcd}(\omega,\mathbf{p})e^{i\mathbf{p}\cdot(\mathbf{r}-\mathbf{r}')}, \tag{85}$$

where

$$\begin{aligned}M_{abcd}=&[p_ap_c(p^2\delta_{bd}-p_bp_d)+p_bp_d(p^2\delta_{ac}-p_ap_c)\\&+p_ap_d(p^2\delta_{bc}-p_bp_c)+p_bp_c(p^2\delta_{ad}-p_ap_d)]/\\&[p^4(p^2-\omega^2c_t^2)]+[(1-2\sigma)p_ap_b+\sigma p^2\delta_{ab}]\\&\times[(1-2\sigma)p_cp_d+\sigma p^2\delta_{cd}]/[(1-\sigma)^2p^4(p^2-\omega^2/c_l^2)]\\&-(c_t/\omega)^2[(p_ap_d/p^2-\delta_{ad})(p_cp_b/p^2-\delta_{bc})\\&+(p_ap_c/p^2-\delta_{ac})(p_dp_b/p^2-\delta_{bd})\\&+2\sigma(p_dp_b/p^2-\delta_{ab})(p_cp_d/p^2-\delta_{cd})/(1-\sigma)].\end{aligned} \tag{86}$$

The usual Green's function [Eq. (52)] is obtained quite easily from Eq. (86). Most terms vanish, and the rest are easily integrated to give

$$\begin{aligned}g_{mn}(\omega,R)=&[\nabla_m\nabla_nR^{-1}\exp(i\omega R/c_l)+(\delta_{mn}\nabla^2-\nabla_m\nabla_n)\\&\times R^{-1}\exp(i\omega R/c_t)]/(4\pi\rho\omega^2).\end{aligned} \tag{87}$$

The integral can be done analytically, but it is more useful for application to the plate problem to do only the integral over the z component of the wavevector p_a. There are really only two integrals needed:

$$\begin{aligned}\psi(q_\perp,z)=&\int\frac{dq}{2\pi}\frac{\exp(iqz)}{(q^2+q_\perp^2)^2}\\=&\{\theta(z)(1+zq_\perp)\exp(-q_\perp z)+\theta(-z)(1-zq_\perp)\exp(q_\perp z)\}/\\&(4q_\perp^3)\end{aligned} \tag{88}$$

and

$$\begin{aligned}\phi(\omega/c,q_\perp,z)=&\int\frac{dq}{2\pi}\frac{\exp(iqz)}{(q^2+q_\perp^2)^2}\frac{1}{(q^2+q_\perp^2-\omega^2/c^2}\\=&\frac{1}{2}\left(\frac{c}{\omega}\right)^4\left[\theta(z)\left(\frac{i\exp(i\alpha z)}{\alpha}-\frac{\exp(-q_\perp z)}{q_\perp}\right)\right.\\&\left.+\theta(-z)\left(\frac{i\exp(-i\alpha z)}{\alpha}-\frac{\exp(q_\perp z)}{q_\perp}\right)\right]-\frac{c^2}{\omega^2}\psi(q_\perp,z),\end{aligned} \tag{89}$$

where

$$\alpha=(\omega^2/c^2-q_\perp^2)^{1/2},\quad \mathrm{Im}\,\alpha>0, \tag{90}$$

and θ is the Heaviside step function. In numerical applications to the plate problem one needs ψ,ϕ, and their first four derivatives with respect to z.

4. GREEN'S FUNCTION FOR THE PLATE

We consider the calculation of $\mathbf{H}$ for an isotropic slab, infinite in the xy plane and of thickness d in the z direction. We take the free space Green's function $\mathbf{H}^\circ$ calculated in the previous section as a particular integral of Eq. (33) and search for a complementary function $R_{abcd}(\omega,\mathbf{r},\mathbf{r}')$ satisfying

$$\mathbf{D}^s(\omega,\nabla')\mathbf{R}=0 \tag{91}$$

for which $\mathbf{H}$,

$$\mathbf{H}=\mathbf{H}^\circ+\mathbf{R}, \tag{92}$$

satisfies [cf. Eq. (31)]

$$H_{abc3}(\omega;\mathbf{r};x',y',0)=H_{abc3}(\omega;\mathbf{r};x',y',d)=0. \tag{93}$$

Because there are no dyads in this section we can discontinue the use of the cumbersome "×" symbol and let juxtaposition denote the operator product.

We achieve a significant economy at this point by realizing that we can require

$$H_{abcd}=H_{bacd}=H_{abdc} \tag{94}$$

and introduce six-component notation, thereby reducing the 81 components of $\mathbf{H}$ to 36. Our matrices are more nearly diagonal if we make a slight departure from convention and define our basis by

$$xx=1,\quad yy=2,\quad xy=3,\quad xz=4,\quad yz=5,\quad zz=6. \tag{95}$$

The plate is translation invariant in the horizontal plane, so the Green's function depends not on the transverse coordinates $\mathbf{r}_\perp$ and $\mathbf{r}'_\perp$ separately, but only on their difference. We exploit this difference by Fourier transformation in the difference of transverse coordinates,

$$\mathbf{R}(\omega;\mathbf{r};\mathbf{r}')=\int\frac{d^2Q}{(2\pi)^2}\mathbf{F}(\omega;\mathbf{Q};z,z')\exp[i\mathbf{Q}(\mathbf{r}'_\perp-\mathbf{r}_\perp)]. \tag{96}$$

It is also conceptually and computationally useful to treat the 36 components of $\mathbf{F}$ as a column vector $\mathbf{F}_V$ whose nth component is

$$F_V(n)=F_{(n-1)\,\mathrm{div}6+1,\ (n-1)\,\mathrm{mod}6+1}, \tag{97}$$

where "div" and "mod" refer to integer division and remainder,

$$N=6(N\,\mathrm{div}6)+N\,\mathrm{mod}6. \tag{98}$$

The inverse relation is

$$F_{ij}=F_V(6(i-1)+j). \tag{99}$$

The matrix of $\mathbf{D}^s$ in the column vector representation is $\mathbf{D}^s\otimes\mathbf{1}$ with "1" denoting yet another identity matrix, the six by six identity matrix. The tensor product of second-rank tensors on a six-dimensional space is constructed by the same rule used for three-dimensional spaces, Eq. (55), and $\mathbf{D}^s\otimes\mathbf{1}$ is thirty-six by thirty-six, but many of its 1296 elements are zero. When operating on $\mathbf{F}_V$ it has the simple form

$$\mathbf{D}^s\otimes\mathbf{1}=\left(\mathbf{A}\frac{d^2}{dz^2}+\mathbf{B}\frac{d}{dz}+\mathbf{C}\right)\otimes\mathbf{1}, \tag{100}$$

where the six by six matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ (given explicitly in Appendix B) can be decomposed into three by three matrices:

$$\mathbf{A}=\begin{bmatrix}0&0\\0&\mathbf{A}_{22}\end{bmatrix}, \tag{101}$$

$$\mathbf{B}=\begin{bmatrix}0&\mathbf{B}_{12}\\\mathbf{B}_{21}&\mathbf{B}_{22}\end{bmatrix}, \tag{102}$$

$$\mathbf{C}=\begin{bmatrix}\mathbf{C}_{11}&\mathbf{C}_{12}\\\mathbf{C}_{21}&\mathbf{C}_{22}\end{bmatrix}. \tag{103}$$

Equation (91) can now be greatly simplified. Only components 19–36 of $\mathbf{F}_V$ satisfy a second-order differential equation in z, and only these components are constrained by the boundary condition Eq. (93). We want to eliminate the first 18 components of $\mathbf{F}_V$ from the differential equation, so we define a column vector $\mathbf{F}_U$ consisting of the

upper 18 components and $\mathbf{F}_L$ consisting of the second (lower) 18 components. Thus $\mathbf{F}_V$ is given by

$$\mathbf{F}_V = \mathbf{F}_U \oplus \mathbf{F}_L = \begin{bmatrix} \mathbf{F}_U \\ \mathbf{F}_V \end{bmatrix}. \tag{104}$$

We partition $\mathbf{D}^s \otimes \mathbf{1}$ into the corresponding 18 by 18 blocks,

$$\mathbf{D}^s \otimes \mathbf{1}\ \mathbf{F}_V = \begin{bmatrix} \mathbf{d}_{11} & \mathbf{d}_{12} \\ \mathbf{d}_{21} & \mathbf{d}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{F}_U \\ \mathbf{F}_L \end{bmatrix} = 0, \tag{105}$$

to obtain the two equations

$$\mathbf{F}_U = -\mathbf{d}_{11}^{-1}\mathbf{d}_{12}\mathbf{F}_L \tag{106}$$

and

$$(\mathbf{d}_{22} - \mathbf{d}_{21}\mathbf{d}_{11}^{-1}\mathbf{d}_{12})\mathbf{F}_L = 0. \tag{107}$$

Equation (107) is an 18 by 18 matrix second-order ordinary differential equation for the eighteen-component column vector $\mathbf{F}_L(z')$. The solution must satisfy the boundary conditions

$$\mathbf{F}_L(0) = -(\mathbf{H}_V^\circ(\omega, z, 0))_L \tag{108}$$

and

$$\mathbf{F}_L(d) = -(\mathbf{H}_V^\circ(\omega, z, d))_L. \tag{109}$$

The solution of Eq. (107) subject to the boundary conditions of Eqs. (108) and (109) can then be inserted into Eq. (106) ($\mathbf{d}_{11}$ is a numerical matrix) to obtain by differentiation and matrix multiplication the remaining components of $\mathbf{F}_V$. Equation (107) is far simpler than it seems. When written in terms of the three by three submatrices of **A**, **B**, and **C**, it becomes

$$\begin{aligned} &\left\{(\mathbf{A}_{22} + \mathbf{Q}\circ\mathbf{B}_{21}\mathbf{C}_{11}^{-1}\mathbf{Q}\cdot\mathbf{B}_{12})\frac{d^2}{dz^2} + i(\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right. \\ &\left. - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{B}_{12})\frac{d}{dz} + \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right\} \otimes \mathbf{1}\,\mathbf{F}_V \\ &= 0 \\ &= \left(\alpha\frac{d^2}{dz^2} + i\beta\frac{d}{dz} + \gamma\right) \otimes \mathbf{1}\,\mathbf{F}_V. \end{aligned} \tag{110}$$

The explicit form for these three by three matrices is given in Appendix B. This equation is easily solved if we define

$$\frac{d\mathbf{F}_L}{dz'} = \mathbf{F}_L^{(1)} \tag{111}$$

and

$$\mathbf{f} = \begin{bmatrix} \mathbf{F}_L \\ \mathbf{F}_L^{(1)} \end{bmatrix}. \tag{112}$$

Then the differential equation becomes

$$\frac{d\mathbf{f}}{dz'} = \mathbf{M} \otimes \mathbf{1}\mathbf{f}, \tag{113}$$

where M is a six by six constant matrix (depending on $\mathbf{Q}_i$) given in Appendix B. The solution to Eq. (113) is

$$\mathbf{f}(z') = \mathbf{U}(z') \otimes \mathbf{1}\mathbf{f}(0), \tag{114}$$

where the "evolution operator" $\mathbf{U}(z')$ is given by

$$\mathbf{U}(z') = \exp(z'\mathbf{M}). \tag{115}$$

Equation (114) expresses f in terms of $\mathbf{F}_L(0)$ and $\mathbf{F}_L^{(1)}(0)$. However, the boundary conditions specify $\mathbf{F}_L(0)$ and $\mathbf{F}_L(\mathbf{d})$, not $\mathbf{F}_L^{(1)}(0)$, so we use [from Eq. (114)]

$$\mathbf{F}_L(d) = \mathbf{U}_{11}(d) \otimes \mathbf{1}\mathbf{F}_L(0) + \mathbf{U}_{12}(d) \otimes \mathbf{1}\mathbf{F}_L^{(1)}(0) \tag{116}$$

to eliminate $\mathbf{F}_V^{(1)}(0)$,

$$\mathbf{F}_V^{(1)}(0) = \mathbf{U}_{12}(d)^{-1} \otimes \mathbf{1}(\mathbf{F}_L(d) - \mathbf{U}_{11}(d) \otimes \mathbf{1}\mathbf{F}_L(0)). \tag{117}$$

Of course, $\mathbf{U}_{ab}(z')^{-1}$ denotes the inverse of the three by three submatrix $\mathbf{U}_{ab}(z')$ of the six by six matrix $\mathbf{U}(z')$ and must not be confused with $(\mathbf{U}(z')^{-1})_{ab}$, which is the a-b three by three submatrix of the inverse of $\mathbf{U}(z')$.

We can now obtain explicit expression for $\mathbf{F}_L(z')$ and its first derivative,

$$\mathbf{F}_L(z') = [\mathbf{U}_{11}(z') - \mathbf{U}_{12}(z')\mathbf{U}_{12}(d)^{-1}\mathbf{U}_{11}(d)] \otimes \mathbf{1}(\mathbf{H}_V^\circ(\omega, \mathbf{Q}, z))_L \tag{118}$$

and

$$\begin{aligned} \mathbf{F}_L^{(1)}(z') = {} & [\mathbf{U}_{21}(z') - \mathbf{U}_{22}(z')\mathbf{U}_{12}(d)^{-1}\mathbf{U}_{11}(d)] \\ & \otimes \mathbf{1}(\mathbf{H}_V^\circ(\omega;\mathbf{Q};z))_L + (\mathbf{U}_{22}(z')\mathbf{U}_{12}(d)^{-1}) \\ & \otimes \mathbf{1}(\mathbf{H}_V^\circ(\omega;\mathbf{Q};z-d))_L. \end{aligned} \tag{119}$$

Thr remarkable feature of this solution of the thirty-six component second-order ordinary differential equation, Eq. (105), is that it involves at worst only six by six matrices, and most of the result is expressed in terms of three by three matrices. The solution is completed by obtaining the upper components $\mathbf{F}_U$ from Eq. (106),

$$\begin{aligned} \mathbf{F}_U(\omega;\mathbf{Q};z,z') = {} & -i(\mathbf{C}_{11}^{-1}\mathbf{Q}\cdot\mathbf{B}_{12}) \otimes \mathbf{1}\mathbf{F}_L^{(1)}(z') \\ & - (\mathbf{C}_{11}^{-1}\mathbf{C}_{12}) \otimes \mathbf{1}\mathbf{F}_L(z'), \end{aligned} \tag{120}$$

and the final result for the Green's function at the plate is

$$\begin{aligned} \mathbf{H}(\omega;\mathbf{r};\mathbf{r}') = {} & \int \frac{d^2Q}{(2\pi)^2} \exp[i\mathbf{Q}\circ(\mathbf{r}-\mathbf{r}')][\mathbf{H}_V^\circ(\omega;\mathbf{Q};z-z') \\ & + \mathbf{F}_V(\omega;\mathbf{Q};z;z')]. \end{aligned} \tag{121}$$

There are four distinct tensor Green's functions of interest; the basic two are those for stress or displacement with a stress source, but one may also be interested in the same two quantities with a body force source. These are obtained [Eq. (52)] by suitable derivatives of Eq. (121), all of which can be done analytically either by differentiating $\mathbf{H}^\circ$ with respect to z or by using Eq. (113). Of course the higher derivatives are obtained from, for example,

$$\frac{d^2 f}{dz'^2} = \mathbf{M}^2 \otimes \mathbf{1}\mathbf{f}. \tag{122}$$

Inspection of the resulting formula for $g_{ij}(\omega, \mathbf{r}, \mathbf{r}')$ reveals that the introduction of the stress Green's function $\mathbf{H}(\omega, \mathbf{r}, \mathbf{r}')$, motivated by the resulting simplicity of the boundary conditions, serves a profound role because of the simplicity of its evolution from the boundary [Eq. (114)].

5. CONCLUSION

The objective of the work reported here was to provide, within the limits of continuum mechanics, an analytical basis for the numerical calculations of spectrally analyzed acoustic emisssion signals. Subsequent

reports will deal with two kinds of numerical studies with these Green's functions. The first will apply it to simple models for various physical sources and seek to explain present experimental spectra.[5] An application which may eventually be more profitable is to explore the tensor and spatial distribution of signals from theoretically understood sources, in order to identify characteristics which are sufficiently different to warrant the construction of specialized transducers for nondestructive testing applications.

The reader interested in seismological applications should consult also the Pilant thesis.[16] Pilant calculated, as a sum of normal modes, the Green's function for a plate subjected to a uniform line body force parallel to its surface. The AE problem is rather different, in that such a line force is probably never realistic, and the observations are made in the far field region at high frequencies where many normal modes contribute.[17]

ACKNOWLEDGMENT

I am grateful to E. A. Kraut for suggesting improvements in the manuscript.

APPENDIX A

The fourth-rank tensor operators **D** and $\mathbf{D}^s$ can be conveniently expressed in terms of the simple second-rank tensor operators **P** and **Q** with Cartesian components

$$P_{ij} = \nabla_i \nabla_j / \nabla^2 \tag{A1}$$

and

$$Q_{ij} = \delta_{ij} - P_{ij}. \tag{A2}$$

Repeated indices are always summed from 1 to 3 unless stated otherwise. We use "×" to denote operator multiplication; for any two second-rank tensors **A** and **B**, **A**×**B** is a second rank tensor with components

$$(\mathbf{A}\times\mathbf{B})_{ij} = A_{ik}B_{kj}, \tag{A3}$$

and for any two fourth-rank tensors **M** and **N**, **M**×**N** is a fourth-rank tensor with components

$$(\mathbf{M}\times\mathbf{N})_{abcd} = M_{abij}N_{ijcd}. \tag{A4}$$

The projections **P** and **Q** satisfy

$$\mathbf{P}\times\mathbf{Q} = \mathbf{Q}\times\mathbf{P} = 0, \tag{A5}$$

$$\mathbf{P}\times\mathbf{P} = \mathbf{P}, \tag{A6}$$

$$\mathbf{Q}\times\mathbf{Q} = \mathbf{Q}. \tag{A7}$$

We can construct fourth-rank tensor operators from **P** and **Q** in two ways, by defining dyads, and from the tensor product. For any two second rank tensors **A** and **B** we can form a dyad **AB** with components

$$(\mathbf{AB})_{abcd} = A_{ab}B_{cd}. \tag{A8}$$

The tensor product $\mathbf{A}\otimes\mathbf{B}$ is a fourth-rank tensor with components

$$(\mathbf{A}\otimes\mathbf{B})_{abcd} = A_{ac}B_{bd}. \tag{A9}$$

The tensor product and dyad behave simply under multiplication,

$$(\mathbf{A}\otimes\mathbf{B})\times(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{A}\times\mathbf{C})\otimes(\mathbf{B}\times\mathbf{D}) \tag{A10}$$

and

$$(\mathbf{AB})\times(\mathbf{CD}) = \mathbf{A}\,\mathrm{tr}(\tilde{\mathbf{B}}\times\mathbf{C})\mathbf{D}, \tag{A11}$$

where

$$\tilde{B}_{ij} = B_{ji} \tag{A12}$$

is the transpose and

$$\mathrm{tr}\mathbf{B} = B_{ii}. \tag{A13}$$

The remaining multiplication rules are easily derived,

$$(\mathbf{A}\otimes\mathbf{B})\times(\mathbf{CD}) = (\mathbf{A}\times\mathbf{C}\times\tilde{\mathbf{B}})\mathbf{D}, \tag{A14}$$

$$(\mathbf{AB})\times(\mathbf{C}\otimes\mathbf{D}) = \mathbf{A}(\mathbf{C}\times\mathbf{B}\times\mathbf{D}). \tag{A15}$$

This operator calculus makes it easy to invert $\mathbf{D}^s$ in the absence of boundaries. In terms of dyads and tensor products of **P** and **Q**, $\mathbf{D}^s$ becomes

$$\begin{aligned}\mathbf{D}^s = {} & (\rho\omega^2/E + \nabla^2)\,\mathbf{P}\otimes\mathbf{P} + \tfrac{1}{2}(\rho\omega^2/c_t^2 + \nabla^2)(\mathbf{P}\otimes\mathbf{Q} + \mathbf{Q}\otimes\mathbf{P}) \\ & + \tfrac{1}{2}(\omega^2/c_t^2)\,\mathbf{Q}\otimes\mathbf{Q} - \sigma(\rho\omega^2/E)(\mathbf{PQ} + \mathbf{QP} + \mathbf{QQ}).\end{aligned} \tag{A16}$$

We look for a solution to

$$\begin{aligned}\mathbf{D}^s\times\mathbf{H} = -\mathbf{I} &= -(\mathbf{P}+\mathbf{Q})\otimes(\mathbf{P}+\mathbf{Q}) \\ &= -(\mathbf{P}\otimes\mathbf{P} + \mathbf{P}\otimes\mathbf{Q} + \mathbf{Q}\otimes\mathbf{P} + \mathbf{Q}\otimes\mathbf{Q})\end{aligned} \tag{A17}$$

of the form

$$\begin{aligned}\mathbf{H} = {} & -2\frac{(\mathbf{P}\otimes\mathbf{Q} + \mathbf{Q}\otimes\mathbf{P})}{\omega^2/c_t^2 + \nabla^2} - 2\frac{\mathbf{Q}\otimes\mathbf{Q}}{\omega^2/c_t^2} \\ & + a_1\,\mathbf{P}\otimes\mathbf{P} + a_2\,\mathbf{QP} + b_1\,\mathbf{PQ} + b_2\mathbf{QQ}.\end{aligned} \tag{A18}$$

Equation (A17) can be easily multiplied out, producing four equations for the scalar coefficients. We note that for the operator **P** (only!)

$$\mathbf{P}\otimes\mathbf{P} = \mathbf{PP}. \tag{A19}$$

The four equations are

$$(\rho\omega^2/E + \nabla^2)\,a_1 - 2\sigma\rho\omega^2 a_2/E = -1 \tag{A20}$$

from the coefficient of $\mathbf{P}\times\mathbf{P}$,

$$(\rho\omega^2/E)(\tfrac{1}{2}E/\mu - 2\sigma)\,a_2 - \sigma\rho\omega^2\,a_1/E = 0 \tag{A21}$$

from the coefficient of **QP**,

$$(\rho\omega^2/E + \nabla^2)\,b_1 - 2\sigma\rho\omega^2 b_2/E = -2\sigma\mu/E \tag{A22}$$

from the coefficient of **PQ**, and

$$(\rho\omega^2/E)(\tfrac{1}{2}E/\mu - 2\sigma)\,b_2 - \sigma\rho\omega^2 b_1/E = -2\sigma\mu/E \tag{A23}$$

form the coefficient of **QQ**. Using the relations

$$E/\mu = \tfrac{1}{2}(1+\sigma) \tag{A24}$$

and

$$\rho(1+\sigma)(1-2\sigma)/(E(1-\sigma)) = c_l^{-2} \tag{A25}$$

one easily obtains

$$a_1 = -1/(\nabla^2 + \omega^2/c_l^2), \tag{A26}$$

$$a_2 = \sigma a_1/(1-\sigma) = b_1, \tag{A27}$$

$$b_2 = -(\sigma/(1-\sigma))^2/(\nabla^2 + \omega^2/c_l^2) - 2\sigma c_t^2/((1-\sigma)\omega^2). \tag{A28}$$

APPENDIX B

Expressing $\mathbf{D}^3$ in terms of derivatives with respect to z with coefficient polynomials in $\mathbf{Q}$ is a task which can be done analytically, but since it really calls only for a series of logical decisions and simple counting, it is quickly and reliably done by a computer. These results are conveniently expressed in terms of three by three submatrices of the six by six matrices in Eq. (100). That is, one writes

$$\mathbf{A}=\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix}, \tag{B1}$$

where the A_{ij} are three by three matrices. The results are

$$\mathbf{A}_{11}=\mathbf{A}_{12}=\mathbf{A}_{21}=0, \tag{B2}$$

$$\mathbf{A}_{22}=\begin{bmatrix}\frac{1}{4} & 0 & 0\\ 0 & \frac{1}{4} & 0\\ 0 & 0 & 1\end{bmatrix}, \tag{B3}$$

$$\mathbf{B}_{11}=0, \tag{B4}$$

$$\mathbf{B}_{12}=\tilde{\mathbf{B}}_{21}=\begin{bmatrix}\frac{1}{2}Q_x & 0 & 0\\ 0 & \frac{1}{2}Q_y & 0\\ \frac{1}{2}Q_y & \frac{1}{4}Q_x & 0\end{bmatrix}, \tag{B5}$$

$$\mathbf{B}_{22}=\begin{bmatrix}0 & 0 & \frac{1}{2}Q_x\\ 0 & 0 & \frac{1}{2}Q_y\\ \frac{1}{2}Q_x & \frac{1}{2}Q_y & 0\end{bmatrix}, \tag{B6}$$

$$\mathbf{C}_{12}=\tilde{\mathbf{C}}_{21}=-\frac{\rho\omega^2\sigma}{E}\begin{bmatrix}0 & 0 & 1\\ 0 & 0 & 1\\ 0 & 0 & 0\end{bmatrix}, \tag{B7}$$

$$\mathbf{C}_{11}=\begin{bmatrix}\rho\omega^2/E-Q_x^2 & -\rho\omega^2\sigma/E & -\frac{1}{2}Q_xQ_y\\ -\rho\omega^2\sigma/E & \rho\omega^2/E-Q_y^2 & -\frac{1}{2}Q_xQ_y\\ -\frac{1}{2}Q_xQ_y & -\frac{1}{2}Q_xQ_y & \frac{1}{4}(\omega^2/c_t^2-Q^2)\end{bmatrix}, \tag{B8}$$

and

$$\mathbf{C}_{22}=\frac{1}{4}\begin{bmatrix}\omega^2/c_t^2-Q_x^2 & -Q_xQ_y & 0\\ -Q_xQ_y & \omega^2/c_t^2-Q_x^2 & 0\\ 0 & 0 & 4\rho\omega^2/E\end{bmatrix}. \tag{B9}$$

The mass density is ρ, E is Young's modulus, σ is Poisson's ratio, and the transverse velocity c_t given by

$$c_t=(\mu/\rho)^{1/2} \tag{B10}$$

with μ the shear modulus. The determinant of $\mathbf{C}_{11}$ is given by

$$\det\mathbf{C}_{11}=\tfrac{1}{4}(\rho\omega^2/E)J, \tag{B11}$$

where

$$J=(\omega^2/c_t^2-Q^2)(\omega^2/v_l^2-Q^2) \tag{B12}$$

with[18]

$$v_l=(\rho(1-\sigma^2)/E)^{-1/2} \tag{B13}$$

the velocity of a longitudinal plate plane wave; that is, a longitudinal wave with displacement independent of z. The matrix $\boldsymbol{\alpha}$ of Eq. (110) of the text is then

$$\alpha=\begin{bmatrix}\alpha' & & 0\\ & & 0\\ 0 & 0 & 1\end{bmatrix}, \tag{B14}$$

where

$$\alpha'=\frac{(\omega/c_t)^2}{4J}\times\begin{bmatrix}\frac{1}{2}(1-\sigma)((\omega/c_t)^2-Q_x^2)-Q_y^2 & \frac{1}{2}(1+\sigma)Q_xQ_y\\ \frac{1}{2}(1+\sigma)Q_xQ_y & \frac{1}{2}(1-\sigma)((\omega/c_t)^2-Q_y^2)-Q_x^2\end{bmatrix}. \tag{B15}$$

The most useful matrices are not β and γ, but $\alpha^{-1}\beta$ and $\alpha^{-1}\gamma$ given by

$$\alpha^{-1}\beta=\frac{2}{1+\sigma}\,\frac{1-(QV_L/\omega)^2}{(\omega/v_l)^2-Q^2}\times\begin{bmatrix}0 & 0 & (\rho\omega^2/E-Q^2)Q_x\\ 0 & 0 & (\rho\omega^2/E-Q^2)Q_y\\ \frac{1}{4}\rho\omega^2Q_x/E & \frac{1}{4}\rho\omega^2Q_y/E & 0\end{bmatrix}, \tag{B16}$$

where the symbol V_L has been used for

$$V_L=(2\mu/\rho)^{1/2}, \tag{B17}$$

the Lame velocity[19] (the velocity of a pure shear vertical plate mode), and

$$\alpha^{-1}\gamma=\begin{bmatrix}\Xi & & 0\\ & & 0\\ 0 & 0 & \xi\end{bmatrix} \tag{B18}$$

with

$$\Xi=\frac{2}{1-\sigma}(1-(QV_L/\omega)^2)\begin{bmatrix}(\omega/V_p)^2-Q_x^2 & -Q_xQ_y\\ -Q_xQ_y & (\omega/V_p)^2-Q_y^2\end{bmatrix} \tag{B19}$$

and

$$\xi=\frac{(\omega/c_l)^2-Q^2}{1-(Q_p/\omega)^2} \tag{B20}$$

where the "plate velocity" V_p is given by

$$V_p^{-2}=\tfrac{1}{2}\rho(1-\sigma)/\mu. \tag{B21}$$

The plate velocity is the low frequency limiting velocity of the first longitudinal mode. These velocities are introduced largely to simplify the formulas. The six by six infinitesimal generator M is given by

$$\mathbf{M}=\begin{bmatrix}0 & -1\\ \alpha^{-1}\gamma & i\alpha^{-1}\beta\end{bmatrix}. \tag{B22}$$

[1]A.A. Pollock, in *Acoustics and Vibration Progress* 1, edited by R.B.W. Stephens and H.G. Leventhall (Chapman and Hall, London, Wiley, New York, 1974).
[2]K.E. Stahlkopf and G.J. Dau, Nuclear Safety **17**, 33 (1976).
[3]R.G. Liptai, D.O. Harris, R.B. Engle, and C.A. Tatro, Int. J. Nondestructive Test. **3**, 215 (1971).
[4]W.J. Pardee, "Theoretical Analysis of Acoustic Emission Spectra," ARPA/AFML Review of Quantitative NDE, July 15—17, 1975, ARPA Technical Report AFML-TR-75-212, 667 (1975).
[5]W.J. Pardee and L.J. Graham, 1975 Ultrasonics Symposium Proceedings, IEEE Cat. #75 CHO 994-4SU, 597 (1975).
[6]W.J. Pardee and L.J. Graham, "Two Types of Simulated Acoustic Emission," submitted to J. Acoust. Soc. Am.
[7]L.J. Graham and G.A. Alers, Mater. Eval. **XXXII**, 31 (1974).

[8]L. Knopoff and F. Gilbert, Bull. Seismol. Soc. Am. **50**, 117 (1960).
[9]The small amplitude of the AE signal does not imply the adequacy of linear elasticity. In many applications the entire medium is strained. However, the additional mathematical problems encountered when the medium is in a state of finite strain are not intrinsic to the phenomenon.
[10]C. Truesdell and R. Toupin, *Handbuch der Physik*, edited by S. Flügge (Springer, Berlin, 1960), Vol. III/1, p. 226.
[11]The same result can be obtained by extending the essentially geometrically proof given by R. N. Thurston, *Physical Acoustics*, edited by W. P. Mason (Academic, New York, 1964), Vol. I, Part A, Chap. 1, pp. 91ff, to include dissipative contributions to **T**.
[12]There is a host of adequate references to the properties of Hilbert space. The basic properties are adequately described in either F. Riesz and B. Sz. Nagy, *Functional Analysis*, translated by L. F. Boron (Ungar, New York, 1955), p. 197ff or N. Dunford and J. T. Schwartz, *Linear Operators*, *Part* 1 (Interscience, New York, 1966), p. 242ff.
[13]P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, II (McGraw-Hill, New York, 1953), p. 1779.
[14]I am indebted to J. M. Richardson for introducing me to **P** and **Q**.
[15]The expression for M_{abcd} given in Ref. 5 contains two typographical errors; the expression in Ref. 6 is Eq. (86) before symmetrization [Eq. (84)].
[16]W. Pilant, Ph. D. Thesis, Physics Dept., UCLA (1960).
[17]The lowest frequency mode for a $\frac{1}{2}$ in.-thick steel plate occurs at 0.2 MHz; there are four families of modes. Experimental results are currently frequency analyzed from 0.1 to 3.0 MHz.
[18]L. D. Landau and E. M. Lifshitz, *Theory of Elasticity*, Course of Theoretical Physics **7**, transl. by J. B. Sykes and W. H. Reid (Pergamon, Oxford and New York, 1970), 2nd ed., p. 114.
[19]T. R. Meeker and A. H. Meitzler, in *Physical Acoustics*, edited by W. P. Mason (Academic, New York, 1964), Vol. I, Part A, Chap. 2, p. 126.

Ergodic state with least mean recurrence time in Markovian kinetics

Koichiro Matsuno

Central Research Laboratories, Nippon Electric Co., Ltd., Kawasaki 213, Japan
(Received 1 October 1976)

We present a proof of the theorem that the ergodic state with the least mean recurrence time in an irreducible Markov process having countable states is identical, with probability unity, to the state with the least irreversible decay rate if neither microscopic reversibility nor the doubly stochastic property holds.

1. INTRODUCTION

We reported in a previous article[1] (referred to as I hereafter) a proof of the theorem that the ergodic state with the least mean recurrence time in an irreducible Markov process with countable states is the state with the least irreversible decay rate if neither microscopic reversibility nor the doubly stochastic property holds. Since then, it has come to our attention that the proof developed in I is incomplete[2] in some respect. The aim of this note is to present a proof of an almost identical theorem to the previous one and to correct the incompleteness found in I.

The theorem we try to prove is as follows:

Theorem A: Suppose that a Markov process having countable states $\{i\}$ $(i=1,2,\cdots)$, each of which is an ergodic state, follows the Pauli master equation

$$\frac{\partial}{\partial t}P(i,t)=\sum_{j\neq i}W(i\leftarrow j)P(j,t)-\sum_{j\neq i}W(j\leftarrow i)P(i,t) \tag{1.1}$$

with the set of transition probability rates $\{W(i\leftarrow j)\}$ each of which is independent of time, where $P(i,t)$ is the probability density of the state i at time t and $W(i\leftarrow j)$ is the probability of transition from j to i per unit time. Then, it occurs with probability unity that the mean recurrence time $\tau(i\rightarrow i)$ of the state i has its minimum value at the state $i=i_m$ having the least irreversible decay rate

$$i_m\equiv(i\,|\min\{W(i)\}) \tag{1.2}$$

with

$$W(i)\equiv\sum_{j\neq i}W(j\leftarrow i) \tag{1.3}$$

if the auxiliary condition

$$\tilde{W}(i_m)-W(i_m)>0 \tag{1.4}$$

with

$$\tilde{W}(i)\equiv\sum_{j\neq i}W(i\leftarrow j) \tag{1.5}$$

is satisfied.

2. PROOF

We consider time evolution of a sample process following (1.1) over the infinite time interval $(-\infty,\infty)$ and introduce a quantity

$$\mu[i;T\,|\,t_{in},\tau_0]\equiv[T]/T \tag{2.1}$$

as in I. The meanings of symbols are as follows: We call a unit-continuous event (UCE) of the state i the event that the state i is continuously realized from the moment the transition to the state i from any other state $j(\neq i)$ occurs till the moment the state i again undergoes the transition from i to any other $j(\neq i)$; $[T]$ is the total time duration of UCE's of the state i appearing during the time interval $(t_{in},t_{in}+T)$. Here, the time duration of UCE of i is counted as zero if it is less than an appropriate time unit τ_0. $[T]$ is consecutively counted by units of τ_0 in a single sequence of an interval T from an arbitrary initial time $t=t_{in}$, while, if two successive UCE's of any j with the intermission of the interval τ' occur with the time durations $\tau_1(>0)$ and $\tau_2(>0)$, respectively, and if $\tau'<\tau_0$, the successive events are regarded as if a single UCE of j with the time duration $(\tau_1+\tau'+\tau_2)$ had occurred.

Next, we introduce another quantity $P_c(j\,|\,i;T\,|\,t_{in})$ for a sample process appearing over the interval $(-\infty,\infty)$ as follows:

$$P_c(j\,|\,i;T\,|\,t_{in})\equiv\psi(j;t_{in},t_{in}+T)P_j(T)/P_i(T) \tag{2.2}$$

with

$$P_i(T)\equiv\exp[-W(i)T], \tag{2.3}$$

where $\psi(j;t_{in},t_{in}+T)=1$ if and only if a local sample process appearing over the interval $(t_{in},t_{in}+T)$ realizes the state j without undergoing any transition during the same interval, and $\psi(j;t_{in},t_{in}+T)\equiv 0$ otherwise. The quantity $P_i(T)$ is the probability of the event that the state i is continuously realized over the interval T under the condition that the state initially found is i. When $\psi(j;t_{in},t_{in}+T)=1$, the asymptotic relation

$$\lim_{T\to\infty}P_c(j\,|\,i;T\,|\,t_{in})=0 \tag{2.4}$$

holds if

$$W(i)<W(j). \tag{2.5}$$

This result tells that the event $\psi(j;t_{in},t_{in}+T)=1$ would occur with probability zero compared to the event $\psi(i;t'_{in},t'_{in}+T)=1$ in the limit $T\to\infty$, that is to say, the event $\psi(i;t'_{in},t'_{in}+T)=1$ would occur with probability unity compared to the event $\psi(j;t_{in},t_{in}+T)=1$ in the limit $T\to\infty$, where either $t'_{in}>t_{in}+T$ or $t_{in}>t'_{in}+T$ is satisfied.

Since each state of $\{i\}$ is supposed to be ergodic, it follows that

$$\tau(i\rightarrow i)<\infty \tag{2.6}$$

and

$$\tau(i\longrightarrow i)\geq 0 \tag{2.7}$$

 Copyright © 1977 American Institute of Physics

for any i, where $\tau(i \longrightarrow i)$ is the average holding time of the state i. Hence, one readily notes that the state

$$k_m \equiv (i \,|\, \min\{\tau(i \to i)\}) \tag{2.8}$$

with the least mean recurrence time of the sample process continuing over $(-\infty, \infty)$ will lead to

$$\lim_{T\to\infty} \mu[k_m; T \,|\, t_{in}, \tau_0] = 1 \quad \text{for any } t_{in} \tag{2.9}$$

if both the conditions

$$\tau_0 > \tau(k_m \to k_m) \tag{2.10}$$

and

$$\tau(k_m \longrightarrow k_m) > 0 \tag{2.11}$$

are satisfied. One can fix the arbitrary parameter τ_0 so as to satisfy (2.10) because of (2.6). However, the inequality of (2.11) is not self-evident under the premise (2.7).

We next present a few lemmas.

Lemma A: Given a sample process continuing over the infinite time interval $(-\infty, \infty)$ following the Markov process (1.1), then, if

$$\mu[j;\ T \,|\, t_{in}, \tau_0] = 1 \quad \text{for any } t_{in} \text{ and } T \tag{2.12}$$

holds with respect to the sample process, there exists a certain t'_{in} such that

$$P_c(j \,|\, i;\ T \,|\, t'_{in}) \neq 0. \tag{2.13}$$

Proof: The probability of the event that the state j occurs without undergoing any transition during an interval T from an arbitrary initial time t_{in} under the condition that the state found at $t = t_{in}$ is j is

$$P_j(T) = \exp[-W(j)T], \tag{2.14}$$

which never vanishes for finite T. The number of recurrences to any state j over the interval $(-\infty, \infty)$ exceeds any finite value because of (2.6). Hence, one can find a certain t'_{in} such that the state j is continuously realized over the interval $(t'_{in}, t'_{in} + T)$ because of $P_j(T) \neq 0$. If it never occurred that the state j keeps itself during a finite T in the interval $(-\infty, \infty)$, the probability $P_j(T)$ would have to vanish. This leads to the conclusion that there exists t'_{in} such that

$$\psi(j; t'_{in}, t'_{in} + T) = 1. \tag{2.15}$$

Lemma B: Suppose a set of independent sample processes $\{\Omega_l\}$ $(l = 1, 2, \cdots)$ each of which could continue over the infinite time interval $(-\infty, \infty)$ and further suppose that only one process out of $\{\Omega_l\}$ is allowed on the real-time axis over the interval $(-\infty, \infty)$ with probability unity compared to other processes. Then, if a local sample process $\omega_n(t_1, t_2)$, which is a local part of Ω_n over the interval (t_1, t_2), occurs with probability unity compared to a set of other local processes $\omega_m(t_1, t_2)$ as local parts of Ω_m over (t_1, t_2) for all $m(\neq n)$, the sample process Ω_n will occur with probability unity compared to other processes $\Omega_{m(\neq n)}$ on the real-time axis over the interval $(-\infty, \infty)$.

Proof: If a sample process Ω_x occurs with probability unity compared to other processes on the real-time axis, the intermediate process $\omega_x(t_1, t_2)$ of Ω_x has to occur, also, with probability unity. Since $\omega_n(t_1, t_2)$ of Ω_n is supposed to occur with probability unity and since only one process out of $\{\Omega_l\}$ is allowed with probability unity, Ω_x must be Ω_n.

Proof of Theorem A: Suppose a set of independent sample processes $\{S_i(T; T_0)\}$ $(i = 1, 2, \cdots)$, in which each of $S_i(T; T_0)$ could continue over the time interval $(-T_0, T_0)$ being subject to

$$\mu[i; T \,|\, t_{in}, \tau_0] = 1 \tag{2.16}$$

for any t_{in} satisfying

$$-T_0 < t_{in} < t_{in} + T < T_0. \tag{2.17}$$

Since the state $i = k_m$ with the least mean recurrence time satisfies the asymptotic relation

$$\lim_{T\to\infty}\left(\lim_{T_0\to\infty} \mu[k_m; T \,|\, t_{in}, \tau_0]\right) = 1 \tag{2.18}$$

because of (2.9) if the inequality of (2.11) holds, the event giving (2.18) coincides with the sample process $S_i(\infty;\infty)$ with $i = k_m$, in which it is understood that

$$S_i(\infty;\infty) \equiv \lim_{T\to\infty}\left(\lim_{T_0\to\infty} S_i(T; T_0)\right) \quad \text{for any } i \tag{2.19}$$

when both the limits of T and T_0 are taken. For the time being, we suppose the inequality of (2.11). Its justification will be given at the final stage of this proof. The inequality of (2.11) yields that the sample process $S_i(\infty;\infty)$ with $i = k_m$ certainly occurs on the real-time axis. Other sample processes $S_j(\infty;\infty)$ with $j \neq k_m$ are never allowed on the real-time axis; otherwise the state with the least mean recurrence time could not be defined.

Next, given a set of independent sample processes $\{S_i(T;\infty)\}$ $(i = 1, 2, \cdots)$, one can always find t_i such that

$$\psi(i; t_i, t_i + T) = 1 \tag{2.20}$$

for each process $(i = 1, 2, \cdots)$ due to Lemma A. Since each of $\{S_i(T;\infty)\}$ is independent of others, one can shift the origins of time axis in such a way that

$$t_1 = t_2 = \cdots = t \tag{2.21}$$

without loss of generality. Hence, the local sample process $s_i(t, t + T)$, which is a local part of $S_i(T;\infty)$ over the interval $(t, t + T)$, gives

$$P_c(i \,|\, k; T \,|\, t) = P_i(T)/P_k(T). \tag{2.22}$$

If the limit $T \to \infty$ is taken, the local sample process $s_i(t, t + T)$ with $i = i_m$ would occur with probability unity compared to other processes $s_j(t, t + T)$ $(j \neq i_m)$ because of the asymptotic relation

$$\lim_{T\to\infty} P_c(j(\neq i_m) \,|\, i_m; T \,|\, t) = 0 \tag{2.23}$$

[cf. (1.2) and (2.4)]. As a result, if one and only one process out of $\{S_i(\infty;\infty)\}$ is allowed on the real-time axis with probability unity compared to the remaining other processes, such a process has to be $S_i(\infty;\infty)$ with $i = i_m$ because of Lemma B. The candidate for the sample process $S_{k_m}(\infty;\infty)$, which certainly occurs over the interval $(-\infty, \infty)$ if (2.11) holds, is found to be $S_{i_m}(\infty;\infty)$ with probability unity.

Hence, a necessary prerequisite for that the sample process $S_{i_m}(\infty;\infty)$ may become a candidate for the event $S_{k_m}(\infty;\infty)$ is the inequality of average holding time

$$\tau(i_m \longleftarrow i_m) > 0 \tag{2.24}$$

due to (2.11). If $\tau(i_m \longleftarrow i_m) = 0$ held, it would follow that

$$\lim_{T\to\infty}\left(\lim_{T_0\to-\infty}\mu[i_m;T\,|\,t_{in},T_0]\right)=0 \quad \text{for any } t_{in} \tag{2.25}$$

with respect to $S_{i_m}(T;T_0)$ because of the definition (2.1). The present consideration leads us to the fact that the sample process $S_{k_m}(\infty;\infty)$ will be identical to $S_{i_m}(\infty;\infty)$ with probability unity if (2.24) is proved, since the inequality of (2.11) guarantees the occurrence of $S_{k_m}(\infty;\infty)$. When one estimates the average holding time $\tau(i \longleftarrow i)$ of the state i, the balancing between both the incoming and the outgoing transitions of i must be investigated. If the *a priori* probabilities of expecting each state are $\{\alpha_i\}$ $(i=1,2,\cdots)$, the probability rate of transition from i to any other state per unit time will be

$$q_i = \alpha_i W(i). \tag{2.26}$$

Similarly, the probability current of transition from any state other than i to i during one recurrence cycle of i will be

$$\tilde{q}_i = \exp[-W(i)\tau(i \longleftarrow i)]\sum_{j\neq i}\left\{\frac{\alpha_j}{\sum_{k\neq i}\alpha_k}W(i \leftarrow j)\right\}. \tag{2.27}$$

The balancing between the incoming and the outgoing probability currents of transition gives

$$q_i = \tilde{q}_i \tag{2.28}$$

because of the conservation of probability. Since the *a priori* probabilities of expectation $\{\alpha_i\}$ are shown to be uniform if only Markov process is concerned,[3] expression (2.28) leads to

$$W(i) = \exp[-W(i)\tau(i \longleftarrow i)]\tilde{W}(i)/\sum_{k\neq i}\alpha_k. \tag{2.29}$$

Expression (2.29) implies

$$\tau(i \longleftarrow i) \geq [1/W(i)]\ln[\tilde{W}(i)/W(i)], \tag{2.30}$$

where the equality is the case if the number of the states $\{i\}$ approaches infinity. Consequently, the inequality

$$\tau(i_m \longleftarrow i_m) > 0 \tag{2.24}$$

certainly holds under the premise (1.4). This completes the prooof of Theorem A.

3. CONCLUDING REMARKS

The theorem we have proved is that the ergodic state with the least mean recurrence time in an irreducible Markov process having countable states is identical, with probability unity, to the state with the least irreversible decay rate if neither microscopic reversibility nor the doubly stochastic property holds. The term *with probability unity* is incorrectly missing in the old theorem quoted in I. We, however, maintain that discussions and concluding remarks made in the previous article I remain unchanged.

[1]K. Matsuno, J. Math. Phys. 15, 1950 (1974).
[2]R. Kubo, Math. Rev. 51, 1710 (1976).
[3]K. Matsuno, J. Math. Phys. 16, 604 (1975).

Generating exactly soluble nonlinear discrete evolution equations by a generalized Wronskian technique

S. -C. Chiu* and J. F. Ladik†

The Rockefeller University, New York, New York 10021
(Received 20 September 1976)

By means of generalized Wronskian relations an operator formulation is developed allowing one to generate a class of discrete evolution equations which are soluble by inverse scattering. This class includes nonlinear difference–difference equations as well as nonlinear differential–difference equations. The development of the method exhibits the significance of the dispersion relation of the associated linearized equation. The generalized Wronskian technique also permits one to obtain generalized Bäcklund transformations.

I. INTRODUCTION

It is well known that nonlinear difference equations have many important physical applications. For example, nonlinear differential—difference equations have been used to model Ladder type electric circuits[1] and vibration of particles,[2] and nonlinear difference—difference equations have been used as computer models of their continuous counterparts.[3] Until recently there has been no theoretical approach which could be utilized to completely analyze these nonlinear problems. However, since the discovery of the technique of inverse scattering by Gardner, Greene, Kruskal, and Miura[4,5] in their classic work on the Korteweg—de Vries equation (KdV), a nonlinear extension of Fourier analysis has been developed. This method is applicable to a class of both continuous[6] and discrete[7] nonlinear equations. The inverse scattering method allows one to obtain exact solutions to a class of nonlinear evolution equations, i. e., equations describing the evolution in time of a specified quantity from its initial state.

The solutions to these classes of equations are constructed through a succession of linear computations. Basically this operation can be broken down into three steps.

(1) Direct scattering: By way of a specified linear scattering (eigenvalue) problem one calculates the scattering parameters, such as the eigenvalues, reflection coefficients, and bound state normalization constants, from a knowledge of the initial conditions of the evolution equations.

(2) Time evolution of scattering parameters: By solving a system of linear differential (or difference) equations one obtains the time dependence of the scattering parameters.

(3) Inverse scattering: Knowing the time dependence of the scattering parameters one can reconstruct the solution by the formulas of inverse scattering theory (see Refs. 7, 8).

As one can easily see from the solution process, a key step in using this method is to obtain a relationship between the nonlinear evolution equation and a linear scattering problem. The point of view that will be used here is to start with a given scattering problem for which the scattering theory has been developed, and determine the class of evolution equations associated with it.

The scattering problem that will be used as the starting point is a discrete version of the Zakharov—Shabat eigenvalue problem considered by Ablowitz and Ladik,[8]

$$v_1(n+1) = z v_1(n) + Q(n) v_2(n),$$

$$v_2(n+1) = (1/z) v_2(n) + R(n) v_1(n). \tag{1.1}$$

Here z is the eigenvalue, and the potentials $Q(n)$ and $R(n)$ are defined on the interval $|n| < \infty$. It is understood that the potentials are also functions of time, and this temporal dependence can be considered to be continuous or discrete. The class of evolution equations which is associated with (1.1) includes, among others, the discrete versions of the nonlinear Schrödinger and modified KdV equations. The relationship between the evolution equations and (1.1) was first obtained by an expansion process developed in Ref. 8.

The main result of this paper is that every equation in this class can be characterized by the dispersion relation of its associated linearized version and a discrete sum—difference operator. A similar result has been found in the continuous theory first by Albowitz, Kaup, Newell, and Segur[9] by making use of the time dependent eigenfunctions and a squared eigenvalue problem. Later, Calogero and Degasperis[10] obtained a more general class of solvable equations by a generalized Wronskian technique. The approach of this paper is similar, in spirit, to that of Calogero and Degasperis. The class of soluble partial differential—difference equations generated by a discrete version of the generalized Wronkian equations is given by

$$\frac{\partial}{\partial t}\begin{pmatrix} R(n) \\ Q(n) \end{pmatrix} + \omega(L)\begin{pmatrix} R(n) \\ -Q(n) \end{pmatrix} = 0. \tag{1.2}$$

In the above representation for the evolution equations ω is the dispersion relation of the linearized version of (1.2) and L is a sum—difference operator given by

$$L = \begin{bmatrix} E^-(\,) - R(n-1)a(n)\mathcal{G}_n(Q(j)) - R(n)\mathcal{G}_n^-(Q(j)) & R(n-1)a(n)\mathcal{G}_n(R(j)) - R(n)\mathcal{G}_n^+(R(j)) \\ -Q(n+1)a(n)\mathcal{G}_{n+1}(Q(j)) - Q(n)\mathcal{G}_{n+1}^-(Q(j)) & E^+(\,) + Q(n+1)a(n)\mathcal{G}_{n+1}(R(j)) + Q(n)\mathcal{G}_{n+1}^+(R(j)) \end{bmatrix}, \tag{1.3}$$

 Copyright © 1977 American Institute of Physics

where $a(n)=1-R(n)Q(n)$ and $E^{\pm}(\ )$ is defined by $E^{\pm}(u(n))=u(n\pm1)$. The operators $\mathcal{J}_n$ and $\mathcal{J}_n^{\pm}$ are defined by

$$\mathcal{J}_n(x(j))y(n)\equiv\sum_{j=n}^{\infty}\frac{x(j)y(j)}{a(j)},$$
$$\mathcal{J}_n^{\pm}(x(j))y(n)\equiv\sum_{j=n}^{\infty}x(j)y(j\pm1). \tag{1.4}$$

The discrete version of the generalized Wronskian equations yields not only exactly soluble nonlinear differential—difference equations, but exactly soluble difference—difference equations as well. A direct consequence of generating each evolution equation is an ordinary linear differential (or difference) equation involving the time evolution of the scattering parameters of Eq. (1.1). The flexibility of this method also allows us to obtain Bäcklund transformations for this class of evolution equations. In fact, the partial difference equations can be viewed as a special case of Bäcklund transformation.

In this theory the fundamental quantity is the operator L which is found through a generalized Wronskian technique. However, before going into the details of computing L we wish to elucidate upon what we mean by the general Wronskian technique.

In the discrete case the usual Wronskian relation is constructed from two different solutions of the same difference problem. In this instance we will consider the scattering problem (1.1) in conjunction with two sets of potentials, first with the set $[Q(n),R(n)]$ and then with the set $[Q'(n),R'(n)]$. Now utilizing two solutions of (1.1), one corresponding to the potentials $[Q(n),R(n)]$ and the other to the potentials $[Q'(n),R'(n)]$, we construct a generalized Wronskian relation.

Once we have the generalized Wronskian relation, Eq. (2.5), we employ two iteration procedures, one in powers of z^2 and the other in powers of $1/z^2$, to determine an operator corresponding to all four of the potentials. Upon relating the primed and nonprimed potentials we can immediately deduce the operator L.

In Sec. II we develop in detail the discrete generalized Wronskian equations which lead to the operator given by (1.3). Section III is devoted to obtaining evolution equations, both partial differential—difference equations and partial difference equations. Also in this section we see how the generalized Wronskian equations lead to Bäcklund transformations. All the results of this section are the consequence of simply forming linear combinations of the generalized Wronskian equations of Sec. II. Section IV is reserved for concluding remarks.

II. GENERALIZED WRONSKIAN TECHNIQUE

We start by rewriting the discrete Zakharov—Shabat eigenvalue problem (1.1) in the form

$$\Phi(n+1)=[Z+Q(n)I_1+R(n)I_2]\Phi(n), \tag{2.1}$$

where $\Phi(n)$ is the 2×2 matrix solution of the wavefunctions and the matrices Z, I_1, and I_2 are defined by

$$Z=\begin{bmatrix}z&0\\0&1/z\end{bmatrix},\quad I_1=\begin{bmatrix}0&1\\0&0\end{bmatrix},\quad I_2=\begin{bmatrix}0&0\\1&0\end{bmatrix}. \tag{2.2}$$

In the analysis it is assumed that the potentials $Q(n)$ and $R(n)$ decay exponentially fast as $|n|\to\infty$. Under this assumption, the asymptotic boundary conditions of (2.1) are given by

$$n\to-\infty:\ \Phi(n)\sim\begin{bmatrix}\alpha_1z^n&0\\0&\alpha_2z^{-n}\end{bmatrix},$$
$$n\to+\infty:\ \Phi(n)\sim\begin{bmatrix}z^n&\beta_2z^n\\\beta_1z^{-n}&z^{-n}\end{bmatrix}. \tag{2.3}$$

If we think of (2.1) as an iteration scheme for determining $\Phi(n+1)$ in terms of $\Phi(n)$, we can also invert the interation and determine $\Phi(n)$ in terms of $\Phi(n+1)$; thus

$$\Phi(n)=\frac{1}{1-R(n)Q(n)}[Z-Q(n)I_1-R(n)I_2]\Phi(n+1). \tag{2.4}$$

Let us now consider Eqs. (2.1) and (2.3) for a second set of potentials and scattering parameters denoted by $Q'(n)$, $R'(n)$ and $\alpha_1',\alpha_2',\beta_1',\beta_2'$, respectively. Corresponding to the primed eigenvalue problem and boundary conditions we have the solution $\Phi'(n)$. The following generalized Wronskian relation is a direct consequence of Eq. (2.1),

$$\sum_{n=-\infty}^{\infty}P(n+1)\Phi'^{t}(n)[(Z+Q'(n)I_2+R'(n)I_1)F(n+1)$$
$$-F(n)(Z^{-1}-Q(n)I_1-R(n)I_2)]\Phi(n+1)$$
$$=\Phi'^{t}(n)P(n)F(n)\Phi(n)\Big|_{-\infty}^{\infty}, \tag{2.5}$$

where the superscript t denotes the usual matrix transpose. In Eq. (2.5) $P(n)$ is the scalar $P(n)=\prod_{j=n}^{\infty}\times(1-R(j)Q(j))$ and $F(n)$ is an arbitrary 2×2 matrix with elements given by

$$F(n)=\begin{bmatrix}A(n)&B(n)\\C(n)&D(n)\end{bmatrix}. \tag{2.6}$$

In deriving the operator form of the evolution equations it is convenient to introduce the matrix $M(n)$ defined by

$$M(n)=[Z+Q'(n)I_2+R'(n)I_1]F(n+1)$$
$$-F(n)[Z^{-1}-Q(n)I_1-R(n)I_2]. \tag{2.7}$$

Using the definition of $F(n)$, the elements of $M(n)$ are explicity given by

$$M_{11}(n)=zA(n+1)-(1/z)A(n)+R'(n)C(n+1)+R(n)B(n),$$
$$M_{12}(n)=z\,\Delta_nB(n)+R'(n)D(n+1)+Q(n)A(n),$$
$$M_{21}(n)=(1/z)\Delta_nC(n)+R(n)D(n)+Q'(n)A(n+1),$$
$$M_{22}(n)=(1/z)D(n+1)-zD(n)+Q'(n)B(n+1)+Q(n)C(n), \tag{2.8}$$

where Δ_n is the forward difference operator defined by $\Delta_nu(n)=u(n+1)-u(n)$.

The technique used from this point on is basically a double iteration procedure. Since $F(n)$ is arbitrary, we will first start from a specific F and iterate in terms of z^2; then, second, we will start from a different F and iterate in terms of $1/z^2$. In both instances we will require $M_{12}(n)=M_{21}(n)=0$ and then determine $M_{11}(n)$ and $M_{22}(n)$. The notation that we will use is as follows. For the $(l+1)$th iteration in terms of z^2 let $M_{11}(n)=z^2H_1^{(l)}(n)+H_1^{(l+1)}(n)$, $M_{22}(n)=z^2H_4^{(l)}(n)+H_4^{(l+1)}(n)$, and for the

$(l+1)$th iteration in terms of $1/z^2$ let $M_{11}(n)=1/z^2h_1^{(l)}(n)+h_1^{(l+1)}(n)$, $M_{22}(n)=1/z^2h_4^{(l)}(n)+h_4^{(l+1)}(n)$. In both instances we will assume we know the terms corresponding to l, and we must determine the terms corresponding to $l+1$.

As our starting point for the z^2 iteration, we choose the initial elements of F to be $A^{(0)}(n)=C^{(0)}(n)=D^{(0)}(n)=0$ and $B^{(0)}(n)=1$, and for beginning the $1/z^2$ iteration the initial elements of F are chosen to be $A^{(0)}(n)=B^{(0)}(n)=D^{(0)}(n)=0$ and $C^{(0)}(n)=1$. This selection yields

$$H^0(n)=\begin{pmatrix}H_1^{(0)}(n)\\H_4^{(0)}(n)\end{pmatrix}=\begin{pmatrix}R(n)\\Q'(n)\end{pmatrix},$$
$$h^{(0)}(n)=\begin{pmatrix}h_1^{(0)}(n)\\h_4^{(0)}(n)\end{pmatrix}=\begin{pmatrix}R'(n)\\Q(n)\end{pmatrix}. \tag{2.9}$$

Substituting these values into Eq. (2.5) gives the two general Wronskian relations,

$$\sum_{n=-\infty}^{\infty}P(n+1)\Phi'^{t}(n)IH^{(0)}(n)\Phi(n+1)=\begin{bmatrix}\beta_1 & 1-P(-\infty)\alpha_1\alpha_2\\ \beta_2'\beta_1 & \beta_2'\end{bmatrix}, \tag{2.10}$$

and

$$\sum_{n=-\infty}^{\infty}P(n+1)\Phi'^{t}(n)Ih^{(0)}(n)\Phi(n+1)=\begin{bmatrix}\beta_1' & \beta_1'\beta_2\\ 1-P(-\infty)\alpha_1\alpha_1' & \beta_2\end{bmatrix}, \tag{2.11}$$

where I is the usual 2×2 identity matrix.

We next consider the z^2 iterations for $l>0$. Letting $A^{(l)}(n)=zG_1^{(l)}(n)$, $D^{(l)}(n)=zG_4^{(l)}(n)$, and setting $M_{12}=M_{21}=0$ we can solve the resulting equations for $B^{(l)}(n)$ and $C^{(l)}(n)$,

$$B^{(l)}(n)=\sum_{j=n}^{\infty}[R'(j)G_4^{(l)}(j+1)+Q(j)G_1^{(l)}(j)],$$
$$C^{(l)}(n)=z^2\sum_{j=n}^{\infty}[Q'(j)G_1^{(l)}(j+1)+R(j)G_4^{(l)}(j)]. \tag{2.12}$$

Substituting these two results into the equations corresponding to $M_{11}(n)$ and $M_{22}(n)$ yields

$$\begin{aligned}&z^2H_1^{(l)}(n)-H_1^{(l+1)}(n)\\&\quad=z^2G_1^{(l)}(n+1)-G_1^{(l)}(n)+z^2R'(n)\\&\qquad\times\sum_{j=n+1}^{\infty}[Q'(j)G_1^{(l)}(j+1)+R(j)G_4^{(l)}(j)]+R(n)\\&\qquad\times\sum_{j=n}^{\infty}[R'(j)G_4^{(l)}(j+1)+Q(j)G_1^{(l)}(j)],\\&z^2H_4^{(l)}(n)-H_4^{(l+1)}(n)\\&\quad=G_4^{(l)}(n+1)-z^2G_4^{(l)}(n)+Q'(n)\\&\qquad\times\sum_{j=n+1}^{\infty}[R'(j)G_4^{(l)}(j+1)+Q(j)G_1^{(l)}(j)]+z^2Q(n)\\&\qquad\times\sum_{j=n}^{\infty}[Q(j)G_1^{(l)}(j+1)+R(j)G_4^{(l)}(j)].\end{aligned} \tag{2.13}$$

If corresponding powers of z^2 and z^0 are set equal, the following four equations result:

$$H_1^{(l)}(n)=G_1^{(l)}(n+1)-R'(n)\times\sum_{j=n+1}^{\infty}[Q'(j)G_1^{(l)}(j+1)+R(j)G_4^{(l)}(j)], \tag{2.14}$$

$$H_1^{(l+1)}(n)=G_1^{(l)}(n)-R(n)\times\sum_{j=n}^{\infty}[R'(j)G_4^{(l)}(j+1)+Q(j)G_3^{(l)}(j)], \tag{2.15}$$

$$H_4^{(l)}(n)=-G_4^{(l)}(n)+Q(n)\times\sum_{j=n}^{\infty}[Q'(j)G_1^{(l)}(j+1)+R(j)G_4^{(l)}(j)], \tag{2.16}$$

$$H_4^{(l+1)}(n)=-G_4^{(l)}(n+1)-Q'(n)\times\sum_{j=n+1}^{\infty}[R'(j)G_4^{(l)}(j+1)+Q(j)G_3^{(l)}(j)]. \tag{2.17}$$

Since it is assumed that we know $H^{(l)}$, we can solve Eqs. (2.14) and (2.16) for $G_1^{(l)}(n)$ and $G_4^{(l)}(n)$ in terms of the known quantities and then substitute back into Eqs. (2.15) and (2.17) and obtain $H^{(l+1)}$ in terms of $H^{(l)}$.

To compute $H^{(l+1)}$ in terms of $H^{(l)}$, we begin by shifting the index of Eq. (2.14) down by one and comparing the result to (2.16). This calculation yields

$$Q(n)[H_1^{(l)}(n-1)-G_1^{(l)}(n)]=R'(n-1)[H_4^{(l)}(n)+G_4^{(l)}(n)]. \tag{2.18}$$

If we subtract Eq. (2.14) from (2.16) and use the relation given by (2.18), we arrive at the following difference equation:

$$a'(n)\frac{H_1^{(l)}(n)-G_1^{(l)}(n+1)}{R'(n)}-a(n)\frac{H_1^{(l)}(n-1)-G_1^{(l)}(n)}{R'(n-1)}=R(n)H_4^{(l)}(n)-Q'(n)H_1^{(l)}(n), \tag{2.19}$$

where $a(n)=1-R(n)Q(n)$ and $a'(n)=1-R'(n)Q'(n)$. Solving the difference Eq. (2.19) and using (2.18) yields

$$G_1^{(l)}(n)=H_1^{(l)}(n-1)-R'(n-1)p(n)+R'(n-1)\times\sum_{j=n}^{\infty}\left[\left(\prod_{k=n}^{j}\frac{a'(k)}{a(k)}\right)\frac{R(j)H_4^{(l)}(j)-Q'(j)H_2^{(l)}(j)}{a'(j)}\right], \tag{2.20}$$

$$G_4^{(l)}(n)=-H_4^{(l)}(n)+Q(n)p(n)-Q(n)\times\sum_{j=n}^{\infty}\left[\left(\prod_{k=n}^{j}\frac{a'(k)}{a(k)}\right)\frac{R(j)H_4^{(l)}(j)-Q'(j)H_1^{(l)}(j)}{a'(j)}\right], \tag{2.21}$$

where $p(n)=p\prod_{j=n}^{\infty}a'(j)/a(j)$ (p an arbitrary constant). If the values of $G_1^{(l)}(n)$ and $G_4^{(l)}(n)$ are substituted in Eqs. (2.15) and (2.17), we find

$$\begin{pmatrix}H_1^{(l+1)}(n)\\H_4^{(l+1)}(n)\end{pmatrix}=\Lambda\begin{pmatrix}H_1^{(l)}(n)\\H_4^{(l)}(n)\end{pmatrix}, \tag{2.22}$$

where the 2×2 matrix operator Λ is given by

$$\Lambda=\begin{bmatrix}E^-(\ )+R'(n-1)a(n)\mathcal{G}_n(-Q'(j))-R(n)\mathcal{G}_n^-(Q(j)) & R'(n-1)a(n)\mathcal{G}_n(R(j))+R(n)\mathcal{G}_n^+(R(j))\\ Q(n+1)a'(n)\mathcal{G}_{n+1}(-Q'(j))-Q'(n)\mathcal{G}_{n+1}^-(Q(j)) & E^+(\ )+Q(n+1)a'(n)\mathcal{G}_{n+1}(R(j))+Q'(n)\mathcal{G}_{n+1}^+(R(j))\end{bmatrix}. \tag{2.23}$$

Again $E^{\pm}$ is the shift operator defined by $E^{\pm}(U(n)) \equiv U(n \pm 1)$ and the operator $\mathcal{J}_n$ and $\mathcal{J}_n^{\pm}$ are defined by

$$\mathcal{J}_n(x(j))y(n) \equiv \left[\frac{p(n)}{2} y(n) + \sum_{j=n}^{\infty} \left(\prod_{i=n}^{j} \frac{a'(i)}{a(i)} \frac{x(j)y(j)}{a'(j)}\right)\right], \quad \mathcal{J}_n^{\pm}(x(j)y(j)) \equiv \sum_{j=n}^{\infty} x(j)y(j \pm 1). \tag{2.24}$$

After l iterations we can write the generalized Wronskian equation (2.5) as

$$\sum_{n=-\infty}^{\infty} P(n+1)\Phi'^{t}(n)[IH^{(l+1)}(n)]\Phi(n+1) - z^2 \sum_{n=-\infty}^{\infty} P(n+1)\Phi'^{t}(n)[IH^{(l)}(n)]\Phi(n+1) = S^{(l+1)}, \tag{2.25}$$

where

$$S^{(l+1)} = \Phi'^{t}(n)P(n)F^{(l+1)}(n)\Phi(n)\Big|_{-\infty}^{\infty}, \tag{2.26}$$

and the values of $G_1^{(l+1)}(n)$ and $G_4^{(l+1)}(n)$ are used to calculate the elements of $F^{(l+1)}(n)$. Since (2.26) is a difference equation with respect to the index l, we can combine its solution, with Eqs. (2.22) and (2.10), to obtain

$$\sum_{n=-\infty}^{\infty} P(n+1)\Phi'^{t}(n)[I\Lambda^{l}H^{(0)}(n)]\Phi(n+1) = (z^2)^{l}\begin{bmatrix} \beta_1 & 1 - P(-\infty)\alpha_1'\alpha_2 \\ \beta_2'\beta_1 & \beta_2' \end{bmatrix} + \sum_{j=1}^{l} (z^2)^{l-j} S^{(j)} \quad (l \geq 0). \tag{2.27}$$

Equation (2.27) is the generalized Wronskian equation with respect to the z^2 iteration.

The next step is to proceed in an analogous manner for the $1/z^2$ iterations. Having already found $h^{(0)}(n)$, given by Eq. (2.9), we let $A^{(l)}(n) = (1/z)g_1^{(l)}(n)$, $D(n) = (1/z)g_4^{(l)}(n)$, and set $M_{12} = M_{21} = 0$. This restriction allows us to determine $B^{(l)}(n)$ and $C^{(l)}(n)$,

$$B^{(l)}(n) = (1/z^2) \sum_{j=n}^{\infty} [R'(j)g_4^{(l)}(j+1) + Q(j)g_1^{(l)}(j)], \quad C^{(l)}(n) = \sum_{j=n}^{\infty} [Q'(j)g_1^{(l)}(j+1) + R(j)g_4^{(l)}(j)]. \tag{2.28}$$

The results of (2.28) are inserted into the equations for $M_{11}(n)$ and $M_{12}(n)$, and after equating corresponding powers of z^0 and $1/z^2$ we obtain the following four equations:

$$h_1^{(l)}(n) = g_1^{(l)}(n) - R(n) \sum_{j=n}^{\infty} [R'(j)g_4^{(l)}(j+1) + Q(j)g_1^{(l)}(j)], \tag{2.29}$$

$$h_1^{(l+1)}(n) = g_1^{(l)}(n+1) + R'(n) \sum_{j=n+1}^{\infty} [Q'(j)g_1^{(l)}(j+1) + R(j)g_4^{(l)}(j)], \tag{2.30}$$

$$h_4^{(l)}(n) = -g_4^{(l)}(n+1) - Q'(n) \sum_{j=n+1}^{\infty} [R'(j)g_4^{(l)}(j+1) + Q(j)g_1^{(l)}(j)], \tag{2.31}$$

$$h_4^{(l+1)}(n) = -g_4^{(l)}(n) + Q(n) \sum_{j=n}^{\infty} [Q'(j)g_1^{(l)}(j+1) + R(j)g_4^{(l)}(j)]. \tag{2.32}$$

These four equations can be inverted to determine $h^{(l+1)}(n)$ in terms of $h^{(l)}(n)$ in a similar manner as before. Using Eqs. (2.29) and (2.31), we first obtain $g_1^{(l)}$ and $g_4^{(l)}(n)$ which are given by

$$g_1^{(l)}(n) = h_1^{(l)}(n) - R(n)p(n) + R(n) \sum_{j=n}^{\infty} \left[\left(\prod_{i=n}^{\infty} \frac{a'(i)}{a(i)}\right)\left(\frac{Q(j)h_1^{(l)}(j) - R'(j)h_4^{(l)}(j)}{a'(j)}\right)\right], \tag{2.33}$$

$$g_4^{(l)}(n) = -h_4^{(l)}(n-1) + Q'(n-1)p(n) - Q'(n-1) \sum_{j=n}^{\infty} \left[\left(\prod_{i=n}^{i} \frac{a'(i)}{a(i)}\right)\left(\frac{Q(j)h_1^{(l)}(j) - R'(j)h_4^{(l)}(j)}{a'(j)}\right)\right]. \tag{2.34}$$

Substituting the above two equations into (2.30) and (2.32) yields

$$\begin{pmatrix} h_1^{(l+1)}(n) \\ h_4^{(l+1)}(n) \end{pmatrix} = \tilde{\Lambda} \begin{pmatrix} h_1^{(l)}(n) \\ h_4^{(l)}(n) \end{pmatrix}, \tag{2.35}$$

where the operator $\tilde{\Lambda}$ is given by

$$\tilde{\Lambda} = \begin{bmatrix} E^{+}(\,) + R(n+1)a'(n)\mathcal{J}_{n+1}(Q(j)) + R'(n)\mathcal{J}_{n+1}^{+}(Q'(j)) & R(n+1)a'(n)\mathcal{J}_{n+1}(-R'(j)) - R'(n)\mathcal{J}_{n+1}^{-}(R(j)) \\ Q'(n-1)a(n)\mathcal{J}_n(Q(j)) + Q(n)\mathcal{J}_n^{+}(Q'(j)) & E^{-}(\,) - Q'(n-1)a(n)\mathcal{J}_n(-R'(j)) - Q(n)\mathcal{J}_n^{-}(R(j)) \end{bmatrix}. \tag{2.36}$$

Again $E^{\pm}(\,)$ is the shift operator and the operators $\mathcal{J}_n$ and $\mathcal{J}_n^{\pm}$ are defined in (2.24). Corresponding to the $1/z^2$ iteration, the generalized Wronkian equation (2.5) now takes the form

$$\sum_{n=-\infty}^{\infty} P(n+1)\Phi'^{t}(n)[Ih^{(l+1)}(n)]\Phi(n+1) - z^2 \sum_{n=-\infty}^{\infty} P(n+1)\Phi'^{t}(n)[Ih^{(l)}(n)]\Phi(n+1) = s^{(l+1)}, \tag{2.37}$$

where

$$s^{(l+1)} = \Phi'^{t}(n)P(n)F^{(l+1)}(n)\Phi(n)\Big|_{-\infty}^{\infty}, \tag{2.38}$$

and the values of $g_1^{(l+1)}(n)$ and $g_4^{(l+1)}(n)$ are used to compute the elements of $F^{(l+1)}(n)$. Solving, with respect to the

index l, the difference equation (2.37), and combining its solution with (2.35) and (2.11) we find,

$$\sum_{n=-\infty}^{\infty} P(n+1)\Phi'^{t}(n)[I\tilde{\Lambda}^{l}h^{(0)}(n)]\Phi(n+1) = (1/z^2)^{l}\begin{bmatrix} \beta_1' & \beta_1'\beta_2 \\ 1-P(-\infty)\alpha_1\alpha_2' & \beta_2 \end{bmatrix} + \sum_{j=1}^{\infty}(1/z^2)^{l-j}s^{(j)} \quad (l \geq 0). \tag{2.39}$$

Equation (2.39) is the generalized Wronskian equation corresponding to the $1/z^2$ iteration.

It is also possible to relate the two operators Λ and $\tilde{\Lambda}$. To do this, we first use Eqs. (2.14)–(2.17) and solve for $H^{(l)}(n)$ in terms of $H^{(l+1)}(n)$. This results in the equation

$$H^{(l)}(n) = \tilde{\Lambda}H^{(l+1)}(n). \tag{2.40}$$

Next, the set of Eqs. (2.29)–(2.32) are used to determine $h^{(l)}(n)$ in terms of $h^{(l+1)}(n)$. This calculation yields

$$h^{(l)}(n) = \Lambda h^{(l+1)}(n). \tag{2.41}$$

Using (2.22), (2.35), (2.40), and (2.41), it is easy to see that

$$\tilde{\Lambda} = \Lambda^{-1}. \tag{2.42}$$

Before obtaining evolution equations, it is convenient to gain further insight into the structure $S^{(l)}$ and $s^{(l)}$. The values of these functions are obtained by examining the limit as $|n| \to \infty$ of Eqs. (2.20), (2.21), (2.23), and (2.34) in conjunction with (2.22) and (2.35). Immediately it is seen that $G_1^{(l)}(n)$, $G_4^{(l)}(n)$, $g_1^{(l)}(n)$, $g_4^{(l)}(n)$ vanish as $n \to \infty$. Thus, for $l > 0$,

$$S^{(l)} = \begin{bmatrix} 0 & -P(-\infty)\alpha_1'\alpha_2 \sum_{j=-\infty}^{\infty}[R'(j)G_4^{(l)}(j+1) + Q(j)G_1^{(l)}(j)] \\ -P(-\infty)\alpha_1\alpha_2' z^2 \sum_{j=-\infty}^{\infty}[Q'(j)G_1^{(l)}(j+1) + R(j)G_4^{(l)}(j)] & 0 \end{bmatrix}, \tag{2.43}$$

$$s^{(l)} = \begin{bmatrix} 0 & -P(-\infty)\alpha_1'\alpha_2(1/z^2) \sum_{j=-\infty}^{\infty}[R'(j)g_4^{(l)}(j+1) + Q(j)g_1^{(l)}(j)] \\ -P(-\infty)\alpha_1\alpha_2' \sum_{j=-\infty}^{\infty}[Q'(j)g_1^{(l)}(j+1) + R(j)g_4^{(l)}(j)] & 0 \end{bmatrix}. \tag{2.44}$$

We can simplify $S^{(l)}$ and $s^{(l)}$ further by considering the elements $s_{21}^{(l)}$ and $s_{12}^{(l)}$. It can be shown that the terms $G_1^{(l)}(n+1)$, $G_4^{(l)}(n)$, $g_1^{(l)}(n)$, and $g_4^{(l)}(n+1)$ can be expressed in the form

$$\begin{aligned} G_1^{(l)}(n+1) &= -\sum_{k=0}^{l-1} R(n-k)f_{(k)}{}^{(l)}(n), \quad G_4^{(l)}(n) = \sum_{k=0}^{l-1} Q'(n+k)f_{(k)}{}^{(l)}(n+k), \\ g_1^{(l)}(n) &= \sum_{k=0}^{l-1} R'(n+k)\tilde{f}_{(k)}{}^{(l)}(n+k), \quad g_4^{(l)}(n+1) = -\sum_{k=0}^{l-1} Q(n-k)\tilde{f}_{(k)}{}^{(l)}(n), \end{aligned} \tag{2.45}$$

where

$$\begin{aligned} f_{(k)}{}^{(l)}(n) &= f_{(k-1)}{}^{(l-1)}(n) - Q(n-k)\sum_{j=1}^{l-k-1} R(n-k-j)\left(\prod_{i=1}^{j-1} a(n-k-i)\right) f_{(k+j-1)}{}^{(l-1)}(n), \\ \tilde{f}_{(k)}{}^{(l)}(n) &= \tilde{f}_{(k-1)}{}^{(l-1)}(n) - R'(n-k)\sum_{j=1}^{l-k-1} Q'(n-k-j)\left(\prod_{i=1}^{j-1} a'(n-k-i)\right)\tilde{f}_{(k+j-1)}{}^{(l-1)}(n), \end{aligned} \tag{2.46}$$

by substituting the appropriate equations, (2.45), into (2.14)–(2.17) and (2.29)–(2.32). It is easily seen that if the expressions given by (2.45) are substituted into $S_{21}^{(l)}$ and $s_{12}^{(l)}$, the summation terms telescope. The result is that each infinite sum reduces to a finite number of terms each of which vanishes as $|n| \to \infty$; hence $S_{21}^{(l)}$ and $s_{12}^{(l)}$ both vanish, and thus

$$S^{(l)} = \begin{bmatrix} 0 & -P(-\infty)\alpha_1'\alpha_2 \sum_{j=-\infty}^{\infty}[R'(j)G_4^{(l)}(j+1) + Q(j)G_1^{(l)}(j)] \\ 0 & 0 \end{bmatrix}, \tag{2.47}$$

$$s^{(l)} = \begin{bmatrix} 0 & 0 \\ -P(-\infty)\alpha_1\alpha_2' \sum_{j=-\infty}^{\infty}[Q'(j)g_1^{(l)}(j+1) + R(j)g_4^{(l)}(j)] & 0 \end{bmatrix}. \tag{2.48}$$

We now have all the necessary equations to obtain the final form for the generalized Wronskian relation. Since Eqs. (2.27) and (2.39) are valid for all values of $l \geq 0$, we can take linear combinations of the two equations. This results in

$$\begin{aligned} &\sum_{n=-\infty}^{\infty} P(n+1)\Phi'^{t}(n)I(\omega_1(\Lambda)H^{(0)}(n) + \omega_2(\Lambda^{-1})h^{(0)}(n))\Phi(n+1) \\ &= \omega_1(z^2)\begin{bmatrix} \beta_1 & 1-P(-\infty)\alpha_1'\alpha_2 \\ \beta_2'\beta_1 & \beta_2' \end{bmatrix} + \omega_2(1/z^2)\begin{bmatrix} \beta_1' & \beta_1'\beta_2 \\ 1-P(-\infty)\alpha_1\alpha_2' & \beta_2 \end{bmatrix} + \sum_{l=1}^{N_1}(b_l y^{(l)}) + \sum_{l=1}^{N_2}(\tilde{b}_l\tilde{y}_l^{(l)}). \end{aligned} \tag{2.49}$$

In Eq. (2.49) $\omega_1(x)$ and $\omega_2(x)$ are polynomials of degree N_1 and N_2,

$$\omega_1(x) = \sum_{i=0}^{N_1} b_i x^i, \quad \omega_2(x) = \sum_{i=0}^{N_2}\tilde{b}_i\tilde{x}^i, \tag{2.50}$$

and

$$y^{(l)}=\sum_{j=1}^{l}(z^2)^{l-j}S^{(j)},\quad \tilde{y}^{(l)}=\sum_{j=1}^{l}(1/z^2)^{l-j}S^{(j)}. \tag{2.51}$$

Equation (2.49) constitutes the major result of the generalized Wronskian technique.

III. EVOLUTION EQUATIONS

In this section the primary interest will be focused on generating solvable nonlinear evolution equations using the operator equation (2.49). We will also see how this operator equation leads to generalized Bäcklund transformations. Before proceeding, it should be emphasized that all the terms of (2.49) which involve the potentials depend also on time. Up to this point, the analysis of Sec. II has been independent of any temporal considerations. In this section we will include the time dependence in our analysis, and it is because of this that we will be able to consider two classes of evolution equations; the first being differential–difference equations and the second being difference–difference equations.

For the first class of equations we begin by setting

$$R'(n)=R(n)+\dot{R}(n)\,dt,\quad Q'(n)=Q(n)+\dot{Q}(n)\,dt, \tag{3.1}$$

where the dot denotes partial differentiation with respect to time. Substituting (3.1) and the corresponding representations of the primed scattering parameters into Eqs. (2.10) and (2.11) and then subtracting (2.11) from (2.10), we find

$$\sum_{n=-\infty}^{\infty}P(n+1)\Phi'^{t}(n)\begin{bmatrix}-\dot{R}(n)\,dt & 0\\ 0 & \dot{Q}(n)\,dt\end{bmatrix}\Phi(n+1)$$
$$=\begin{bmatrix}-\dot{\beta}_1\,dt & 1-P(-\infty)\alpha_1\alpha_2-\beta_1\beta_2-[P(-\infty)\dot{\alpha}_1\alpha_2+\dot{\beta}_1\beta_2]\,dt\\ -[1-P(-\infty)\alpha_1\alpha_2-\beta_1\beta_2]+[\dot{\beta}_2\beta_1+P(-\infty)\dot{\alpha}_2\alpha_1]\,dt & \dot{\beta}_2\,dt\end{bmatrix}. \tag{3.2}$$

We next obtain two equations by ordering terms in (3.2). The first relates the scattering parameters

$$1=P(-\infty)\alpha_1\alpha_2+\beta_1\beta_2, \tag{3.3}$$

and the second relates the time derivatives of the potentials and scattering parameters,

$$\sum_{n=-\infty}^{\infty}P(n+1)\Phi'^{t}(n)\begin{bmatrix}-\dot{R}(n) & 0\\ 0 & \dot{Q}(n)\end{bmatrix}\Phi(n+1)=\begin{bmatrix}-\dot{\beta}_1 & -[P(-\infty)\dot{\alpha}_1\alpha_2+\dot{\beta}_1\beta_2]\\ \dot{\beta}_2\beta_1+P(-\infty)\dot{\alpha}_2\alpha_1 & \dot{\beta}_2\end{bmatrix}. \tag{3.4}$$

If Eq. (3.1) is also substituted into (2.49) and only first order terms are kept, and we note that $\Lambda=L$, then we arrive at the equation

$$\sum_{n=-\infty}^{\infty}P(n+1)\Phi'^{t}(n)\omega(L)\begin{pmatrix}R(n)\\ Q(n)\end{pmatrix}\Phi(n+1)=\omega(z^2)\begin{bmatrix}\beta_1 & \beta_1\beta_2\\ \beta_1\beta_2 & \beta_2\end{bmatrix}, \tag{3.5}$$

where $\omega(x)=\omega_1(x)+\omega_2(x^{-1})$. The general evolution equation is a direct consequence of equating (3.4) and (3.5). Thus, the potentials evolve according to

$$\frac{\partial}{\partial t}\begin{pmatrix}R(n)\\ Q(n)\end{pmatrix}+\omega(L)\begin{pmatrix}R(n)\\ -Q(n)\end{pmatrix}=0, \tag{3.6}$$

and the time evolution of the scattering parameters are given by

$$\frac{\partial}{\partial t}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}+\omega(z^2)\begin{pmatrix}\beta_1\\ -\beta_2\end{pmatrix}=0, \tag{3.7}$$

$$\frac{\partial}{\partial t}\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}=0. \tag{3.8}$$

If the scattering problem has discrete eigenvalues $z_{k,1}$ and $z_{k,2}$, then we must also find the time evolution of the bound state normalization constants $\gamma_{k,1}$ and $\gamma_{k,2}$. For potentials which decay faster than an exponential, we can analytically continue the β_i's off the unit circle, and we find

$$\frac{\partial}{\partial t}\begin{pmatrix}\gamma_{k,1}\\ \gamma_{k,2}\end{pmatrix}+\begin{pmatrix}\omega(z_{k,1}^2)\gamma_{k,1}\\ -\omega(z_{k,2}^2)\gamma_{k,2}\end{pmatrix}=0. \tag{3.9}$$

For slower decaying potentials, Eq. (3.3) can be arrived at by other means.

By considering the linearized version of (3.6) it is immediately evident that $\omega(L)$ is directly related to the dispersion relations of the $R(n)$ and $Q(n)$ equations. As examples, if we let

$$\omega(z^2)=-i(z^2-2+1/z^2), \tag{3.10}$$

and $R(n)=\mp Q^*(n)$, then Eqs. (3.6) reduce to a differential–difference version of the nonlinear Schrödinger equation,

$$iQ(n)_t=[Q(n+1)-Q(n-1)][1\pm Q(n)Q^*(n)]-2Q(n). \tag{3.11}$$

Also included in this class are two versions of the modified KdV equation. The first version is obtained by letting $R(n) = \mp Q(n)$ and

$$\omega(z^2) = (z^2 - 1/z^2). \tag{3.12}$$

Using this dispersion relation in (3.6) yields

$$Q(n)_t = [1 \pm Q^2(n)][Q(n+1) - Q(n-1)], \tag{3.13}$$

which is related, in the continuum limit, to

$$f_T = f_{XXX} \pm 6f^2 f_X, \tag{3.14}$$

by letting $Q(n) \sim \Delta x f(X, T)$ and scaling x and t as $x = X - 2\Delta x t$ and $t = 3/(\Delta x)^3 T$, respectively. The second version of (3.14) is obtained from (3.6) by choosing

$$\omega(z^2) = z^4/2 + z^2 - 1/z^2 + 1/(2z^4). \tag{3.15}$$

Corresponding to (3.15) we find

$$\begin{aligned} Q(n)_t = {} & [1 \pm Q^2(n)]\{Q(n-1) - Q(n+1) + 1/2Q(n+2)[1 \pm Q^2(n+1)] - 1/2Q(n-2)[1 \pm Q^2(n-1)] \pm 1/2Q(n) \\ & \times [Q^2(n+1) - Q^2(n-1)]\}. \end{aligned} \tag{3.16}$$

In the case of (3.16) the continuum limit is more straightforward. Simply letting $Q(n) \sim (\Delta x)^2 f(x, T)$, $t = (1/\Delta x)^3 T$, we find that (3.16) reduces to (3.14). If we let $Q(n) = e^{\omega t} z^{2n}$, it can be directly verified that (3.10), (3.12), and (3.15) are the dispersion relations of the linearized versions of (3.11), (3.13), and (3.16) respectively.

For the second class of equations, we will demonstrate by an example how equations (2.27) and (2.39) can be used to generate soluble nonlinear difference-difference equations. For the previous class of equations considered, (2.27) and (2.39) were used only when the index l was positive, however they can also be used when l is negative. The only change we must make is to start the sum in the last term at -1, hence

$$\sum_{n=-\infty}^{\infty} P(n+1)\Phi'^t(n)[I\Lambda^l H^{(0)}(n)]\Phi(n+1) = (z^2)^l \begin{bmatrix} \beta_1 & 1 - P(-\infty)\alpha_1'\alpha_2 \\ \beta_2'\beta_1 & \beta_2' \end{bmatrix} + \sum_{j=-1}^{\infty} (z^2)^{l-j} S^{(j)} \quad (l \le 0), \tag{3.17}$$

$$\sum_{n=-\infty}^{\infty} P(n+1)\Phi'^t(n)[I\tilde{\Lambda}^l h^{(0)}(n)]\Phi(n+1) = (1/z^2)^l \begin{bmatrix} \beta_1' & \beta_1'\beta_2 \\ 1 - P(-\infty)\alpha_1\alpha_2' & \beta_2 \end{bmatrix} + \sum_{j=-1}^{l} (1/z^2)^{-j} S^{(j)} \quad (l \le 0). \tag{3.18}$$

To illustrate the method, we will find a general evolution equation corresponding to discretizing the linear portion of the nonlinear Schrödinger equation

$$iq_t = q_{xx} \pm q^2 q^*, \tag{3.19}$$

according to the following six-point scheme:

$$i\Delta_m Q^m(n) = \delta_1 Q^m(n+1) + \delta_2 Q^m(n) + \delta_3 Q^m(n-1) + \epsilon_1 Q^{m+1}(n+1) + \epsilon_2 Q^{m+1}(n) + \epsilon_3 Q^{m+1}(n-1). \tag{3.20}$$

In the above equation, the index n corresponds to the spatial variable, the index m corresponds to the time variable, and Δ_m is defined by $\Delta_m Q^m(n) \equiv Q^{m+1}(n) - Q^m(n)$.

Through the example one will clearly see the crucial role the dispersion relation $\Omega(z^2)$ plays in generating the evolution equations. We begin by seeking a solution of (3.20) of the form $Q^m(n) = \Omega^m z^{2n}$. Thus $\Omega(z^2)$ must satisfy

$$\Omega(z^2) = \frac{i + \delta_1 z^2 + \delta_2 + \delta_3/z^2}{i - \epsilon_1 z^2 - \epsilon_2 - \epsilon_3/z^2} = \frac{\Omega_1(z^2)}{\Omega_2(z^2)}. \tag{3.21}$$

Next we form linear combinations of equations (2.29), (2.39), (3.17), and (3.18) to obtain

$$\begin{aligned} & \sum_{n=-\infty}^{\infty} P(n+1)\Phi'^t(n)\{I[\Omega_2(\Lambda^{-1})h^{(0)}(n) - \Omega_1(\Lambda)H^0(n)]\}\Phi(n+1) \\ & = \Omega_2(z^2) \begin{bmatrix} \beta_1' & \beta_1'\beta_2 \\ 1 - P(-\infty)\alpha_1\alpha_2' & \beta_2 \end{bmatrix} - \Omega_1(z^2) \begin{bmatrix} \beta_1 & 1 - P(-\infty)\alpha_1'\alpha_2 \\ \beta_2'\beta_1 & \beta_2' \end{bmatrix} - \epsilon_1 S^{(-1)} - \epsilon_3 S^{(1)} - \delta_1 S^{(1)} - \delta_1 S^{(-1)}. \end{aligned} \tag{3.22}$$

In computing the various quantities of (3.22) we identify the primed terms with time level m and the unprimed terms with time level $m+1$. Starting all calculations with

$$H^{(0)}(n) = \begin{pmatrix} R^{m+1}(n) \\ Q^m(n) \end{pmatrix}, \quad h^{(0)}(n) = \begin{pmatrix} R^m(n) \\ Q^{m+1}(n) \end{pmatrix}, \tag{3.23}$$

we compute $H^{(1)}(n)$ and $h^{(1)}(n)$ by using (2.22), (2.23), and (2.35), (2.36) respectively. If we let $p = 0$ in (2.23) and (2.36), we find

$$H^{(1)}(n) = \begin{bmatrix} R^{m+1}(n-1) - R^{m+1}(n) \sum_{j=n}^{\infty} [Q^{m+1}(j)R^{m+1}(j-1) - Q^m(j+1)R^m(j)] \\ Q^m(n+1) - Q^m(n) \sum_{j=n+1}^{\infty} [Q^{m+1}(j)R^{m+1}(j-1) - Q^m(j+1)R^m(j)] \end{bmatrix}, \tag{3.24}$$

$$h^{(1)}(n)=\begin{bmatrix} R^m(n+1)-R^m(n)\sum_{j=n+1}^{\infty}[Q^{m+1}(j-1)R^{m+1}(j)-Q^m(j)R^m(j+1)] \\ Q^{m+1}(n-1)-Q^{m+1}(n)\sum_{j=n}^{\infty}[Q^{m+1}(j-1)R^{m+1}(j)-Q^m(j)R^m(j+1)] \end{bmatrix}.$$

Unlike the case for partial differential–difference equations, we must also compute $H^{(-1)}$ and $h^{(-1)}(n)$. This is accomplished via Eqs. (2.40) and (2.41). In order to obtain nontrivial results, we will let $p=1$ in the definitions of the operators Λ and Λ^{-1}. Carrying out the computations, we arrive at

$$H^{(-1)}(n)=\begin{bmatrix} R^{m+1}(n+1)a^m(n)\prod_{j=n+1}^{\infty}a^m(j)/a^{m+1}(j) \\ Q^m(n-1)a^{m+1}(n)\prod_{j=n}^{\infty}a^m(j)/a^{m+1}(j) \end{bmatrix},\quad h^{(-1)}(n)=\begin{bmatrix} R^m(n-1)a^{m+1}(n)\prod_{j=n}^{\infty}a^m(j)/a^{m+1}(j) \\ Q^{m+1}(n+1)a^m(n)\prod_{j=n+1}^{\infty}a^m(j)/a^{m+1}(j) \end{bmatrix}. \tag{3.25}$$

If we get both sides of Eq. (3.22) equal to the zero matrix, substitute in Eqs. (3.23)–(3.25), and let $R^m(n)=\mp Q^m(n)^*$, then the left side of (3.22) yields the evolution equation

$$\begin{aligned} &i\Delta_m Q^m(n)-\delta_3 Q^{m+1}(n+1)a^m(n)\prod_{j=n+1}^{\infty}\frac{a^m(j)}{a^{m+1}(j)}-\delta_2 Q^{m+1}(n)-\delta_1 Q^{m+1}(n-1)-\delta_1 Q^m(n+1)-\delta_2 Q^m(n)\\ &-\delta_3 Q^m(n-1)a^{m+1}(n)\prod_{j=n}^{\infty}\frac{a^m(j)}{a^{m+1}(j)}\mp\delta_1\Big(Q^{m+1}(n)\sum_{j=n}^{\infty}[Q^{m+1*}(j)Q^{m+1}(j-1)-Q^{m*}(j+1)Q^m(j)]\\ &+Q^m(n)\sum_{j=n+1}^{\infty}[Q^{m+1}(j)Q^{m+1*}(j-1)-Q^m(j+1)Q^{m*}(j)]\Big)=0, \end{aligned} \tag{3.26}$$

where $\epsilon_1=\delta_3$, $\epsilon_2=\delta_2$, and $\epsilon_3=\delta_1$ for consistancy. By setting the right-hand side of (3.22) equal to the zero matrix and using (3.26), we obtain difference equations for the scattering parameters,

$$\beta_1^{m+1}-[1/\Omega(z^2)]\beta_1^m=0,\quad \beta_2^{m+1}-\Omega(z^2)\beta_2^m=0,\quad \alpha_1^{m+1}-\alpha_1^m=0,\quad \alpha_2^{m+1}-\alpha_2^m=0. \tag{3.27}$$

As before, if the scattering problem has discrete eigenvalues $z_{k,1}$ and $z_{k,2}$, then we must also find the time evolution of the bound state normalization constants. In this case it is found that they obey the difference equations

$$\gamma_{k,1}^{m+1}-[1/\Omega(z_{k,1}^2)]\gamma_{k,1}^m=0,\quad \gamma_{k,2}^{m+1}-\Omega(z_{k,2}^2)\gamma_{k,2}^m=0. \tag{3.28}$$

Thus we can readily see that the scattering parameters can be expressed in terms of their initial values by

$$\beta_1^m=\beta_1^0[\Omega(z^2)]^{-m},\quad \gamma_{k,1}^m=\gamma_{k,1}^0[\Omega(z_{k,1}^2)]^{-m},\quad \alpha_1^m=\alpha_1^0,\quad \beta_2^m=\beta_2^0[\Omega(z^2)]^m,\quad \gamma_{k,2}^m=\gamma_{k,2}^0[\Omega(z_{k,2}^2)]^m,\quad \alpha_2^m=\alpha_2^0. \tag{3.29}$$

As a special case of (3.26), if we let

$$\delta_1=\sigma,\quad \delta_2=-\sigma,\quad \delta_3=0, \tag{3.30}$$

then (3.20) can be written, after a summation by parts, as

$$\begin{aligned} i\Delta_m Q^m(n)=\sigma\Big[&\Delta_n Q^m(n)-\Delta_n Q^{m+1}(n-1)\pm\frac{1}{2}\Big(Q^m(n)[Q^m(n+1)Q^{m*}(n)+Q^{m+1}(n+1)Q^{m+1*}(n)]\\ &+Q^{m+1}(n)[Q^m(n-1)Q^{m*}(n)+Q^{m+1}(n-1)Q^{m+1*}(n)]-Q^m(n)\sum_{k=n+1}^{\infty}\Delta_m W^m(k)-Q^{m+1}(n)\sum_{k=n}^{\infty}\Delta_m W^{m*}(k)\Big)\Big], \end{aligned} \tag{3.31}$$

where $W^m(k)=Q^m(k)Q^{m*}(k-1)+Q^m(k+1)Q^{m*}(k)$. It is easy to see from (3.31) that if $\sigma=\Delta t$, $Q^{m+1}(n)=Q(n)+\Delta tQ(n)$, then (3.31) reduces to (3.11), and, if $\sigma=\Delta t/(\Delta x)^2$, $Q^m(n)=\Delta x q(n\Delta x, m\Delta t)$, then (3.31) has as its continuum limit the nonlinear Schrödinger equation (3.19). It is interesting to note, from a numerical analysis point of view, that the choice of constants, (3.30), guarantees that $|\Omega(z^2)|=1$, hence the linear portion of (3.31) is neutrally stable. For a more detailed investigation of equations of the form (3.21) we call the reader's attention to Ref. 11.

It should be emphasized that each evolution equation is generated by simply forming linear combinations to arrive at equations of the form (2.49) and (3.22). The identical procedure also allows us to obtain Bäcklund transformations. By simply equating both sides of Eq. (2.49) to the zero matrix we obtain generalized Bäcklund transformations. The left side yields the general form for the Bäcklund transformation

$$\omega_1(\Lambda)\begin{pmatrix}R(n)\\ Q'(n)\end{pmatrix}+\omega_2(\Lambda^{-1})\begin{pmatrix}R'(n)\\ Q(n)\end{pmatrix}=0, \tag{3.32}$$

where $\omega_1(\Lambda)$ and $\omega_2(\Lambda^{-1})$ are arbitrary polynomials in Λ and Λ^{-1} of, say, degree N_1 and N_2, i.e., $\omega_1(\Lambda)=\sum_{i=0}^{N_1}\omega_{i,1}\Lambda^i$ and $\omega_2(\Lambda)=\sum_{i=0}^{N_2}\omega_{2,i}\Lambda^{-i}$. The right-hand side yields relations between the scattering parameters. The diagonal terms yield

$$\beta_1'=-[\omega_1(z^2)/\omega_2(z^{-2})]\beta_1,\quad \beta_2'=[\omega_2(z^{-2})/\omega_1(z^2)]\beta_2, \tag{3.33}$$

and the off-diagonal terms yield

$$\alpha_1' = \frac{\omega_1(z^2)}{\omega_1(z^2) + \sum_{l=1}^{N_1}\{\omega_{1,l}\sum_{k=1}^{l}(z^2)^{l-k}\sum_{j=-\infty}^{\infty}[R'(j)G_4^{(k)}(j+1) + Q(j)G_1^{(k)}(j)]\}}\alpha_1,$$
$$\alpha_2' = \frac{\omega_2(z^{-2})}{\omega_2(z^{-2}) + \sum_{l=1}^{N_2}\{\omega_{2,l}\sum_{k=1}^{l}(z^{-2})^{l-k}\sum_{j=-\infty}^{\infty}[Q'(j)g_1^{(k)}(j+1) + R(j)g_4^{(k)}(j)]\}}\alpha_2, \tag{3.34}$$

Although the derivation of (3.32) and (3.33) is based only on the spatial dependence of the potentials, the above equations are also valid if we include the fact that the potentials depend on time. In this situation the scattering parameters may also depend on time. (It should be noted that for the first few iterations we have found that the summation terms contained in Eq. (3.34) are independent of time. Since α_i $(i=1,2)$ are time independent by (3.8) it follows that α_i' will also be time independent.]

If the Bäcklund transformation (3.32) is used to generate a new solution $R'(n)$, $Q'(n)$ from a given solution $R(n)$, $Q(n)$, there will be a dependence on t due to the time dependence of $R(n)$ and $Q(n)$; however, the "constants" of summation that are obtained in solving for the primed potentials will, in general, also depend on time. This time dependence can be obtained by substituting the new solution into the original evolution equation. This is slightly different from the usual formulation of the Bäcklund transformation (see, for example, Refs. 12, 13), where it is not necessary to use the evolution equation to determine the time dependence.

From the theory of inverse scattering it is well known that the reconstruction of the potentials does not involve the time variable explicitly. The time dependence only enters in through the scattering parameters β_i, which satisfy (3.7), and the bound state constants $\gamma_{k,i}$, which satisfy (3.9). Thus any relationship between the primed and unprimed betas implies a relationship between the primed and unprimed potentials, the link being the dispersion relation.

It is clear from Eqs. (3.33) that if β_i $(i=1,2)$ satisfies the linear differential equation (3.7), then β_i' is also a solution since β_i and β_i' differ only by a constant of proportionality which is independent of time. From generating evolution equations we know that the linear equation (3.7) corresponds to the nonlinear evolution equation (3.6). We can therefore conclude that for two pairs of potentials related by (3.32), if $R(n)$ and $Q(n)$ satisfy (3.6), then $R'(n)$ and $Q'(n)$ satisfy the same equation. We can also turn the argument around and say that since the primed and unprimed potentials satisfy the same evolution equation, the β_i (in addition to the β_i') have the same dependence; hence the summation terms in (3.34) must be constant.

For the important special case of $R(n) = \mp Q^*(n)$, the polynomials $\omega_1(z^2)$ and $\omega_2(z^{-2})$ are related by

$$\omega_1(z^2) = -(\omega_2(z^{-2}))^*. \tag{3.35}$$

Due to the arbitrariness of the polynomial $\omega_1(z^2)$, the Bäcklund transformations obtained from (3.31) correspond to adding a specific number of solitons to the new solution. The exact number of solitons added is equal to the degree of the polynomial $\omega_1(z^2)$. This can easily be seen from Eqs. (3.34). Since the number of solitons depends on the number of poles of α_i (α_i'), increasing the degree of $\omega_1(z^2)$ also measures the number of poles of α_i' (α_i). In this regard, Eq. (3.31) can be viewed as a nonlinear superposition principle.

Increasing the degree of $\omega_1(z^2)$ also intensifies the difficulty of generating new solutions from given solutions. This is exemplified if we consider the following. Let $R'(n) = -Q'^*(n)$, $R(n) = Q(n) = 0$, and $\omega_1(z^2) = C_1 z^2 + C_0$. Equation (3.32) becomes

$$C_1 Q'(n+1) - C_1 Q'(n)\sum_{j=n+1}^{\infty} Q'(j+1)Q'^*(j) + C_0 Q'(n) = 0. \tag{3.36}$$

If we divide Eq. (3.36) by $C_1 Q'(n)$ and take the difference, we obtain

$$1 + |Q'(n+1)|^2 = Q'^2(n+1)/[Q'(n)Q'(n+2)]. \tag{3.37}$$

Since the left side of (3.37) is real, let

$$Q'(n) = \exp[i\phi(n)]\rho(n). \tag{3.38}$$

Substituting (3.38) into (3.37) yields

$$\phi(n+2) + \phi(n) - 2\phi(n+1) = 0, \tag{3.39}$$

and

$$1 + \rho^2(n+1) = \rho^2(n+1)/[\rho(n)\rho(n+2)]. \tag{3.40}$$

Equation (3.39) is readily solved to yield

$$\phi(n) = n\phi + \phi_1, \tag{3.41}$$

where ϕ and ϕ_1 are constants with respect to n but may depend on time. A solution to (3.40) is obtained by sub-

stituting in the ansatz,

$$1/\rho(n) = d[\exp(2n\vartheta + \vartheta_1) + \exp(2n\vartheta - \vartheta_1)]. \tag{3.42}$$

The ansatz satisfies (3.40) provided

$$1/d = \exp(2\vartheta) - \exp(-2\vartheta). \tag{3.43}$$

Thus we find

$$Q'(n) = \exp[i(n\phi + \phi_1)] \sinh 2\vartheta \operatorname{sech}(2n\vartheta + \vartheta_1). \tag{3.44}$$

The time dependence of ϕ, ϕ_1, and ϑ_1 can be found by substituting (3.44) into Eq. (3.11). Upon choosing the plus sign in (3.11) we find

$$\frac{\partial t}{\partial \phi} = 0, \quad \frac{\partial t}{\partial \phi_1} = -\cos 2\vartheta \cos 2\phi + 2, \quad \frac{\partial t}{\partial \vartheta_1} = \sin 2\vartheta \sin 2\phi. \tag{3.45}$$

Substituting the solution of (3.45) into (3.44), we find the single soliton solution of (3.11) to be

$$Q'(n) = \exp\{i[n\phi - (\cos 2\vartheta \cos 2\phi - 2)t + \phi_0]\} \sin 2\vartheta \operatorname{sech}[2n\vartheta + (\sin 2\vartheta \sin 2\phi)t + \vartheta_0], \tag{3.46}$$

where ϕ_0 and ϑ_0 are constants independent of n and t. This solution agrees with that found by the inverse scattering method.

It is interesting to note the similarity between constructing Bäcklund transformations and generating partial difference–difference equations. In general, the difference–difference equations are obtained by setting the left side of (3.22), with appropriate $\Omega_1(Z^2)$ and $\Omega_2(Z^2)$, equal to the zero matrix. This yields

$$\Omega_2(\Lambda^{-1}) \begin{pmatrix} R^m(n) \\ Q^{m+1}(n) \end{pmatrix} - \Omega_1(\Lambda) \begin{pmatrix} R^{m+1}(n) \\ Q^m(n) \end{pmatrix} = 0, \tag{3.47}$$

which is similar to the Bäcklund transformations given by (3.31). If we make the identification

$$R(n) = \mp Q^*(n), \quad R'(n) = \mp Q'^*(n), \quad \omega_1(z^2) = \Omega_1(z^2) = 1 - i(z^2 - 1), \quad \omega_2(z^{-2}) = -\Omega_2(z^{-2}) = -1 - i(1/z^2 - 1), \tag{3.48}$$

along with primed potentials corresponding to time level m and unprimed potentials corresponding to time level $m+1$, then the equation for the Bäcklund transformation, (3.31), reduces to the partial difference equation (3.30), for $\sigma = 1$. Also corresponding to $\omega(z^2) = \omega_1(z^2) + \omega_2(z^{-2})$ we have the evolution equation (3.11). Hence we may view (3.30), for $\sigma = 1$, as a special case of a Bäcklund transformation relating the solution of (3.11) at different time levels. We also have, in the limit as $\Delta t \to 0$, Eq. (3.30) with $\sigma = \Delta t$ reducing to (3.11). Hence, we may view differential–difference equations as being infinitesimal Bäcklund transformations.

IV. CONCLUSION

In this paper we have added an alternate means of generating evolution equation to the theory of nonlinear discrete equations soluble by the inverse scattering method. Through the development of a discrete Wronskian technique we have been able to generate partial difference equations via an operator formulation. This generalized Wronskian approach has several advantages over the expansion process of Refs. 8, 11.

One such advantage is that the relation between the operator and the dispersion relation of the linearized equation becomes clear. Just as in the continuous theory, the role of the dispersion relation is of paramount importance. A second advantage is found in the computational aspects of generating the evolution equations. The operator approach has the feature that more complex equations are constructed out of existing equations while in the expansion procedure a fresh start must be made to generate each new equation. The manner in which the generalized Wronskian equations are derived leads to a third advantage. The derivation of equations of the form (2.49) and (3.22) does not involve any time dependent equations. The time dependence is treated as a parameter and is explicity included after one obtains the generalized Wronskian Eqs. (2.49) and (3.22). This gives the operator method the added flexibility of generating a class of difference–difference equations, as well as a class of differential–difference equations. Furthermore, since the time variable is treated as a parameter, Eqs. (2.49) and (3.21) lead immediately to generalized Bäcklund transformations. Since the transformations are highly nonlinear they are difficult to use; however, their construction reveals the connection between the two classes of discrete evolution equations. One class is a special case of the Bäcklund transformation associated with the second class.

*Supported in part by the Air Force Office of Scientific Research Grant 72-2187.
†Supported in part by the National Science Foundation under Grant MPS74-13252.

[1]R. Hirota, J. Phys. Soc. Japan 35, 289 (1973).
[2]M. Toda, Prog. Theor. Phys. Suppl. 45, 1974 (1970).
[3]F. Tappert, Numerical solution of the Korteweg—de Vries equation and its generalizations by the split step Fourier method, in *Nonlinear Wave Motion*, edited by A. C. Newell (American Mathematical Society, Providence, R.I., 1974), pp. 215—6.
[4]C. Gardner, J. Greene, M. Kruskal, and R. Miura, Phys. Rev. Lett. 19, 1095 (1967).
[5]C. Gardner, J. Greene, M. Kruskal, and R. Miura, Comm. Pure Appl. Math. 27, 97 (1974).
[6]M. Ablowitz, D. Kaup, A. Newell, and H. Segur, Phys. Rev. Lett. 31, 125 (1973).
[7]M. Ablowitz and J. Ladik, J. Math. Phys. 16, 598 (1975).
[8]M. Ablowitz and J. Ladik, J. Math. Phys. 17, 1011 (1976).
[9]M. Ablowitz, D. Kaup, A. Newell, and H. Segur, Stud. Appl. Math. 53, 249 (1974).
[10]F. Calogero and A. Degasperis, Nuovo Cimento B 32, 201 (1976).
[11]M. Ablowitz and J. Ladik, to appear in Stud. Appl. Math.
[12]G. Lamb Jr., J. Math. Phys. 15, 2157 (1974).
[13]H. Chen and C. Liu, J. Math. Phys. 16, 1428 (1975).

Solutions of nonlinear network equations by the inverse scattering method

John F. Ladik* and Shiu-Chu Chiu†

The Rockefeller University, New York, New York 10021
(Received 20 September 1976)

Two coupled equations for a nonlinear network are solved by the inverse-scattering method. N-soliton solutions are obtained. The effect of the continuum spectrum is investigated.

I. INTRODUCTION

Recently, Hirota,[1] and Hirota and Satsuma[2] presented N-soliton solutions of a nonreciprocal nonlinear network equation, or the Volterra competition equation, and a Miura transformation relating this equation to the Toda lattice. The purpose of this paper is to show how this system may be completely solved by the inverse-scattering method.

The nonlinear system of interest is the following:

$$\frac{\partial}{\partial t}I(n) = [Z_0^{-1} + I(n)][V(n) - V(n+1)],$$
$$\frac{\partial}{\partial t}V(n) = [Z_0 + V(n)][I(n-1) - I(n)], \tag{1}$$

where Z_0 is a constant. Let us set

$$a(n) = \tfrac{1}{2}[Z_0 I(n) + 1],$$
$$b(n) = \tfrac{1}{2}[Z_0^{-1} V(n) + 1]. \tag{2}$$

Then Eqs. (1) become

$$\frac{\partial}{\partial t}a(n) = 2Z_0 a(n)[b(n) - b(n+1)],$$
$$\frac{\partial}{\partial t}b(n) = 2Z_0^{-1} b(n)[a(n-1) - a(n)]. \tag{3}$$

As noted by Hirota,[1] Eqs. (3) are related to the Toda equations

$$\frac{\partial}{\partial t}\alpha(n) = \alpha(n)[\beta(n-1) - \beta(n)],$$
$$\frac{\partial}{\partial t}\beta(n) = 2[\alpha(n)^2 - \alpha(n+1)^2], \tag{4}$$

by a Miura transformation[3]

$$\alpha(n)^2 = a(n)b(n),$$
$$\beta(n) = Z_0 b(n+1) + Z_0^{-1} a(n) - \Lambda_0, \tag{5}$$

where

$$\Lambda_0 = \tfrac{1}{2}(Z_0 + Z_0^{-1}). \tag{6}$$

The solutions of Eqs. (4) by the inverse scattering method have been thoroughly studied by Flaschka,[4] and Kac and van Moerbeke.[5] These solutions represent the isospectral flow of the discrete Schrödinger equation,

$$\alpha(n+1)\psi(n+1) + \beta(n)\psi(n) + \alpha(n)\psi(n-1) = \Lambda\psi(n), \tag{7}$$

where the eigenvalue Λ can be expressed as

$$\Lambda = \tfrac{1}{2}(Z + Z^{-1}), \tag{8}$$

the time evolution being determined by[6]

$$\frac{\partial}{\partial t}\Psi(n) = -2\alpha(n+1)\Psi(n+1) + [\delta - \beta(n)]\Psi(n), \tag{9}$$

where δ is a constant.

Let us make the ansatz

$$a(n) = -Z_0\alpha(n)y(n-1)/y(n). \tag{10}$$

Substituting into Eq. (5), we easily see that $y(n)$ satisfies

$$\alpha(n+1)y(n+1) + \beta(n)y(n) + \alpha(n)y(n-1) = -\Lambda_0 y(n). \tag{11}$$

Thus $y(n)$ is a solution of Eq. (7) with eigenvalue $-\Lambda_0$. In the following, we assume $0 < Z_0 \leq 1$. There is no loss of generality, since Eqs. (1) are invariant under the transformations $V'(n) = I(n)$, $I'(n-1) = V(n)$, $Z_0' = Z_0^{-1}$.

We are interested in solutions of (7) associated with boundary conditions

$$\lim_{|n|\to\infty} a(n) = \lim_{|n|\to\infty} b(n) = \tfrac{1}{2}. \tag{12}$$

Correspondingly, $\alpha(n)$ also has the limiting value $\frac{1}{2}$ at infinity. If a pair of functions, $a(n)$ and $b(n)$, is transformed to $\alpha(n)$ and $\beta(n)$ via (5), the spectrum of (7) will consist of the unit circle $|Z| = 1$ and N discrete eigenvalues, which, for $|Z| \leq 1$, are on the real invervals[7]

$$1 \geq Z_i > 0, \text{ or } -Z_0 > Z_i \geq -1, \quad i = 1, \ldots, N. \tag{13}$$

Conversely, whenever (7) has a spectrum as above, the transformation back to $a(n)$ and $b(n)$ is nonsingular. For simplicity, we assume no bound state at $Z = \pm 1$. Aside from the bound states, the Jost solutions $f_1(Z, n)$ and $f_2(Z, n)$, defined by

$$\lim_{n\to\infty} f_1(Z, n) \to Z^n, \tag{14}$$

and

$$\lim_{n\to-\infty} f_2(Z, n) \to Z^{-n}, \tag{15}$$

are two independent solutions. It is well known[8] that they have triangular representations

$$f_1(Z, n) = \sum_{m=n}^{\infty} A(n, m)Z^m, \tag{16}$$

$$f_2(Z, n) = \sum_{m=-\infty}^{n} B(n, m)Z^{-m}, \tag{17}$$

so that $Z^{-n}f_1(Z, n)$, and $Z^n f_2(Z, n)$ are analytic inside the unit circle $|Z| < 1$. Also, direct substitution of (16)

Copyright © 1977 American Institute of Physics

and (17) into (7) shows that

$$\alpha(n)=A(n,n)/[2A(n-1,n-1)],$$
$$\beta(n)=[A(n,n)A(n-1,n)-A(n,n+1)A(n-1,n-1)] \times[A(n-1,n-1)A(n,n)]^{-1}. \tag{18}$$

The boundary condition (12) implies that $y(n)$ in Eq. (10) must be $f_1(-Z_0,n)$. With Eqs. (18), $a(n)$ and $b(n)$ can be expressed as

$$a(n)=-Z_0F_1(-Z_0,n-1)/2F_1(-Z_0,n),$$
$$b(n)=-A^2(n,n)F_1(-Z_0,n) \times[2Z_0A^2(n-1,n-1)F_1(-Z_0,n-1)]^{-1}, \tag{19}$$

where

$$F_1(-Z_0,n)=f_1(-Z_0,n)/A(n,n). \tag{20}$$

Equations (19) are the main expressions we seek to obtain from an inverse-scattering theory. One may proceed as in Refs. 4, 5, and 8 to solve the inverse problem by the Marchenko equation for $A(n,m)$. However, we prefer the formalism of Zakharov and Shabat[9] since it yields $F_1(-Z_0,n)$ and $A(n,n)$ directly. In Sec. III, we shall derive expressions for N solitons and investigate the effect of the continuum spectrum.

II. INVERSE SCATTERING THEORY

In this section, we shall derive representations of $F_1(-Z_0,n)$ and $A(n,n)$ in terms of the scattering data $(\gamma_l, S_{21}(Z))$ defined in Eqs. (21) and (30).

From Eqs. (7), (14), and (15) one directly sees that $f_1(Z^{-1},n)$ and $F_2(Z^{-1},n)$ are solutions which can be analytically continued outside the unit circle, $|Z|>1$. On the unit circle, f_1 and f_2 are related by

$$S_{22}(Z)f_2(Z,n)=f_1(Z^{-1},n)+S_{21}(Z)f_1(Z,n),$$
$$S_{11}(Z)f_1(Z,n)=f_2(Z^{-1},n)+S_{12}(Z)f_2(Z,n), \tag{21}$$

where $S_{ij}(Z)$ are the elements of the unitary S matrix. It is well known that $S_{22}(Z)=S_{11}(Z)$ and that they can be analytically continued into the unit circle $|Z|<1$, except at a finite number of points on the real axis Z_l, $l=1,\ldots,N$ where they have simple poles. At Z_l, we define C_l by

$$C_lf_1(Z_l,n)=f_2(Z_l,n). \tag{22}$$

As shown in Ref. 4, the time dependence of S_{ij} and C_l may be obtained from Eq. (9), and they are as follows:

$$S_{11}=S_{22}=\text{const} \quad \text{(in time)}, \tag{23}$$
$$S_{21}(Z,t)=S_{21}(Z,0)\exp[(Z^{-1}-Z)t], \tag{24}$$
$$S_{12}(Z,t)=S_{12}(Z,0)\exp[(Z-Z^{-1})t], \tag{25}$$

and

$$C_l(t)=C_l(0)\exp[(Z_l^{-1}-Z_l)t]. \tag{26}$$

We note that

$$1/A(n,n)=B(n,n)S_{22}(Z\to 0). \tag{27}$$

By considering the contour intergral

$$I(\zeta,n)=\oint_{C_\epsilon}\frac{dZ}{2\pi i}[S_{22}(Z)f_2(Z,n)Z^n(Z-\zeta)^{-1}A(n,n)^{-1}]$$

where C_ϵ is a small circle such that it excludes all bound states, one obtains

$$F_1(\zeta^{-1},n)\zeta^n-\sum_l\gamma_lF_1(Z_l,n)Z_l^{n+1}(Z_l-\zeta)^{-1} -\frac{P}{2\pi i}\oint_{C_0}\frac{dZ}{(Z-\zeta)}S_{21}(Z)F_1(Z,n)Z^n+\tfrac{1}{2}S_{21}(\zeta)F_1(\zeta,n)\zeta^n =1, \quad \text{for } |\zeta|=1, \tag{28}$$

and

$$F_1(\zeta^{-1},n)\zeta^n-\sum_l\gamma_lF_1(Z_l,n)Z_l^{n+1}(Z_l-\zeta)^{-1} -\frac{1}{2\pi i}\oint_{C_0}\frac{dZ}{(Z-\zeta)}S_{21}(Z)F_1(Z,n)Z^n=1, \quad \text{for } |\zeta|>1. \tag{29}$$

In the above, C_0 is the unit circle, the letter P in front of the integral denotes principal value, and[10]

$$\gamma_l=-\frac{C_l}{Z_l}\frac{d}{dZ}[S_{22}^{-1}(Z)]_{Z=Z_l}. \tag{30}$$

For $\zeta=Z_j^{-1}$, $j=1,\ldots,N$, Eq. (29) yields

$$F_1(Z_j,n)Z_j^{-n}-\sum_l\gamma_lF_1(Z_l,n)Z_l^{n+1}(Z_l-Z_j^{-1})^{-1} -\frac{1}{2\pi i}\oint_{C_0}\frac{dZ}{(Z-Z_j^{-1})}S_{21}(Z)F_1(Z,n)Z^n=1. \tag{31}$$

Equation (29) or (28) implies that, at $\zeta=-Z_0^{-1}$,

$$F_1(-Z_0,n)(-Z_0)^{-n} =1+\frac{P}{2\pi i}\oint_{C_0}\frac{dZ}{(Z+Z_0^{-1})}S_{21}(Z)F_1(Z,n)Z^n -\Theta(1-|Z_0|)S_{21}(-Z_0)F_1(-Z_0,n)(-Z_0)^n +\sum_l\gamma_lF_1(Z_l,n)Z_l^{n+1}(Z_l+Z_0^{-1})^{-1}, \tag{32}$$

where

$$\Theta(x)=\begin{cases}\frac{1}{2}, & x=0,\\ 0, & x>0.\end{cases} \tag{33}$$

Similarly, by considering

$$\tilde{I}(\zeta,n)=\oint_{C_R}\frac{dZ}{2\pi i}[S_{22}(Z^{-1})f_2(Z^{-1},n)Z^{-n}(Z-\zeta)^{-1}A(n,n)^{-1}],$$

where C_R is a circle of radius $R>1$ and including all Z_j^{-1}, one obtains for $|\zeta|<1$

$$[A(n,n)]^{-2} =\sum_l\gamma_lF_1(Z_l,n)Z_l^{n-1}(Z_l^{-1}-\zeta)^{-1}+F_1(\zeta,n)\zeta^{-n} +\frac{1}{2\pi i}\oint_{C_0}\frac{dZ}{(Z-\zeta)}S_{21}(Z^{-1})F_1(Z^{-1},n)Z^{-n}. \tag{34}$$

Using Eq. (29), Eq. (34) may be reduced to

$$[A(n,n)]^{-2}=1+\sum_l\gamma_lF_1(Z_l,n)Z_l^n +\frac{1}{2\pi i}\oint dZ\,S_{21}(Z)F_1(Z,n)Z^{n-1}. \tag{35}$$

Equations (28), (31), (32), and (35) are the equations for inverse scattering. First, Eqs. (28), (31) are solved for $F_1(Z_j,n)$ and $F_1(Z,n)$ on the unit circle. [Although Eqs. (28) and (31) are singular integral equations, they are equivalent to the Gel'fand–Levitan equations, as can be easily shown by Fourier transforming the equations, and hence they are soluable. In practice, explicit solutions for arbitrary $S_{21}(Z)$ is a difficult problem. For small $S_{21}(Z)$, one may use perturbation theory as shown in Sec. III. For arbitrary $S_{21}(Z)$ without bound states, the analysis is expected to be quite complicated and has not been done. For the type of analysis which can be expected, we refer the readers to Ref. 11.] These are substituted into Eqs. (32) and (35) to obtain $F_1(-Z_0,n)$ and $A(n,n)$. Equations (19) then allow us to determine $a(n)$ and $b(n)$.

III. SOLUTIONS

We first consider pure N solitons. In this case $S_{21}(Z)=0$ and Eq. (31) reduce to a set of N linear algebraic equations. Let us define

$$\zeta_j(n)\equiv\sqrt{\gamma_j}Z_j^n, \tag{36}$$

and matrix $M(n)$,

$$M_{ij}(n)\equiv\delta_{ij}+\zeta_i(n)\zeta_j(n)(1-Z_iZ_j)^{-1}. \tag{37}$$

Then, by Cramer's rule,

$$F_1(Z_l,n)=\gamma_l^{-1/2}\det\|M_j(n+1)\|/\det\|M(n+1)\|, \tag{38}$$

where the matrix $M_j(n+1)$ is the matrix $M(n+1)$ with the jth column replaced by

$$\begin{pmatrix}\zeta_1(n)\\ \zeta_2(n)\\ \vdots\\ \zeta_N(n)\end{pmatrix}.$$

Equation (35) now reduces to

$$[A(n,n)]^2=\det\|M(n+1)\|/\det\|M(n)\|. \tag{39}$$

This formula was first obtained by Flaschka[4]; its generalization to include the continuum spectrum has been found by Case.[12] From Eqs. (19) and (32),

$$a(n)=\tfrac{1}{2}\left[1+\sum_l\gamma_lF_1(Z_l,n-1)Z_l^n(Z_l+Z_0^{-1})^{-1}\right]\times\left[1+\sum_l\gamma_lF_1(Z_l,n)Z_l^{n+1}(Z_l+Z_0^{-1})^{-1}\right]. \tag{40}$$

For a single soliton,

$$S_{22}(Z)=(Z_1Z-1)/(Z-Z_1). \tag{41}$$

In this case, let $Z_1=\exp(-\Theta_1)$, $k_1=C_1(0)Z_1^{-1}$, we have

$$2a(n)-1=4k_1\sinh^2\Theta_1\exp(2\omega_1)[1+k_1\exp(2\omega_1)]^{-1}\times[1+\kappa_1k_1\exp(2\omega_1)]^{-1}, \tag{42}$$

where

$$\kappa_1=(Z_0+Z_1^{-1})(Z_0+Z_1)^{-1} \tag{43}$$

and

$$\omega_1=\sinh\Theta_1t-n\Theta_1. \tag{44}$$

This agrees with the single soliton solution obtained by Hirota *et al.*[2] When $1>Z_1>0$, we have a soliton propagating in the forward direction and there is no restriction on the magnitude of Z_1. When $-1<Z_1<0$, we have a soliton propagating backwards. In order to have regular solutions, κ_1 must be positive, which amounts to the restrictions $1>Z_1>0$ and $-Z_0>Z_1>-1$. This is a general restriction, since even for N solitons, the asymptotic behaviors will be N separated single solitons. Collisions of the solitons have been studied in Ref. 2. We shall not go into details, but only mention that the solitons preserve their shapes after collisions, the only effect being a phase shift.

The advantage of the inverse scattering method is that it allows a complete solution to the initial value problem which, in general, also contains the continuum spectrum. We shall investigate the effect of the presence of a small amount of the continuum spectrum. In this case, we may use perturbation theory with the pure solitons as the zeroth approximation. In the case of pure continuum, a steepest decent calculation at a fixed spatial point n shows that the solution decays asymptotically as $t^{-1/2}$. The case of the interaction of a forward soliton with a small admixture of continuum will be studied in some detail.

In this case, the proper reference frame is that which moves with the soliton, i.e., with velocity

$$v_1\equiv(\sinh\Theta_1)/\Theta_1\equiv1/\tau_1. \tag{45}$$

However, due to the discreteness of the spatial coordinate, we are forced to do calculations for only discrete times, j/v_1, where j is an integer. The solution at intermediate times has to be interpolated. In the following, the dependence will be explicitly labeled by j, and the F's are those in the moving frame.

To first order in S_{21},

$$F_1(-Z_0,n;j)\approx F_1^{(0)}(-Z_0,n)+F_1^{(1)}(-Z_0,n;j), \tag{46}$$

where $F_1^{(0)}(-Z_0,n)$ is the pure soliton solution,

$$(-Z_0)^nF_1^{(0)}(-Z_0,n)=1+h(n)/(Z_1+Z_0^{-1}). \tag{47}$$

Here

$$h(n)=\gamma_1(0)Z_1^{2n+1}(Z_1^{-1}-Z_1)\times[(Z_1^{-1}-Z_1)+\gamma_1(0)Z_1^{2n+1}]^{-1}, \tag{48}$$

and $\gamma_1(0)$ is the initial value of the bound state normalization constant. The first order term is given by

$$\begin{aligned}(-Z_0)^nF_1^{(1)}(-Z_0,n;j)&=h(n)(Z_1+Z_0^{-1})^{-1}(2\pi i)^{-1}\oint dZ\{(Z-Z_1^{-1})^{-1}S_{21}(Z,0)Z^{2n}\\&\quad\times\exp[jp(Z)][1+h(n)/(Z_1-Z^{-1})]\}\\&\quad+(2\pi i)^{-1}\oint dZ\{(Z+Z_0^{-1})^{-1}S_{21}(Z,0)Z^{2n}\\&\quad\times\exp[jp(Z)][1+h(n)/(Z_1-Z^{-1})]\},\end{aligned} \tag{49}$$

where

$$p(Z)=2\ln Z+(Z^{-1}-Z)\tau_1. \tag{50}$$

The only time dependence in Eq. (49) is the exponential factor $\exp[jp(Z)]$. Assuming that $S_{21}(Z,0)$ can be analytically continued into the unit circle $|Z|<1$, except at $Z_1>0$, where it may have a simple pole, we may apply steepest descent method to obtain asymptotic behavior of the integrals in Eq. (49) for a fixed n at $j\gg n$. The

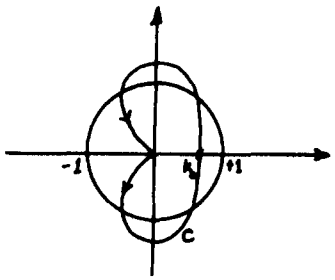


FIG. 1. Contour of steepest descent.

saddle point is at

$$k_0 = v_1 - (v_1^2 - 1)^{1/2} \equiv \exp(-\kappa_0). \tag{51}$$

It is easily seen that

$$1 > k_0 > Z_1. \tag{52}$$

The contour of integration is shown in Fig. 1, starting from the origin into the third quadrant, then passing through the saddle point k_0 perpendicular to the real axis, and going back to the origin from the second quadrant. The asymptotic behavior of an integral of the form

$$I(n,j) = \frac{1}{2\pi i} \oint_C dZ f(Z) Z^{2n} \exp[jp(Z)] \tag{53}$$

for large j is

$$I(n,j) \sim (4\pi j)^{-1/2}(1 - \tau_1^2)^{-1/4} k_0^{2n+1} f(k_0) \times \exp[-2nk_0 - 2j(\kappa_0 - \tau_1 \sinh\kappa_0)]. \tag{54}$$

Noting that $\kappa_0 - \tau_1 \sinh\kappa_0 > 0$, and substituting the appropriate form of $f(Z)$ for Eq. (49), we see that the contribution from the continuum is eventually a backward moving wave (relative to the soliton) whose magnitude decays as $t^{-1/2}$. Physically, the continuum cannot quite catch up with the soliton.

The effect of the continuum spectrum on a soliton moving backwards can be investigated similarly with a different contour. The result is similar.

IV. CONCLUSION

We have shown how the Volterra competition equation can be solved within the framework of the inverse scattering theory of the discrete Schrödinger equation. The Zakharov–Shabat formalism of inverse scattering theory is shown to be more suitable for this purpose. N-soliton solutions are derived and the continuum spectrum is shown to decay as $t^{-1/2}$. We remark that when there is a bound state at $Z = \pm 1$, the effect of the continuum may be quite different. Finally, we note that another nonlinear network equation in Refs. 1, 2,

$$\frac{\partial}{\partial t} U(n) = U(n)^2[W(n-1) - W(n)],$$
$$\frac{\partial}{\partial t} W(n) = W(n)^2[U(n) - U(n+1)], \tag{55}$$

with boundary conditions $\lim_{|n|\to\infty} U(n) = \lim_{|n|\to\infty} W(n) = 1$, can also be solved by the inverse scattering of Eq. (7) by setting

$$U(n+1) = [f_1(-1,n)]^{-2} \tag{56}$$

and

$$W(n) = -2\alpha(n) f_1(-1, n-1) f_1(-1, n). \tag{57}$$

*Supported by the National Science Foundation under Grant MPS74-13252.
†Supported by the Air Force Office of Scientific Research, Grant 72-2187.

[1]R. Hirota, Prog. Theor. Phys. 54, 288 (1975).
[2]R. Hirota and J. Satsuma, J. Phys. Soc. Japan 40, 891 (1976).
[3]The form of the transformation, Eq. (5), is slightly different from Ref. 1 because we have put the Toda equations differently.
[4]H. Flaschka, Prog. Theor. Phys. 51, 703 (1974).
[5]M. Kac and P. van Moerbeke, Advan. Math. 16, 160 (1975).
[6]Equation (9) can be casted in different but equivalent forms. There is no advantage to any particular one. The constant δ, for example, can be any function of λ without affecting the result.
[7]Of course, regular $\alpha(n)$ and $\beta(n)$ corresponding to the Toda lattice have, in general, a discrete spectrum on the real interval $1 \geq Z \geq -1$. This difference in the discrete spectrum between the Toda lattice and Eq. (3) shows that a regular solution of the former may correspond to a singular solution of the latter.
[8]For a detailed theory of the inverse scattering problem of Eq. (7), we refer the readers to K.M. Case, J. Math. Phys. 14, 916 (1973), and K.M. Case and S.C. Chiu, J. Math. Phys. 14, 1643 (1973).
[9]V.E. Zakharov and A.B. Shabat, Zh. Eksp. Teor. Fiz. 61, 118 (1971) [Sov. Phys. JETP 34, 62 (1972)].
[10]As shown in Ref. 8, this is the normalization constant of $f_2(Z_1, n)$, and is therefore positive.
[11]M.J. Ablowitz and A.C. Newell, J. Math. Phys. 14, 1277 (1973).
[12]K.M. Case, J. Math. Phys. 17, 1703 (1976).

General relativistic magnetofluids

F. Paul Esposito*

Department of Physics, University of Cincinnati, Cincinnati, Ohio 45221

E. N. Glass†

Department of Physics, University of Windsor, Windsor, Ontario N9B 3P4, Canada
(Received 28 September 1976)

General relativistic hydromagnetic systems are studied in the infinite conductivity limit. It is shown that when the frozen-in magnetic field lies in surfaces of constant pressure, the fluid acceleration and magnetic field are orthogonal. The theorem is proved that a shear-free infinite conductivity magnetofluid is irrotational if and only if the Weyl tensor is pure electric type. A restricted steady state is formulated and discussed, and within that state it is shown that an infinite conductivity magnetofluid cannot be electrically neutral.

1. INTRODUCTION

General relativistic hydromagnetic systems are physically important in several contexts: cosmology, the late stages of stellar collapse, and equilibrium configurations of compact objects. In this work, we consider the idealized case of a single component, infinitely conducting hydromagnetic system and derive some general properties of such configurations. The results obtained here are a natural extension of earlier studies of shear-free perfect fluids[1,2] and stationary electrovac spaces.[3]

It is shown that when the frozen-in magnetic field lies in surfaces of constant pressure, the fluid acceleration and magnetic field are orthogonal. The theorem is proved that a necessary and sufficient condition for a shear-free magnetofluid to be irrotational is a pure electric type Weyl tensor. The conservation of magnetic energy is established for shear-free expanding motions. Finally, the notion of a restricted steady state is formulated and discussed. It is shown that when the spacetime has no Killing vectors and the steady state is defined with respect to the only preferred timelike congruence available, namely the fluid 4-velocity, then the magnetic field and fluid vorticity must be aligned and thus form a restricted class of steady motions.

All necessary kinematic relations are collected in appendix.

2. INFINITE CONDUCTIVITY MAGNETOFLUID

A perfect fluid is described by the energy–momentum tensor

$$t_{\alpha\beta} = w u_\alpha u_\beta - p\gamma_{\alpha\beta}, \tag{1}$$

where $u^\alpha u_\alpha = 1$, and $\gamma^\alpha{}_\beta$ is the projection operator onto the 3-spaces quotient to u^α:

$$\gamma_{\alpha\beta} := g_{\alpha\beta} - u_\alpha u_\beta. \tag{2}$$

Projection will be abbreviated by $\perp$ which projects all free indices, i.e.,[4]

$$\perp A_{\alpha\beta;}{}^{\delta\nu} := \gamma_\alpha{}^\mu \gamma_\lambda{}^\nu A_{\mu\beta;}{}^{\delta\lambda}.$$

The kinematic properties of the fluid streamlines are characterized by the usual decomposition for the rate-of-change of the streamline vector[5]:

$$u_{\alpha;\beta} = a_\alpha u_\beta + \omega_{\alpha\beta} + \sigma_{\alpha\beta} + \tfrac{1}{3}\theta\gamma_{\alpha\beta}, \tag{3}$$

with the vorticity vector given by

$$\omega_\mu = \overset{*}{\omega}_{\mu\nu} u^\nu := \tfrac{1}{2}\eta_{\mu\nu\alpha\beta}\omega^{\alpha\beta}u^\nu.$$

The alternating tensor in the quotient space is

$$\eta_{\mu\nu\alpha} := \eta_{\mu\nu\alpha\beta}u^\beta,$$

and has the property

$$\eta_{\mu\alpha\beta}\eta^{\mu\rho\sigma} = -2\gamma^\rho{}_{[\alpha}\gamma^\sigma{}_{\beta]}.$$

When the perfect fluid is an infinite conductivity magnetofluid, the Maxwell field tensor is given by

$$F_{\mu\nu} = -\eta_{\mu\nu\alpha}B^\alpha, \tag{4}$$

where B^α is the magnetic field orthogonal to u^α, and

$$\tfrac{1}{2}F_{\alpha\beta}F^{\alpha\beta} = -B^\alpha B_\alpha =: B^2.$$

(For simplicity it is assumed that the magnetic permeability is the same as that of vacuum.)

The Einstein field equations

$$G_{\mu\nu} = -\kappa[t_{\mu\nu} - (1/4\pi)(F_\mu{}^\alpha F_{\nu\alpha} - \tfrac{1}{4}g_{\mu\nu}F_{\alpha\beta}F^{\alpha\beta})]$$

determine the Ricci tensor:

$$R_{\mu\nu} = -(\kappa/2)[(w + 3p + B^2/4\pi)u_\mu u_\nu + (p - w - B^2/4\pi)\gamma_{\mu\nu} - B_\mu B_\nu/2\pi]. \tag{5}$$

3. EQUATIONS OF MOTION

The source-free part of the Maxwell field equations for infinite conductivity is

$$\overset{*}{F}{}^{\mu\nu}{}_{;\nu} = \overset{\circ}{B}{}^\mu - u^\mu{}_{;\nu}B^\nu + \theta B^\mu - u^\mu B^\nu{}_{;\nu} = 0, \tag{6}$$

with components

$$B^\alpha{}_{;\alpha} = -B^\alpha a_\alpha, \tag{6a}$$

$$\perp \mathcal{L}_u B^\mu + \theta B^\mu = 0. \tag{6b}$$

Equation (6b) can be rewritten as

$$\mathcal{L}_u F_{\mu\nu} = 0, \tag{6c}$$

which expresses the frozen-in property of the magnetic field. The 4-current $j^\mu = qu^\mu$ satisfies

$$F^{\mu\nu}{}_{;\nu} = -4\pi j^\mu, \tag{7}$$

Copyright © 1977 American Institute of Physics

with components

$$2\omega^{\alpha}B_{\alpha} = -4\pi q, \tag{7a}$$

$$\perp B_{[\mu;\nu]} = B_{[\mu}a_{\nu]}. \tag{7b}$$

The magnetofluid equations of motion follow from

$$T^{\mu\nu}{}_{;\nu} = 0, \tag{8}$$

with

$$T^{\mu\nu} = (w + B^2/8\pi)u^{\mu}u^{\nu} - (p + B^2/8\pi)\gamma^{\mu\nu} - B^{\mu}B^{\nu}/4\pi.$$

The timelike component of Eq. (8) is

$$\mathring{w} + (w+p)\theta + (1/8\pi)[(B^2)^{\circ} + \tfrac{4}{3}B^2\theta + 2B_{\alpha}B_{\beta}\sigma^{\alpha\beta}] = 0,$$

or, using Eq. (15) below,

$$\mathring{w} + (w+p)\theta = 0, \tag{8a}$$

and the spatial components are

$$(w + p + B^2/8\pi)a_{\mu} = \gamma_{\mu}{}^{\alpha}\nabla_{\alpha}(p + B^2/8\pi) - B_{\mu}(B_{\alpha}a^{\alpha})/8\pi + \gamma^{\alpha}{}_{\mu}\nabla_{(\alpha}B_{\beta)}B^{\beta}/4\pi, \tag{8b}$$

upon using Eqs. (6a) and (7b). When Eq. (8b) is contracted with B^{μ}, one obtains the result

$$(w+p)a_{\mu}B^{\mu} = B^{\mu}\nabla_{\mu}p, \tag{9}$$

which implies that when the magnetic field lies in surfaces of constant pressure, the fluid acceleration and magnetic field are orthogonal.

4. WEYL TENSOR

The electric and magnetic parts of the Weyl tensor are defined with respect to the magnetofluid:

$$E_{\alpha\beta} := C_{\alpha\mu\beta\nu}u^{\mu}u^{\nu}, \tag{10a}$$

$$B_{\alpha\beta} := \overset{*}{C}_{\alpha\mu\beta\nu}u^{\mu}u^{\nu}. \tag{10b}$$

Both tensors are symmetric, trace-free, and orthogonal to u^{α}.

Theorem: A shear-free infinite conductivity magnetofluid is irrotational if and only if the Weyl tensor is pure electric type.

Proof: Necessity follows when $\omega_{\mu} = \sigma_{\mu\nu} = 0$ in Eq. (A4). To prove sufficiency, consider Eq. (A8) with $B_{\mu\nu} = \sigma_{\mu\nu} = 0$:

$$[3E_{\alpha\beta} + \kappa(w + p + B^2/8\pi)\gamma_{\alpha\beta} - (\kappa/8\pi)B_{\alpha}B_{\beta}]\omega^{\beta} = 0, \tag{11}$$

where Eq. (5) has been used to evaluate the Ricci tensor term. When the Weyl tensor is either pure electric or pure magnetic type it must be[6] Petrov type I, D, or O. It follows that

$$\det[3E_{\alpha\beta} + \kappa(w + p + B^2/8\pi)\gamma_{\alpha\beta} - (\kappa/8\pi)B_{\alpha}B_{\beta}] \neq 0 \tag{12}$$

since the vanishing of the determinant implies the eigenvalue equation

$$3E_{\alpha\beta} - (\kappa/8\pi)B_{\alpha}B = -\kappa(w + p + B^2/8\pi)\gamma_{\alpha\beta}, \tag{13}$$

which leads to a contradiction upon taking the trace. Thus $\omega^{\alpha} = 0$. ■

5. CONSERVATION LAWS

Charge conservation implies

$$\nabla_{\alpha}(\omega_{\beta}B^{\beta}u^{\alpha}) = 0. \tag{14}$$

The frozen-in property of the magnetic field allows one to write

$$\gamma^{\mu\alpha}\gamma^{\nu\beta}\pounds_u(F_{\mu\nu}F_{\alpha\beta}) = 0,$$

and since

$$\pounds_u\gamma^{\mu\alpha} = -2[a^{(\mu}u^{\alpha)} + \sigma^{\mu\alpha} + \tfrac{1}{3}\theta\gamma^{\mu\alpha}]$$

it follows that

$$\pounds_u(B^2) + \tfrac{4}{3}\theta B^2 + 2B_{\alpha}B_{\beta}\sigma^{\alpha\beta} = 0. \tag{15}$$

Furthermore, when any conserved scalar n exists (such as charge density or particle number density) satisfying

$$\nabla_{\alpha}(nu^{\alpha}) = 0, \tag{16}$$

which can be rewritten as

$$\pounds_u(n^{-4/3}) = \tfrac{4}{3}n^{-4/3}\theta,$$

then Eq. (15) becomes

$$\pounds_u(n^{-4/3}B^2) + 2n^{-4/3}B_{\alpha}B_{\beta}\sigma^{\alpha\beta} = 0. \tag{17}$$

For shear-free motion (with possible expansion) the conservation of magnetic energy is generalized to

$$\pounds_u(n^{-4/3}B^2) = 0,$$

or equivalently

$$\nabla_{\alpha}(n^{-1/3}B^2u^{\alpha}) = 0. \tag{18}$$

$\dim(n^{-1/3}B^2) =$ energy/area, and so Eq. (18) expresses the fact that the magnetic energy per particle or per charge is "frozen-in" to the fluid.

6. RESTRICTED STEADY STATE

It is clear from Eq. (17) that the presence of differential rotation causes the magnetic energy to be exchanged back and forth for fluid energy. Using the magnetic energy equation (17) and charge conservation, we define the restricted steady state as

$$\perp\dot{B}^{\mu} := \perp\dot{\omega}^{\mu} := \theta := 0, \tag{19}$$

and are thereby led to distinuish two classes of restricted steady motions:

(a) differential rotation with the magnetic field an eigenvector of the shear with zero eigenvalue

$$B^{\mu}\sigma_{\mu\nu} = 0; \tag{20a}$$

(b) rigid rotation

$$\sigma_{\mu\nu} = 0. \tag{20b}$$

The steady state condition of Banerji[7] and Yodzis,[8] $\mathring{\omega}_{\mu\nu} = 0$, is too special since it implies $\perp a_{[\mu;\nu]} = 0$ by Eq. (A6). In general a_{μ} will not be curl-free in the quotient space. It is sufficient to require $\perp\mathring{\omega}_{\mu} = 0$.

The restricted steady state condition (19) and (20) together with Maxwell's equation (6b) requires

$$B^{\alpha}\omega_{\alpha\beta} = 0. \tag{21}$$

Since $B^{\alpha}\omega_{\alpha\beta}\eta^{\beta\mu\nu} = 2B^{[\mu}\omega^{\nu]}$, it follows that B^{μ} and ω^{μ} must be colinear in the restricted steady state.[9] In addition, Maxwell's equation (7a) implies the following.

Lemma: An infinite conductivity magnetofluid cannot be electrically neutral in the restricted steady state.

The reason for calling the steady state "restricted" here, is apparent from Eq. (21) and the subsequent result that the magnetic field and fluid vorticity must be aligned. This is a subclass of possible steady state Newtonian motions, for which the magnetic field and vorticity are generally not aligned. A steady state is most appropriately defined with respect to an inertial frame. In curved space the appropriate frames are locally nonrotating, and correspond to families of spacelike hypersurfaces. Stationary hydromagnetic spacetimes admit natural spacelike foliations. Such spacetimes will be considered in subsequent work, and it is in that context that the generalization of special Newtonian motions, such as those described by Ferraro's law of isorotation,[10] are best discussed.

ACKNOWLEDGMENT

We are grateful for the hospitality of the Aspen Center for Physics, where this work was completed.

APPENDIX: KINEMATIC RELATIONS

With field equations unspecified, the Ricci identities provide kinematic propagation equations for the vorticity, shear, and expansion of a fluid congruence. The Bianchi identities provide the kinematic equations for the electric and magnetic parts of the Weyl tensor. Sign conventions are fixed by Ricci's identity

$$2u_{\nu;[\alpha\beta]} \equiv u_\mu R^\mu{}_{\nu\alpha\beta},$$

and $R_{\nu\alpha} := R^\beta{}_{\nu\alpha\beta}$. The Riemann and Weyl tensors are related by

$$R^{\mu\nu}{}_{\alpha\beta} = C^{\mu\nu}{}_{\alpha\beta} - 2\delta^{[\mu}{}_{[\alpha}R^{\nu]}{}_{\beta]} + \tfrac{1}{3}\delta^{\mu}{}_{[\alpha}\delta^{\nu}{}_{\beta]}R.$$

Ricci's identity yields the following kinematic relations:

$$R_{\alpha\beta}u^\alpha u^\beta = \dot\theta + \tfrac{1}{3}\theta^2 + 2\sigma^2 - 2\omega^2 - a^\alpha{}_{;\alpha}, \tag{A1}$$

$$-\tfrac{1}{2}\eta_{\mu\rho\sigma}R^\mu{}_\nu u^\nu = \perp\omega_{[\rho;\sigma]} - 2\omega_{[\rho}a_{\sigma]} + \tfrac{1}{2}\eta_{\alpha\rho\sigma}(\sigma^{\alpha\beta}{}_{;\beta} + a_\beta\sigma^{\alpha\beta} - \tfrac{2}{3}\theta^{,\alpha}), \tag{A2}$$

$$E_{\mu\nu} = \tfrac{1}{2}(\perp R_{\mu\nu} - \tfrac{1}{3}\gamma_{\mu\nu}R_{\alpha\beta}\gamma^{\alpha\beta}) + \perp a_{(\mu;\nu)} - a_\mu a_\nu - \perp\dot\sigma_{\mu\nu} - \tfrac{2}{3}\theta\sigma_{\mu\nu} - \sigma_{\mu\alpha}\sigma_\nu{}^\alpha + \omega_{\mu\alpha}\omega_\nu{}^\alpha + \tfrac{1}{3}\gamma_{\mu\nu}(2\sigma^2 - 2\omega^2 - a^\alpha{}_{;\alpha}), \tag{A3}$$

$$B_{\mu\nu} = \perp\omega_{(\mu;\nu)} + \gamma_{\mu\nu}\omega_\alpha a^\alpha - \eta^{\alpha\beta}{}_{(\mu}\gamma^\lambda{}_{\nu)}\sigma_{\lambda\alpha;\beta} - 2a_{(\mu}\omega_{\nu)}, \tag{A4}$$

$$\omega^\alpha{}_{;\alpha} + 2a_\alpha\omega^\alpha = 0, \tag{A5}$$

$$\perp a_{[\mu;\nu]} = \perp\dot\omega_{\mu\nu} + \tfrac{2}{3}\theta\omega_{\mu\nu} - 2\omega_{\alpha[\mu}\sigma_{\nu]}{}^\alpha = \pounds_u\omega_{\mu\nu}. \tag{A6}$$

The Bianchi identities

$$\nabla^\mu C_{\mu\nu\alpha\beta} = R_{\nu[\alpha;\beta]} - \tfrac{1}{6}g_{\nu[\alpha}R_{,\beta]},$$

and the Weyl tensor expressed in terms of its electric and magnetic parts ($g_{\mu\nu\alpha\beta} := g_{\mu\alpha}g_{\nu\beta} - g_{\mu\beta}g_{\nu\alpha}$),

$$C_{\mu\nu\rho\sigma} = (g_{\mu\nu\alpha\beta}g_{\rho\sigma\lambda\gamma} - \eta_{\mu\nu\alpha\beta}\eta_{\rho\sigma\lambda\gamma})u^\alpha u^\lambda E^{\beta\gamma} - (g_{\mu\nu\alpha\beta}\eta_{\rho\sigma\lambda\gamma} + \eta_{\mu\nu\alpha\beta}g_{\rho\sigma\lambda\gamma})u^\alpha u^\lambda B^{\beta\gamma},$$

yield the kinematic relations for $E_{\alpha\beta}$ and $B_{\alpha\beta}$ by straightforward calculation:

$$\perp\nabla^\mu E_{\mu\alpha} = \perp R_{\nu[\alpha;\beta]}u^\nu u^\beta + \tfrac{1}{12}\gamma_\alpha{}^\beta R_{,\beta} - a^\beta E_{\alpha\beta} + 3\omega^\beta B_{\alpha\beta} - \eta_{\alpha\beta\nu}\sigma^\beta{}_\mu B^{\mu\nu}, \tag{A7}$$

$$\perp\nabla^\mu B_{\mu\alpha} = \tfrac{1}{2}\eta_\alpha{}^{\mu\nu}u^\beta R_{\beta\mu;\nu} - a^\beta B_{\alpha\beta} - 3\omega^\beta E_{\alpha\beta} + \eta_{\alpha\beta\nu}\sigma^\beta{}_\mu E^{\mu\nu}, \tag{A8}$$

$$\perp\dot E_{\alpha\beta} = 3\sigma_{\mu(\alpha}E^\mu{}_{\beta)} - \gamma_{\alpha\beta}\sigma^{\mu\nu}E_{\mu\nu} - \theta E_{\alpha\beta} - \omega^\mu{}_{(\alpha}E_{\beta)\mu} - 2a^\mu B^\nu{}_{(\alpha}\eta_{\beta)\mu\nu} + \eta_{\mu\nu(\alpha}\nabla^\mu B_{\beta)}{}^\nu + \tfrac{1}{2}\perp[u^\mu\nabla_{(\alpha}R_{\beta)\mu} - \dot R_{\alpha\beta} + \tfrac{1}{6}g_{\alpha\beta}\dot R], \tag{A9}$$

$$\perp\dot B_{\alpha\beta} = 3\sigma_{\mu(\alpha}B^\mu{}_{\beta)} - \gamma_{\alpha\beta}\sigma^{\mu\nu}B_{\mu\nu} - \theta B_{\alpha\beta} - \omega^\mu{}_{(\alpha}B_{\beta)\mu} + 2a^\mu E^\nu{}_{(\alpha}\eta_{\beta)\mu\nu} - \eta_{\mu\nu(\alpha}\nabla^\mu E_{\beta)}{}^\nu - \tfrac{1}{2}\perp[\nabla_\nu(R_{\mu(\alpha}\eta_{\beta)}{}^{\mu\nu}) - u^\mu R_{\mu(\alpha}\omega_{\beta)} + R_{\mu(\alpha}\eta_{\beta)}{}^{\mu\nu}a_\nu]. \tag{A10}$$

(The vacuum equations must be preserved under $E_{\alpha\beta} \to B_{\alpha\beta}$, $B_{\alpha\beta} \to -E_{\alpha\beta}$.)

*Supported in part by NSF Grant INT 75-18457.
†Supported in part by an NRC of Canada grant.

[1] E.N. Glass, J. Math. Phys. 15, 1930 (1974).
[2] E.N. Glass, J. Math. Phys. 16, 2361 (1975).
[3] F.P. Esposito and E.N. Glass, J. Math. Phys. 17, 282 (1976).
[4] Covariant derivatives are denoted by semicolons and ∇_α. Parentheses around indices denote symmetrization and brackets around indices denote antisymmetrization. The dual operator is $\frac{1}{2}\eta^{\mu\nu\alpha\beta}$ and the dual is denoted by an asterisk. $\pounds_u$ symbolizes the Lie derivative with respect to u^α. An overhead dot denotes $u^\alpha\nabla_\alpha$.
[5] We adhere to all the conventions and definitions of Ref. 2.
[6] P. Jordan, J. Ehlers, and W. Kundt, Akad. Wiss. Lit. Mainz Abh. Math. Naturwiss. Kl., No. 2 (1960).
[7] S. Banerji, Nuovo Cimento B 23, 345 (1974).
[8] P. Yodzis, Phys. Rev. D 3, 2941 (1971).
[9] Since $\pounds_u\omega_{\alpha\beta} = \theta\eta_{\alpha\beta\mu}\omega^\mu + \eta_{\alpha\beta\mu}\pounds_u\omega^\mu$, it follows from the restricted steady state conditions (19) and (20a), and the alignment of B^μ and ω^μ, that $\pounds_u\omega_{\alpha\beta} = 0$. This in turn implies $\perp a_{[\alpha;\beta]} = 0$, which is a strong restriction on the magnetofluid. When the spacetime admits a timelike Killing field, and the steady state is defined with respect to its locally nonrotating frames, then curl a need not vanish.
[10] V.C.A. Ferraro, Monthly Not. R. Astron. Soc. 97, 458 (1937).

Rotating steady state configurations in general relativity

E. N. Glass*

Department of Physics, University of Windsor, Windsor, Ontario N9B 3P4, Canada
(Received 29 November 1976)

Spacetimes with axial and timelike Killing fields η^μ and ξ^μ respectively are studied. A natural basis is constructed for the steady state configurations of matter with 4-velocity $u^\mu = \alpha\xi^\mu + \beta\eta^\mu$. It is shown that the rate of shear tensor always has an eigenvector with zero eigenvalue lying in the 2-space of Killing orbits. A natural coordinate system is constructed on the 2-space of Killing orbits for rigidly rotating perfect matter. Gravitational isorotation is proved for the case of a perfect fluid with a one-parameter equation of state. Locally nonrotating frames are constructed for purely stationary spacetimes.

1. INTRODUCTION

There has been considerable recent progress in the study of stationary, axially symmetric, configurations of self-gravitating matter. Geroch[1] has developed a formalism which "factors out" the two Killing vectors. The dynamical problem then becomes one of solving Einstein's equations on the two-dimensional manifold of Killing orbits. Working in this formalism, Winicour[2] has found all the dust metrics by reducing the problem to quadratures. The known exact perfect fluid solutions have recently been surveyed by Krasinski,[3] and there is a new study of the matching problem with emphasis on the Kerr metric and its possible interiors by Roos.[4] As yet, no one has found a rotating perfect fluid solution which can be matched to an asymptotically flat exterior.

In this work the kinematics of the matter orbits are studied for the case where the orbit tangents are a linear combination of the timelike and axial Killing vectors. The field equations are used weakly and only for *perfect matter*—matter whose 4-velocity is an eigenvector of the Ricci tensor, thus including both perfect fluids and perfect magnetofluids.

A natural basis is constructed for the spacetime, and it is shown that the rate of shear tensor always has an eigenvector with zero eigenvalue lying in the 2-space of Killing orbits. For perfect matter in rigid rotation, a natural coordinate system is constructed on the 2-space of Killing orbits. Relaxing the condition of axial symmetry, it is shown how to construct a nonrotating frame for any stationary spacetime.

2. THE SPACETIME

Consider a spacetime $(M, g_{\alpha\beta})$ admitting a pair of commuting Killing vectors ξ^α and η^α. ξ^α is everywhere timelike and has open timelike lines as orbits. η^α is everywhere spacelike with closed rotational orbits (topologically circles) for the parameter range $0 \leq \varphi \leq 2\pi$, where $\eta^\alpha = dx^\alpha/d\varphi$. It is required that the spacetime be asymptotically flat at spatial infinity, and in the Newtonian limit of a fluid it is required that

$$\lambda_{00} := \xi^\alpha \xi_\alpha \to -1 + 2U,$$

where U is the Newtonian potential and $U \leq 0$,

$$\lambda_{01} := \xi^\alpha \eta_\alpha \to 0,$$

$$\lambda_{11} := \eta^\alpha \eta_\alpha \to -\rho^2,$$

where ρ is the orthogonal distance from the rotation axis.

To describe rotating steady state configurations, it is assumed that the matter orbits have unit tangent

$$u^\mu = \alpha\xi^\mu + \beta\eta^\mu, \tag{1}$$

with zero Lie derivative with respect to each Killing vector, and whose rate of change is given by[5]

$$u_{\mu;\nu} = a_\mu u_\nu + \omega_{\mu\nu} + \sigma_{\mu\nu} + \tfrac{1}{3}\theta\gamma_{\mu\nu}, \tag{2}$$

where

$$\gamma_{\mu\nu} := g_{\mu\nu} - u_\mu u_\nu, \quad a_\mu := u_{\mu;\nu}u^\nu,$$

$$\theta := u^\mu{}_{;\mu}, \quad \omega_{\mu\nu} := u_{[\mu;\nu]} - a_{[\mu}u_{\nu]},$$

$$\sigma_{\mu\nu} := u_{(\mu;\nu)} - a_{(\mu}u_{\nu)} - \tfrac{1}{3}\theta\gamma_{\mu\nu}.$$

Here $\theta = 0$ as a consequence of Eq. (1). Since u^μ has unit length, it follows that

$$\alpha^2\lambda_{00} + 2\alpha\beta\lambda_{01} + \beta^2\lambda_{11} = 1. \tag{3}$$

It is now useful to construct a spacelike vector V^μ orthogonal to u^μ. Define the scalars[6]

$$\delta := \xi^\alpha u_\alpha = \alpha\lambda_{00} + \beta\lambda_{01}, \tag{4a}$$

$$-\gamma := \eta^\alpha u_\alpha = \alpha\lambda_{01} + \beta\lambda_{11}, \tag{4b}$$

and the vector

$$V^\mu := \gamma\xi^\mu + \delta\eta^\mu. \tag{5}$$

As a consequence of the definitions of γ and δ, V^μ is orthogonal to u^μ and

$$\alpha\delta - \beta\gamma = 1.$$

The norm of the spacelike vector V^μ is

$$V^2 := -V^\mu V_\mu = \lambda_{01}^2 - \lambda_{00}\lambda_{11}, \tag{6}$$

and it is assumed that V is zero only on the axis of rotation.[7]

The orbits of ξ^μ and η^μ form a two-dimensional manifold with induced metric

$$h_{\mu\nu} = g_{\mu\nu} + V^{-2}[\lambda_{11}\xi_\mu\xi_\nu + \lambda_{00}\eta_\mu\eta_\nu - 2\lambda_{01}\xi_{(\mu}\eta_{\nu)}],$$

which can be rewritten as

$$h_{\mu\nu} = g_{\mu\nu} - u_\mu u_\nu + \hat{V}_\mu\hat{V}_\nu, \tag{7}$$

where unit vectors will henceforth be denoted by an overhead caret (with the exception of u^μ), i.e.,

$$\hat{V}_\mu := V^{-1}V_\mu.$$

 Copyright © 1977 American Institute of Physics

The dynamical properties of the spacetime can be totally described by Einstein's equations on the two-dimensional manifold[8] with metric $h_{\mu\nu}$; however, the matter flow has rate of shear and vorticity with components in the 3-space of metric $\gamma_{\mu\nu}$, and so our considerations will not be restricted to the manifold of Killing trajectories. In the following, we will refer to "the 2-space" and "the 3-space" with the meaning above.

The alternating tensor on the 3-space is

$$\eta_{\mu\alpha\beta} := \eta_{\mu\alpha\beta\nu}u^{\nu},$$

with the property

$$\eta_{\mu\alpha\beta}\eta^{\mu\rho\sigma} = -2\gamma^{\rho}{}_{[\alpha}\gamma^{\sigma}{}_{\beta]}.$$

The alternating tensor on the 2-space is

$$\epsilon_{\mu\nu} := \eta_{\mu\nu\alpha\beta}u^{\alpha}\hat{V}^{\beta} = V^{-1}\eta_{\mu\nu\alpha\beta}\xi^{\alpha}\eta^{\beta},$$

with the property

$$\epsilon_{\mu\alpha}\epsilon^{\mu\beta} = h_{\alpha}{}^{\beta}.$$

There are eight scalars $\alpha, \beta, \gamma, \delta, \lambda_{00}, \lambda_{01}, \lambda_{11}$, and V, connected by four relations, Eqs. (3), (4), and (6). We choose the set α, β, γ, δ, V to describe the properties of the rotating steady state configurations (among α, β, γ, δ only three are independent) while other authors[8] have chosen $\lambda_{00}, \lambda_{01}, \lambda_{11}$, and $\Omega := \beta/\alpha$. The relations connecting the scalars can be inverted to give

$$\lambda_{00} = \delta^2 - V^2\beta^2, \tag{8a}$$

$$\lambda_{01} = V^2\alpha\beta - \gamma\delta, \tag{8b}$$

$$\lambda_{11} = \gamma^2 - V^2\alpha^2. \tag{8c}$$

3. KINEMATICS

A. Acceleration

Killing's equations and the commutativity of ξ^{μ} and η^{μ} imply that

$$\xi^{\alpha}\nabla_{\alpha}\xi_{\beta} = -\tfrac{1}{2}\nabla_{\beta}(\lambda_{00}), \tag{9a}$$

$$\eta^{\alpha}\nabla_{\alpha}\eta_{\beta} = -\tfrac{1}{2}\nabla_{\beta}(\lambda_{11}), \tag{9b}$$

$$\eta^{\alpha}\nabla_{\alpha}\xi_{\beta} = \xi^{\alpha}\nabla_{\alpha}\eta_{\beta} = -\tfrac{1}{2}\nabla_{\beta}(\lambda_{01}). \tag{9c}$$

Therefore, using Eqs. (3), (4), and the definition of acceration

$$a_{\mu} = \delta\alpha_{,\mu} - \gamma\beta_{,\mu}. \tag{10}$$

For rigid rotation, with $\beta = \Omega_0\alpha$,

$$a_{\mu} = (\log\alpha)_{,\mu}.$$

B. Shear

From the definition of the rate of shear and Eqs. (1) and (5) it is seen that

$$\sigma_{\mu\nu} = \alpha^2 V_{(\mu}W_{\nu)}, \tag{11}$$

where the "differential rotation vector" W_{ν} is defined as

$$W_{\nu} := (\beta/\alpha)_{,\nu}. \tag{12}$$

A necessary and sufficient condition for rigid rotation is that W_{ν} vanish, i.e., $\beta = \Omega_0\alpha$, where Ω_0 is constant. When α and β are functionally dependent, then the relation

$$a_{[\mu}W_{\nu]} = (\alpha^{-2})\,\alpha_{,[\mu}\beta_{,\nu]}$$

implies that a_{μ} and W_{μ} must be parallel. With $W^2 := -W^{\alpha}W_{\alpha}$, the shear scalar is

$$\sigma^2 := \tfrac{1}{2}\sigma_{\mu\nu}\sigma^{\mu\nu} = \tfrac{1}{4}\alpha^4V^2W^2. \tag{13}$$

The rate of shear tensor has two eigenvectors which lie in the 3-space and an eigenvector with zero eigenvalue which lies in the 2-space. Let ${}_{(3)}\hat{e}_{\mu}$ be a unit vector in the 2-space and orthogonal to W^{μ}, i.e.,

$${}_{(3)}\hat{e}_{\mu}u^{\mu} = {}_{(3)}\hat{e}_{\mu}V^{\mu} = {}_{(3)}\hat{e}_{\mu}W^{\mu} = 0;$$

then the spectral decomposition of $\sigma_{\mu\nu}$ is

$$\sigma_{\mu\nu} = \tfrac{1}{2}\alpha^2VW[{}_{(1)}\hat{e}_{\mu}\,{}_{(1)}\hat{e}_{\nu} - {}_{(2)}\hat{e}_{\mu}\,{}_{(2)}\hat{e}_{\nu}], \tag{14}$$

where

$${}_{(1)}\hat{e}_{\mu} := (2)^{-1/2}(\hat{V}_{\mu} + \hat{W}_{\mu}), \quad {}_{(2)}\hat{e}_{\mu} := (2)^{-1/2}(\hat{V}_{\mu} - \hat{W}_{\mu}).$$

The ${}_{(i)}\hat{e}_{\mu}$ form an orthonormal triad of eigenvectors with ${}_{(3)}\hat{e}_{\mu}$ having zero eigenvalue. A nonshearing spatial direction exists since the matter flow lies partly along the orbits of the axial Killing vector.

C. Vorticity

From the definition of the vorticity bivector, Eqs. (1), (5), (9), $\omega_{\mu\nu}\xi^{\nu} = -(\beta/\alpha)\,\omega_{\mu\nu}\xi^{\mu}\eta^{\nu}$, and $\omega_{\mu\nu}\xi^{\mu}\eta^{\nu} = 0$, it follows that

$$\omega_{\mu\nu} = \delta^2V^{-2}V_{[\mu}X_{\nu]}, \tag{15}$$

where the "local rotation arm" X_{ν} is defined as

$$X_{\nu} := (\gamma/\delta)_{,\nu}, \tag{16}$$

with $X^2 := -X^{\nu}X_{\nu}$. The vorticity vector lies in the 2-space and is given by

$$\omega_{\mu} = \overset{*}{\omega}_{\mu\nu}u^{\nu} = \tfrac{1}{2}\delta^2V^{-1}\epsilon_{\mu\nu}X^{\nu}, \tag{17}$$

and so it is clear that the local rotation arm and the vorticity vector are *always* orthogonal:

$$\omega_{\mu}X^{\mu} = 0.$$

The eigenvectors of $\omega_{\mu\nu}$ are the two complex null vectors $\hat{X}^{\nu} \pm i\hat{V}^{\nu}$ with eigenvalues $\pm(i/2)\,\delta^2V^{-1}X$, and ω^{ν} with eigenvalue zero. The vorticity scalar is

$$\omega^2 := -\omega^{\mu}\omega_{\mu} = \tfrac{1}{2}\omega_{\mu\nu}\omega^{\mu\nu} = \tfrac{1}{4}\delta^4V^{-2}X^2. \tag{18}$$

4. PERFECT FLUID EQUATION OF STATE

The relation

$$a_{[\mu;\nu]} = \alpha^2\delta^2X_{[\mu}W_{\nu]} \tag{19}$$

implies that X_{μ} and the differential rotation vector W_{μ} are parallel for a perfect fluid when a one-parameter equation of state exists and the equations of motion relate the pressure gradient with acceleration. Furthermore, since $X_{\mu} = (\gamma/\delta)_{,\mu}$ and $W_{\mu} = (\beta/\alpha)_{,\mu}$, if one chooses $y = \gamma/\delta$ as a coordinate on the 2-space then the colinearity of X_{μ} and W_{μ} implies the well-known result $\Omega := \beta/\alpha = \Omega(y)$. In this case, the vorticity vector lies along the eigenvector of the shear tensor which has zero eigenvalue. The result $\omega^{\mu}\Omega_{,\mu} = 0$ is the exact analog of Ferraro's Law of Isorotation[9] in the theory of magnetofluids. Ferraro's law states that, for a magnetofluid with a frozen-in magnetic field, Ω is constant along the magnetic field lines in the steady state. In other words, the magnetofluid differentially rotates as a series of

"shells" with each shell a fluid surface spanned by the magnetic field lines. The same interpretation holds for a perfect fluid with a one-parameter equation of state, since the vorticity vector is frozen-in when

$$\underset{u}{\pounds}\omega_{\mu\nu}=0,$$

just as $\underset{u}{\pounds}F_{\mu\nu}=0$ for an infinite conductivity magnetofluid with a frozen-in magnetic field.

5. RIGID ROTATION

For rigid rotation, the flow vector is proportional to a timelike Killing vector:

$$u^\mu=\alpha(\xi^\mu+\Omega_0\eta^\mu).$$

Ricci's identity then yields the equation

$$\omega_{[\mu;\nu]}-2\omega_{[\mu}a_{\nu]}=-\tfrac{1}{2}\eta_{\mu\nu\alpha}R^\alpha{}_\beta u^\beta. \tag{20}$$

Since $a_\mu=\alpha^{-1}\alpha_{,\mu}$, Eq. (20) be rewritten as

$$\alpha^2(\alpha^{-2}\omega_{[\mu})_{;\nu]}=-\tfrac{1}{2}\eta_{\mu\nu\alpha}R^\alpha{}_\beta u^\beta. \tag{21}$$

For perfect matter (u^β an eigenvector of the Ricci tensor) in rigid rotation it follows that $\alpha^{-2}\omega_\mu$ is locally a gradient.

6. THE METRIC

Since the vorticity and local rotation arm are always orthogonal, an orthonormal tetrad has been constructed which provides a natural basis for the spacetime:

$$g_{\mu\nu}=u_\mu u_\nu-\hat{V}_\mu\hat{V}_\nu-\hat{X}_\mu\hat{X}_\nu-\hat{\omega}_\mu\hat{\omega}_\nu. \tag{22}$$

A. Killing coordinates

Choose $x^0=t$ as a parameter along the timelike Killing orbits with $\xi^\mu=\delta^\mu{}_0$, and $x^1=\varphi$ with $\eta^\mu=\delta^\mu{}_1$. The line element can then be written as (in the orthogonally transitive case)

$$ds^2=\lambda_{00}dt^2+2\lambda_{01}dt\,d\varphi+\lambda_{11}d\varphi^2-(\hat{X}_\mu\hat{X}_\nu+\hat{\omega}_\mu\hat{\omega}_\nu)dx^\mu dx^\nu.$$

B. Comoving coordinates

Choose $u^{\mu'}=\alpha\delta^{\mu'}{}_0$ and $\eta^{\mu'}=\delta^{\mu'}{}_1$, which is equivalent to the coordinate transformation from Killing to comoving coordinates given by

$$dT=dt,\quad d\Phi=d\varphi-(\beta/\alpha)dt,$$

where $d\Phi$ is exact only for rigid rotation. The line element becomes

$$ds^2=\alpha^{-2}(dt-\alpha\gamma\,d\Phi)^2-\alpha^2V^2d\Phi^2-X^{-2}(X_\mu X_\nu+4V^2\delta^{-4}\omega_\mu\omega_\nu)dx^\mu dx^\nu. \tag{24}$$

C. Rigid rotation coordinates

For perfect matter in rigid rotation, $\alpha^{-2}\omega_\mu$ is locally a gradient. In this case, there exists a natural coordinate system on the 2-space. Choose

$$\alpha^{-2}\omega_\mu=\Omega_0\nabla_\mu z\quad(-\infty<z<\infty),$$

$$\gamma/\delta=\Omega_0y^2\quad(0\le y<\infty),$$

and Ω_0 positive. In comoving coordinates the line element is then given by

$$ds^2=\alpha^{-2}(dt-\Omega_0Y^2d\Phi)^2-\alpha^2V^2d\Phi^2-4\Omega_0^2X^{-2}(1-\Omega_0^2y^2)^4(y^{-6}Y^8dy^2+V^2\alpha^8dz^2), \tag{25}$$

where $Y^2=y^2(1-\Omega_0^2y^2)^{-1}$ and where $\delta=\alpha^{-1}(1-\Omega_0^2y^2)^{-1}$ follows from $\alpha\delta-\beta\gamma=1$. The choice of coordinate for γ/δ is constrained by the inequalities of Hansen and Winicour[8]

$$\lambda_{01}+\Omega_0\lambda_{11}\le 0,\quad \lambda_{00}+\Omega_0\lambda_{01}>0,$$

which hold both inside and outside the velocity of light surface (see Sec. 8 below). Thus from Eq. (4)

$$\gamma/\delta\ge 0.$$

The factor of Ω_0 appears so that in the Newtonian limit $y\to\rho$, the usual cylindrical coordinate. Transforming coordinates[10] from $y\to Y$ would serve to push the velocity of light surface from $y=1/\Omega_0$ to $Y\to\infty$. There are only three unknown functions in the metric of rigidly rotating perfect matter: α, V, and X, in agreement with Harrison's earlier work.[11]

7. ORTHOGONAL TRANSITIVITY

For completeness, the conditions under which the 2-spaces are imbedded as 2-surfaces will be presented. The commuting Killing vectors ξ^μ and η^μ are orthogonally transitive to a family of spacelike 2-surfaces if the system is composed of perfect matter and has a rotation axis.

Theorem (*Carter*[12]): Consider the simple bivector $W_{\alpha\beta}:=\xi_{[\alpha}\eta_{\beta]}=u_{[\alpha}V_{\beta]}$. If, on an open set D of M,

(a) $\overset{*}{W}_{\alpha\beta}R^{\beta\mu}W_{\mu\nu}=0$,

(b) $W_{\alpha\beta}=0$ at some point in D,

then $W_{\mu\alpha}\overset{*}{W}{}^{\alpha\beta}{}_{;\beta}=0$ on D, which is the necessary and sufficient Frobenius condition that $W_{\alpha\beta}$ be locally orthogonal to a family of 2-surfaces.

Outline of Proof: Define twist scalars

$$c_0:=\overset{*}{W}{}^{\alpha\beta}\xi_{\alpha;\beta},\quad c_1:=\overset{*}{W}{}^{\alpha\beta}\eta_{\alpha;\beta}.$$

It can be shown that $c_0=c_1=0$ is equivalent to the Frobenius condition. Since the scalars vanish on the rotation axis ($W_{\alpha\beta}=0$), and their gradients are given by

$$c_{0,\alpha}=-2\overset{*}{W}_{\alpha\beta}R^{\beta\mu}\xi_\mu,\quad c_{1,\alpha}=-2\overset{*}{W}_{\alpha\beta}R^{\beta\mu}\eta_\mu,$$

it follows that the scalars are globally zero for Ricci tensors invertible in the group of Killing vectors, i.e., for perfect matter when $u^\mu=\alpha\xi^\mu+\beta\eta^\mu$.

8. LOCALLY NONROTATING FRAMES

In any stationary spacetime it is possible to construct a 1-form $t_\mu dx^\mu$ which is closed and which, in a simply connected region, labels a family of spacelike hypersurfaces whose orthogonal trajectories are the world lines of locally non-rotating observers.

Starting with the timelike Killing vector field ξ^α, define its twist vector and twist bivector:

$$\Omega^\mu:=\tfrac{1}{2}\eta^{\mu\nu\rho\sigma}\xi_\nu\xi_{\rho;\sigma},\quad \Omega_{\mu\nu}:=\lambda_{00}^{-1/2}\eta_{\mu\nu\alpha\beta}\xi^\alpha\Omega^\beta.$$

Let s_μ be a solution of the curl equation

$$s_{[\mu;\nu]} = \lambda_{00}^{-3/2}\Omega_{\mu\nu}, \tag{26}$$

with s_μ and ξ^μ orthogonal ($\pounds_\xi s_\mu = 0$). Solutions always exist (locally) since

$$(\eta^{\mu\nu\alpha\beta}\lambda_{00}^{-3/2}\Omega_{\mu\nu})_{;\beta} = 0,$$

which follows from $\pounds_\xi \Omega^\mu = 0$, and $\Omega^\nu{}_{;\nu} - 2\lambda_{00}^{-1}\lambda_{00,\nu}\Omega^\nu = 0$. The closed 1-form is then constructed as

$$t_\mu\, dx^\mu := \lambda_{00}^{-1}\xi_\mu\, dx^\mu - s_\mu\, dx^\mu. \tag{27}$$

Closure follows from $\xi_{\mu;\nu} = \lambda_{00}^{-1}\xi_{[\mu}\lambda_{00,\nu]} + \lambda_{00}^{-1/2}\Omega_{\mu\nu}$.

For the spacetimes considered here with a pair of commuting Killing vectors ξ^μ, η^μ, it is easily seen that (with $c_0 = c_1 = 0$)

$$t_\mu\, dx^\mu = dt, \quad s_\mu\, dx^\mu = (\lambda_{01}/\lambda_{00})\, d\varphi$$

in Killing coordinates, and generally

$$t_\mu = V^{-2}(-\lambda_{11}\xi_\mu + \lambda_{01}\eta_\mu),$$

$$s_\mu = V^{-2}(\lambda_{01}/\lambda_{00})(\lambda_{01}\xi_\mu - \lambda_{00}\eta_\mu).$$

The 4-velocity u^μ can now be written in the Minkowski-like form

$$u^\mu = (1 - v^2)^{-1/2}(\hat{t}^\mu + v\hat{\eta}^\mu). \tag{28}$$

The speed of matter, v, measured by a locally nonrotating observer is given by

$$(1 - v^2)^{-1/2} := u^\mu \hat{t}_\mu,$$

or $v^2 = V^{-2}\lambda_{11}^2(\beta/\alpha + \lambda_{01}/\lambda_{11})^2$, in agreement with Bardeen.[13] Furthermore, writing

$$v^2 = V^{-2}\alpha^{-2}\gamma^2 = (\lambda_{11}/V^2\alpha^2) + 1$$

shows the velocity of light surface given by the limit $\alpha \to \infty$.

9. SUMMARY

A natural basis has been constructed for the spacetime of rotating steady state configurations of matter with 4-velocity $u^\mu = \alpha\xi^\mu + \beta\eta^\mu$, and orthogonal spacelike vector $V^\mu = \gamma\xi^\mu + \delta\eta^\mu$. The orthonormal tetrad satisfies the completeness relation $g_{\mu\nu} = u_\mu u_\nu - \hat{V}_\mu\hat{V}_\nu - \hat{X}_\mu\hat{X}_\nu - \hat{\omega}\,\hat{\omega}_\nu$, where the tetrad orientation is given by

$$\eta_{\mu\nu\alpha\beta}u^\mu\hat{V}^\nu\hat{X}^\alpha\hat{\omega}^\beta = 1.$$

The metric on the 3-space is $\gamma_{\mu\nu} := g_{\mu\nu} - u_\mu u_\nu$, and the metric on the 2-space of Killing orbits is $h_{\mu\nu} := g_{\mu\nu} - u_\mu u_\nu + \hat{V}_\mu\hat{V}_\nu$.

2-space objects	3-space objects
a_μ acceleration	V_μ
ω_μ vorticity vector	$\sigma_{\mu\nu}$ rate of shear
$X_\mu := (\gamma/\delta)_{,\mu}$ local rotation arm	$\omega_{\mu\nu}$ vorticity bivector
$W_\mu := (\beta/\alpha)_{,\mu}$ differential rotation vector	

Kinematic results which follow from weak use of the field equations are:

(A) perfect fluid with 1-parameter equation of state: W_μ and X_μ parallel, $\sigma_{\mu\nu}\omega^\nu = 0$, $\omega^\nu\Omega_{,\nu} = 0$;

(B) perfect matter in rigid rotation: a_μ and $\alpha^{-2}\omega_\mu$ curl-free.

Locally nonrotating frame: In any stationary spacetime one can construct a closed 1-form orthogonal (in simply connected regions) to a family of spacelike hypersurfaces and given by $t_\mu\, dx^\mu = \lambda_{00}^{-1}\xi_\mu\, dx^\mu - s_\mu\, dx^\mu$.

New results: Identification of ω_μ and X_μ as physical vectors which provide a natural basis for the 2-space. Gravitational isorotation, wherein ω^μ is an eigenvector of the shear tensor with zero eigenvalue and $\omega^\mu\Omega_{,\mu} = 0$, for the case of a perfect fluid with a 1-parameter equation of state. The construction of spacelike hypersurfaces in the purely stationary case.

ACKNOWLEDGMENTS

I am grateful to Professor F. P. Esposito and Professor J. Winicour for many valuable discussions, and to L. Lindblom for noticing a serious flaw in the original manuscript.

*Supported in part by an NRC of Canada grant.

[1]R. Geroch, J. Math. Phys. 13, 394 (1972).
[2]J. Winicour, J. Math. Phys. 16, 1806 (1975).
[3]A. Krasinski, Acta Phys. Polon. B 5, 411 (1974); J. Math. Phys. 16, 125 (1975); Acta Phys. Polon. B 6, 223 (1975); Acta Phys. Polon. B 6, 239 (1975).
[4]W. Roos, Gen. Rel. Grav. 7, 431 (1976).
[5]Covariant derivatives are denoted by semicolons and ∇_α; partial derivatives by commas. Parentheses around indices denote symmetrization and brackets around indices denote antisymmetrization. The dual operator is $\frac{1}{2}\eta^{\mu\nu\alpha\beta}$ and the dual is denoted by an asterisk. $\pounds_\xi$ symbolizes the Lie derivative with respect to ξ^α. An overhead caret denotes unit vectors. The sign conventions for the Riemann and Ricci tensors are fixed by $2A_{\mu;[\sigma\tau]} \equiv A_\alpha R^\alpha{}_{\mu\sigma\tau}$, $R_{\mu\nu} := R^\alpha{}_{\mu\nu\alpha}$.
[6]These scalars were first introduced by R. H. Boyer, Proc. Camb. Philos. Soc. 62, 495 (1966).
[7]For axis regularity conditions see B. Carter, Comm. Math. Phys. 17, 233 (1970).
[8]R. Hansen and J. Winicour, J. Math. Phys. 16, 804 (1975).
[9]V. C. A. Ferraro, Monthly Not. R. Astron. Soc. 97, 458 (1937).
[10]A rigid rotation metric with Y, z coordinates (up to trivial change of notation) that range up to the velocity of light surface appears in M. Abramowicz and B. Muchotrzeb, Gen. Rel. Grav. 7, 371 (1976).
[11]B. K. Harrison, Phys. Rev. D 1, 2269 (1970).
[12]B. Carter, J. Math. Phys. 10, 70 (1969).
[13]J. M. Bardeen, Astrophys. J. 162, 71 (1970).

Path integration of a one-dimensional three-body problem with three-body forces

D. C. Khandekar and S. V. Lawande

Theoretical Reactor Physics Section, Bhabha Atomic Research Centre, Trombay, Bombay 400 085, India
(Received 4 May 1976)

A one-dimensional three-body problem with three-body forces is treated by the path integral approach of Feynman. This involves essentially the path integration of a Lagrangian in plane polar coordinates with a potential energy term depending on both the radial and angular coordinates. An exact expression for the propagator is obtained.

1. INTRODUCTION

Exact propagators for a one-dimensional three-body problem with harmonic and inverse square pairwise interactions[1] were obtained by the authors in the earlier papers.[2,3] A nontrivial extension of the previous results is obtained when the three-body problem is generalized to include additional three-body interaction terms.[4] This is discussed in the present paper.

Although the energy eigenvalues and eigenfunctions of the Schrödinger equation corresponding to the problem are available in explicit forms[1,4] and may be used to construct the Green's function or the propagator,[5] it is quite instructive to obtain the propagator as a "sum over paths" as defined by Feynman.[6] This exercise turns out to be quite difficult in general, the three-body problem being no exception. This is the main motivation for this and the earlier work.

An interesting feature of the three-body problem considered is that it involves path integration of a Lagrangian in plane polar coordinates with a potential which depends not only on the radial coordinates but also on the angle. However, the form of the potential is such that the propagator separates out into the radial and angular part. Further, it is possible to evaluate the radial part analytically in a closed form. In short, an exact expression for the propagator is obtained from which the eigenfunctions of the corresponding Schrödinger equation may be recovered. Throughout this paper we use units such as $\hbar = m = 1$, where m is the mass of the particles.

2. FORMULATION OF THE PATH INTEGRAL

The problem under consideration involves three particles of equal mass moving in one dimension and is characterized by the Lagrangian

$$L = \tfrac{1}{2}[(\dot{\xi}^1)^2 + (\dot{\xi}^2)^2 + (\dot{\xi}^3)^2] - (V_1 + V_2 + V_3), \tag{1}$$

where ξ^1, ξ^2, ξ^3 are the coordinates of the three particles, V_1, V_2 are the harmonic and the inverse square pairwise potentials, and V_3 is the three-body potential. These are defined by the following expressions:

$$V_1 = (1/6)\Omega^2[(\xi^1 - \xi^2)^2 + (\xi^2 - \xi^3)^2 + (\xi^3 - \xi^1)^2], \tag{2}$$

$$V_2 = g[(\xi^1 - \xi^2)^{-2} + (\xi^2 - \xi^3)^{-2} + (\xi^3 - \xi^1)^{-2}], \tag{3}$$

$$V_3 = 3f[(\xi^1 + \xi^2 - 2\xi^3)^{-2} + (\xi^2 + \xi^3 - 2\xi^1)^{-2} + (\xi^3 + \xi^1 - 2\xi^2)^{-2}]. \tag{4}$$

Here Ω is the angular frequency arising from the strength of the harmonic potentials and g and f are real parameters representing the strengths of the two-body and three-body potentials. We assume that $g > -\frac{1}{2}$ and $f > -\frac{1}{2}$ to avoid "fall to the center." The three-body interaction represents essentially a coupling between any two bodies i, j through the presence of the third body k. It depends symmetrically on the distance of i and j from k, vanishes when k is removed to infinity with respect to i and j, and reduces to the two-body inverse square potential whenever $k = i$ or $k = j$.

The propagator is now defined as the path integral

$$K(\xi'', \xi', t) = \lim_{N\to\infty} A_N \int \exp\left[i\sum_{j=1}^{N} S(\xi_j, \xi_{j-1})\right] \prod_{j=1}^{N-1} d\xi_j, \tag{5}$$

where ξ stands for the triplet (ξ^1, ξ^2, ξ^3), $\xi_j = \xi(t_j)$, $\xi_0 = \xi'$, $\xi_N = \xi''$, $t_j - t_{j-1} = \epsilon = t/N$, and A_N is the normalization factor in the Nth approximation. The action $S(\xi_j, \xi_{j-1})$ over a small time interval $t_j - t_{j-1} = \epsilon$ may be approximated by

$$\begin{aligned} S(\xi_j, \xi_{j-1}) &\simeq \epsilon L(\xi_j, \xi_{j-1}, \epsilon) = \frac{1}{2\epsilon}\sum_{i=1}^{3}(\xi_j^i - \xi_{j-1}^i)^2 - \frac{\epsilon\Omega^2}{6}[(\xi_j^1 - \xi_j^2)^2 \\ &+ (\xi_j^2 - \xi_j^3)^2 + (\xi_j^3 - \xi_j^1)^2] - \epsilon g\{[(\xi_j^1 - \xi_j^2)(\xi_{j-1}^1 - \xi_{j-1}^2)]^{-1} \\ &+ [(\xi_j^2 - \xi_j^3)(\xi_{j-1}^2 - \xi_{j-1}^3)]^{-1} + [(\xi_j^3 - \xi_j^1)(\xi_{j-1}^3 - \xi_{j-1}^1)]^{-1}\} \\ &- 3\epsilon f\{[(\xi_j^1 + \xi_j^2 - 2\xi_j^3)(\xi_{j-1}^1 + \xi_{j-1}^2 - 2\xi_{j-1}^3)]^{-1} \\ &+ [(\xi_j^2 + \xi_j^3 - 2\xi_j^1)(\xi_{j-1}^2 + \xi_{j-1}^3 - 2\xi_{j-1}^1)]^{-1} \\ &+ [(\xi_j^3 + \xi_j^1 - 2\xi_j^2)(\xi_{j-1}^3 + \xi_{j-1}^1 - 2\xi_{j-1}^2)]^{-1}\}. \end{aligned} \tag{6}$$

Next, we introduce the center of mass (CM) and Jacobi coordinates

$$\begin{aligned} &\xi^1 + \xi^2 + \xi^3 = 3R, \\ &\xi^1 - \xi^2 = \sqrt{2}x, \\ &\xi^1 + \xi^2 - 2\xi^3 = \sqrt{6}y, \end{aligned} \tag{7}$$

so that the action $S(\xi_j, \xi_{j-1})$ is transformed as

$$S(\xi_j, \xi_{j-1}) = S_0(R_j, R_{j-1}) + \bar{S}(x_j, y_j; x_{j-1}, y_{j-1}), \tag{8}$$

where

$$S_0 = \frac{3}{2\epsilon}(R_j - R_{j-1})^2, \tag{9}$$

Copyright © 1977 American Institute of Physics

$$\bar{S}(x_j, y_j; x_{j-1}, y_{j-1})$$
$$= \tfrac{1}{2}\epsilon\{(x_j - x_{j-1})^2 + (y_j - y_{j-1})^2\}$$
$$- \frac{\epsilon\Omega^2}{2}(x_j^2 + y_j^2) - \frac{g\epsilon}{2}\left\{\frac{1}{x_j x_{j-1}} + \frac{4}{(\sqrt{3}y_j - x_j)(\sqrt{3}y_{j-1} - x_{j-1})}\right.$$
$$\left. + \frac{4}{(\sqrt{3}y_j + x_j)(\sqrt{3}y_{j-1} + x_{j-1})}\right\}$$
$$- \frac{f\epsilon}{2}\left\{\frac{1}{y_j y_{j-1}} + \frac{4}{(y_j + \sqrt{3}x_j)(y_{j-1} + \sqrt{3}x_{j-1})}\right.$$
$$\left. + \frac{4}{(y_j - \sqrt{3}x_j)(y_{j-1} - \sqrt{3}x_{j-1})}\right\}. \tag{10}$$

The separation of the action S into S_0 and $\bar{S}$ allows us to write

$$K(\xi'', \xi', t) = K_0(R'', R', t)\bar{K}(x'', y''; x', y', t), \tag{11}$$

where K_0 is a normalized free particle propagator which corresponds to the motion of the center of mass. The motion relative to the CM is described by the propagator $\bar{K}$,

$$\bar{K}(x'', y''; x', y', t)$$
$$= \lim_{N\to\infty} B_N \int \exp\left[i\sum_{j=1}^{N} \bar{S}(x_j, y_j; x_{j-1}, y_{j-1})\right] \prod_{j=1}^{N-1} dx_j dy_j, \tag{12}$$

where B_N is the new normalization constant.

In order to bring out the natural symmetry involved in this problem, we make a further transformation to plane polar coordinates (r, θ),

$$r = (x^2 + y^2)^{1/2} \quad (0 \le r < \infty),$$
$$\theta = \tan^{-1} x/y \quad (0 \le \theta < 2\pi), \tag{13}$$
$$x = r\sin\theta, \quad y = r\cos\theta.$$

It may be noted that the transformation of Eq. (13) implies the following relations between the new and the old coordinates:

$$r = (1/\sqrt{3})[(\xi^1 - \xi^2)^2 + (\xi^2 - \xi^3)^2 + (\xi^3 - \xi^1)^2]^{1/2}, \tag{14a}$$
$$\theta = \tan^{-1}[\sqrt{3}(\xi^1 - \xi^2)/(\xi^1 + \xi^2 - 2\xi^3)], \tag{14b}$$
$$\sin 3\theta = -\sqrt{2}r^{-3} \cdot (\xi^1 - \xi^2)(\xi^2 - \xi^3)(\xi^3 - \xi^1), \tag{14c}$$
$$\cos 3\theta = \sqrt{2/27}\, r^{-3}(\xi^1 + \xi^2 - 2\xi^3)(\xi^2 + \xi^3 - 2\xi^1)(\xi^3 + \xi^1 - 2\xi^2) \tag{14d}$$

We might add here that Eqs. (14) imply that there is a certain ordering of the particles associated with every value of θ In fact, the two-body inverse cube forces do not permit interpenetration of particles so that a given order will be preserved under the presence of these forces. The three-body forces allow an interchange of two particles while only the third conserves order. Some interesting consequences of this ordering for identical and distinguishable particles have been already pointed out in Ref. 4. In the path integration to follow we shall restrict ourselves to only one of the intervals for θ permitted by the potentials used. This is no loss of generality because the propagators for other ranges of θ may be constructed in essentially the same manner as the wavefunctions of Ref. 4.

Using Eqs. (7), (13), and (14) the Lagrangian of Eq. (1) is transformed to the form

$$L = L_0 + \bar{L}, \tag{15a}$$

where

$$L_0 = (3/2)\dot{R}^2, \tag{15b}$$

$$\bar{L} = \tfrac{1}{2}(\dot{r}^2 + r^2\dot{\theta}^2) - \tfrac{1}{2}\Omega^2 r^2 - \frac{9g}{2r^2\sin^2 3\theta} - \frac{9f}{2r^2\cos^2 3\theta}. \tag{15c}$$

In fact our problem is to path integrate the Lagrangian decribed by Eqs. (15). Since L_0 represents the Lagrangian corresponding to CM motion, its path integration yields the free particle propagator mentioned above. The motion relative to CM is described by $\bar{L}$ and is our main concern here. Classically $\bar{L}$ describes the motion of a single particle in a plane under the action of a potential which contains a harmonic term and two others which depend on both r and θ. It is this dependence of potential on r and θ which makes this problem nontrivial from the path integrals point of view.

With these remarks we proceed to evaluate the propagator of Eq. (12) and write

$$\bar{K}(x'', y''; x', y', t) = K(r'', \theta''; r', \theta', t)$$
$$= \lim_{N\to\infty} B_N \int \exp\left(i\sum_{j=1}^{N} S_j\right) \prod_{j=1}^{N-1} r_j\, dr_j d\theta_j, \tag{16}$$

where the action S_j is obtained from $\bar{S}$ of Eq. (10), using the transformation equations (13) and (14), and is given by

$$S_j = \frac{1}{2\epsilon}[r_j^2 + r_{j-1}^2 - 2r_j r_{j-1}\cos(\theta_j - \theta_{j-1})]$$
$$- \frac{\epsilon}{2}\Omega^2 r_j^2 - \frac{9\epsilon g}{2r_j r_{j-1}\sin 3\theta_j \sin 3\theta_{j-1}}$$
$$- \frac{9\epsilon f}{2r_j r_{j-1}\cos 3\theta_j \cos 3\theta_{j-1}} + O(\epsilon^2). \tag{17}$$

In Eq. (17) we have indicated by $O(\epsilon^2)$ the terms which are of second order in ϵ which arise during the transformation of the inverse square and three-body potentials in Eq. (10). In the context of path integral theory the contributions of such second order terms to the finite time propagator is zero and we need only to compute the action correct to the first order in ϵ.

Next, we have to correctly express $\cos(\theta_j - \theta_{j-1})$ in terms of $\cos 3(\theta_j - \theta_{j-1})$. This is an important step in the subsequent evaluation of the propagator. Writing $\psi = \theta_j - \theta_{j-1}$ for short, and comparing the expansions of $\cos\psi$ and $\cos 3\psi$ in terms of their respective arguments, it is easy to obtain

$$\cos\psi \simeq (1/9)(8 + \cos 3\psi - 3\psi^4) + O(\psi^6). \tag{18}$$

As pointed out by Edwards and Gulyaev,[8] it is necessary to retain the fourth order terms in the angular changes ψ in the kinetic energy part of the action in order to ob-

tain the correct finite time propagator. Using Eq. (18) we may rewrite the action as

$$S_j=(1/2\epsilon)[r_j^2+r_{j-1}^2-(16/9)r_jr_{j-1}-\Omega^2r_j^2\epsilon^2 - \tfrac{2}{9}r_jr_{j-1}(\cos3\psi-3\psi^4)]-9g\epsilon/2r_jr_{j-1}\sin3\theta_j\sin3\theta_{j-1} - 9f\epsilon/2r_jr_{j-2}\cos3\theta_j\cos3\theta_{j-1}. \tag{19}$$

We now concentrate on the evaluation of $\exp(iS_j)$. Writing explicitly

$$\exp(iS_j)=\exp(iW_j)\exp\left(\frac{8u}{\epsilon}\right)\exp\left[\frac{u}{\epsilon}(\cos3\psi-3\psi^4)\right] \times\exp\left(\frac{-\epsilon(a^2-1/4)}{2u\sin3\theta_j\sin3\theta_{j-1}}-\frac{\epsilon(b^2-1/4)}{2u\cos3\theta_j\cos3\theta_{j-1}}\right), \tag{20}$$

where we have set

$$a=\tfrac{1}{2}(1+4g)^{1/2},\quad b=\tfrac{1}{2}(1+4f)^{1/2},\quad u=r_jr_{j-1}/9i, \tag{21}$$

and

$$W_j=(1/2\epsilon)(r_j^2+r_{j-1}^2-\Omega^2r_j^2\epsilon^2). \tag{22}$$

Further simplification of this expression is obtained when we use the result

$$\exp\frac{u}{\epsilon}(\cos3\psi-3\psi^4)\simeq\exp\left(\frac{u}{\epsilon}\cos3\psi-\frac{\epsilon}{9u}\right)+O(\epsilon^2) \tag{23}$$

which is proved in the Appendix. After some rearrangement of the terms we obtain

$$\exp(iS_j)=\exp(iW_j)\exp\left(\frac{8u}{\epsilon}-\frac{\epsilon}{9u}\right) \times\exp\left\{\frac{u}{\epsilon}\sin3\theta_j\sin3\theta_{j-1}-\frac{\epsilon(a^2-1/4)}{2u\sin3\theta_j\sin3\theta_{j-1}}\right\} \times\exp\left\{\frac{u}{\epsilon}\cos3\theta_j\cos3\theta_{j-1}-\frac{\epsilon(b^2-1/4)}{2u\cos3\theta_j\cos3\theta_{j-1}}\right\}. \tag{24}$$

Noting that the asymptotic form of $I_\nu(\omega/\epsilon)$, the modified Bessel function, for small ϵ, is given by

$$I_\nu(\omega/\epsilon)\simeq(\epsilon/2\pi\omega)^{1/2} \times\exp[\omega/\epsilon-\tfrac{1}{2}(\nu^2-\tfrac{1}{4})\epsilon/\omega+O(\epsilon^2)], \tag{25}$$

we may replace the last two exponentials in Eq. (24) by appropriate Bessel functions and arrive at the result

$$\exp(iS_j)=\exp(iW_j)\exp\left(\frac{8u}{\epsilon}-\frac{\epsilon}{9u}\right) \times\left(\frac{2\pi u}{\epsilon}\right)(\cos3\theta_j\cos3\theta_{j-1})^{1/2}(\sin3\theta_j\sin3\theta_{j-1})^{1/2} \times I_b\left(\frac{u}{\epsilon}\cos3\theta_j\cos3\theta_{j-1}\right)I_a\left(\frac{u}{\epsilon}\sin3\theta_j\sin3\theta_{j-1}\right). \tag{26}$$

This is the main result of this paper which will enable us to obtain the correct finite time propagator. At this point it is necessary to distinguish between three cases. The first case is when there are no three-body forces (i.e., $f=0$ or $b=\frac{1}{2}$). This has been discussed earlier in II and will not be considered here. The second case is when the two-body inverse cube forces are absent and we have only the harmonic pair forces and three-body forces. This will be discussed in Sec. 3. Finally, the case when all the three types of forces are present will be considered in Sec. 4.

3. PROPAGATOR FOR THE PROBLEM INVOLVING HARMONIC AND THREE-BODY FORCES

We now obtain an exact propagator for the problem involving harmonic and three-body forces. In this case we set $a=\frac{1}{2}$ in Eq. (24) and use Eq. (25) to replace the last exponential in Eq. (24) by an appropriate Bessel function. Thus, we have

$$\exp(iS_j)=\exp(iW_j)\exp\left(\frac{8u}{\epsilon}-\frac{\epsilon}{9u}\right) \times\left(\frac{2\pi u}{\epsilon}\right)^{1/2}(\cos3\theta_j\cos3\theta_{j-1})^{1/2} \times\exp\left(\frac{u}{\epsilon}\sin3\theta_j\sin3\theta_{j-1}\right) \times I_b\left(\frac{u}{\epsilon}\cos3\theta_j\cos3\theta_{j-1}\right). \tag{27}$$

We may now use the expansion formula[9]

$$(\cos\alpha\cos\beta)^{1/2-\lambda}I_{\lambda-1/2}(\omega\cos\alpha\cos\beta)\exp(\omega\sin\alpha\sin\beta) =2^{2\lambda}(2\pi\omega)^{-1/2}[\Gamma(\lambda)]^2\sum_{l=0}^{\infty}\frac{l!(\lambda+l)!}{\Gamma(2\lambda+l)}I_{l+\lambda}(\omega) \times C_l^\lambda(\sin\alpha)C_l^\lambda(\sin\beta) \tag{28}$$

where $C_l^\lambda(\sin\alpha)$ are the Gegenbauer polynomials, and write

$$\exp(iS_j)=\sum_{l_j=0}^{\infty}N_{l_j}^2(\cos3\theta_j\cos3\theta_{j-1})^{b+1/2} \times C_{l_j}^{b+1/2}(\sin3\theta_j)C_{l_j}^{b+1/2}(\sin3\theta_{j-1})R_{l_j}(r_j,r_{j-1}), \tag{29}$$

where N_{l_j} is the normalization factor of the Gegenbauer polynomial $C_{l_j}^{b+1/2}(\sin3\theta_j)$ over the range[10] $-\pi/6<\theta_j<\pi/6$ and

$$R_{l_j}(r_j,r_{j-1})=2\pi\exp(iW_j)I_{3(l_j+b+1/2)}\left(\frac{r_jr_{j-1}}{i\epsilon}\right). \tag{30}$$

In obtaining Eq. (29) we have used the following result proved in the Appendix:

$$\exp\left(\frac{8u}{\epsilon}-\frac{\epsilon}{9u}\right)I_\nu\left(\frac{u}{\epsilon}\right)\simeq3I_{3\nu}\left(\frac{9u}{\epsilon}\right)+O(\epsilon^2). \tag{31}$$

The result of Eq. (29) is now used in Eq. (16) to obtain the finite time propagator,

$$K(r'',\theta'';r',\theta',t) =\lim_{N\to\infty}B_N\sum_{l_1\cdots l_N}^{\infty}\int\prod_{j=1}^{N}N_{l_j}^2(\cos3\theta_j\cos3\theta_{j-1})^{b+1/2} \times C_{l_j}^{b+1/2}(\sin3\theta_j)C_{l_j}^{b+1/2}(\sin3\theta_{j-1}) \times R_{l_j}(r_j,r_{j-1})\prod_{j=1}^{N-1}r_j\,dr_jd\theta_j. \tag{32}$$

The angular integration in Eq. (32) can easily be performed by using the orthogonality relation

$$\int_{-\pi/6}^{\pi/6} C_l^{b+1/2}(\sin 3\theta) C_{l'}^{b+1/2}(\sin 3\theta)(\cos 3\theta)^{2b+1}\, d\theta = \delta_{ll'}/N_l^2 \tag{33}$$

and the integral in Eq. (32) reduces to

$$\prod_{j=1}^{N-1} \delta_{l_j, l_{j+1}} N_{l_N}^2 (\cos 3\theta'' \cos 3\theta')^{b+1/2} \times C_{l_j}^{b+1/2}(\sin 3\theta') C_{l_N}^{b+1/2}(\sin 3\theta'') . \tag{34}$$

When this expression is substituted in Eq. (32), we see that for each quantum number l, the angular and radial contributions to the propagator do separate. Thus

$$K(r'', \theta''; r', \theta', t) = \sum_{l=0}^{\infty} K_l(r'', r', t) N_l^2 \times (\cos 3\theta'' \cos 3\theta')^{b+1/2} \cdot C_l^{b+1/2}(\sin 3\theta'') C_l^{b+1/2}(\sin 3\theta') , \tag{35}$$

where

$$K_l(r'', r', t) = \lim_{N\to\infty} \left(\frac{1}{2\pi i\epsilon}\right)^N \int \prod_{j=1}^{N} R_{l_j}(r_j, r_{j-1}) \prod_{j=1}^{N-1} r_j\, dr_j \tag{36}$$

is the radial propagator of the l wave.[11] This propagator has been evaluated in closed form in I and we quote here the result,

$$K_l(r'', r', t) = \left(\frac{\Omega}{i \sin\Omega t}\right) \exp\left(\frac{i\Omega}{2}(r'^2 + r''^2)\cot\Omega t\right) \times I_{3(l+b+1/2)}\left(\frac{\Omega r' r''}{i \sin\Omega t}\right) . \tag{37}$$

We might add here that we have already recovered the angular part of the Schrödinger equation wavefunctions for the problem[4] in Eq. (35). It is also possible to obtain the radial wavefunctions from the expansion of $K_l(r'', r', t)$ in terms of Laguerre polynomials using the Hille–Hardy formula[12] in the same manner as in I.

4. PROPAGATOR FOR THE PROBLEM INVOLVING HARMONIC, INVERSE CUBE, AND THREE-BODY FORCES

In this case both g and f are nonzero and we have to consider the entire expression of Eq. (26). This may be cast in an appropriate form with the help of Weber's formula,[13]

$$\tfrac{1}{2} z I_\mu(z\cos\alpha\cos\beta) I_\nu(z\sin\alpha\sin\beta) = (\cos\alpha\cos\beta)^\mu (\sin\alpha\sin\beta)^\nu \times \sum_{l=0}^{\infty} (-1)^l (\mu+\nu+2l+1) I_{\mu+\nu+2l+1}(z) \times P_l^{\mu,\nu}(\cos 2\alpha) \cdot P_l^{\mu,\nu}(\cos 2\beta) , \tag{38}$$

where $P_l^{\mu,\nu}(\cos\phi)$ are the Jacobi polynomials, and subsequently absorbing the second exponential factor in Eq. (26) in the Bessel function using Eq. (31) as in Sec. 3. We thus obtain

$$\exp(iS_j) = \sum_{l_j=0}^{\infty} N_{l_j}^2 (-1)^{l_j} R_{l_j}(r_j, r_{j-1}) \times P_{l_j}^{a,b}(\cos 6\theta_j) P_{l_j}^{a,b}(\cos 6\theta_{j-1}) \times (\cos 3\theta_j \cos 3\theta_{j-1})^{b+1/2} (\sin 3\theta_j \sin 3\theta_{j-1})^{a+1/2} , \tag{39}$$

where N_{l_j} is the normalization factor of the Jacobi polynomials $P_{l_j}^{a,b}(\cos 6\theta_j)$ over the range[14] $0 < \theta_j < \pi/6$ and

$$R_{l_j}(r_j, r_{j-1}) = 2\pi \exp(iW_j) I_{3(2l_j+a+b+1)}(-ir_j r_{j-1}/\epsilon) . \tag{40}$$

After putting this expression in the finite time propagator of Eq. (16) and carrying out the intermediate angular integrations (over the range $0 < \theta < \pi/6$), we arrive at

$$K(r'', \theta''; r', \theta', t) = \sum_{l=0}^{\infty} K_l(r'', r', t) N_l^2 (\cos 3\theta'' \cos 3\theta')^{b+1/2} \times (\sin 3\theta'' \sin 3\theta')^{a+1/2} P_l^{a,b}(\cos 6\theta'') P_l^{a,b}(\cos 6\theta') , \tag{41}$$

where the radial propagator is given by

$$K_l(r'', r', t) = \lim_{N\to\infty} \left(\frac{1}{2\pi i\epsilon}\right)^N \int \prod_{j=1}^{N} R_l(r_j, r_{j-1}) \prod_{j=1}^{N-1} r_j\, dr_j . \tag{42}$$

This may again be evaluated as in I yielding the following closed form:

$$K_l(r'', r', t) = \left(\frac{\Omega}{i \sin\Omega t}\right) \exp\left\{\frac{i\Omega}{2}(r'^2 + r''^2)\cot\Omega t\right\} \times I_{3(2l+a+b+1)}(\Omega r'' r'/i \sin\Omega t) . \tag{43}$$

It is seen from this exact result that the angular part of the Schrödinger equation wavefunctions of Ref. 4 are already contained in the expansion (41). The radial part is obtainable from an expansion of $K_l(r'', r', t)$ in Laguerre polynomials as mentioned earlier.

5. CONCLUSIONS

The main results of this paper are the explicit path integration of a three-body problem and the exact analytical expressions for the propagator. The problem essentially involved the path integration of a Lagrangian which has the interesting feature of having a potential which is the sum of a harmonic term and a "centrifugal" term. The form of the centrifugal term arising from the two-body and the three-body forces given by $h(\theta)/r^2$ is what causes the separability of the propagator into radial and angular parts. Physically this corresponds to the conservation of the classical quantity $p_\theta^2 + h(\theta)/r^2$, where p_θ is the angular momentum. However, the explicit evaluation of the angular part of the propagator in terms of the known polynomials is possible because of the special nature of $h(\theta)$.[15] The exact closed form of the radial propagator becomes possible because it involves only the harmonic part of the potential.

As a final remark, we add that it is possible to obtain exact results even when the strength of the harmonic forces is time-dependent. This will only modify the radial propagators defined in Eqs. (36) and (42). The integrations involved in this case have already been indicated in another paper.[16]

APPENDIX

In this Appendix we derive the results stated in Eqs. (23) and (31).

A. Derivation of Eq. (23)

Consider the expansion

$$\psi^4 \exp\left(\frac{u}{\epsilon}\cos\psi\right) = \sum_{m=-\infty}^{\infty} a_m \exp(im\psi)\,. \tag{A1}$$

The expansion coefficients a_m are given by

$$\begin{aligned} a_m &= \int_0^{2\pi} \psi^4 \exp\left(\frac{u}{\epsilon}\cos\psi\right) e^{-im\psi}\,d\psi \\ &= \frac{\partial^4}{\partial m^4}\frac{1}{2\pi}\int_0^{2\pi} \exp\left(\frac{u}{\epsilon}\cos\psi\right) e^{-im\psi}\,d\psi \\ &= \frac{\partial^4}{\partial m^4} I_m\left(\frac{u}{\epsilon}\right), \end{aligned} \tag{A2}$$

where $I_m(\omega)$ is a modified Bessel function of order m. Explicit differentiation of I_m w.r.t. m yields

$$\frac{\partial^4}{\partial m^4} I_m\left(\frac{u}{\epsilon}\right) = 3\left(\frac{\epsilon}{u}\right)^2 I_m\left(\frac{u}{\epsilon}\right) + O(\epsilon^3)\,. \tag{A3}$$

From Eqs. (A2) and (A3), we obtain

$$\begin{aligned} \psi^4 \exp\left(\frac{u}{\epsilon}\cos\psi\right) &= \sum_{m=-\infty}^{\infty} 3\left(\frac{\epsilon}{u}\right)^2 I_m\left(\frac{u}{\epsilon}\right) e^{im\psi} + O(\epsilon^3) \\ &= 3\left(\frac{\epsilon}{u}\right)^2 e^{(u/\epsilon)\cos\psi} + O(\epsilon^3)\,, \end{aligned} \tag{A4}$$

where we have used the expansion

$$e^{(u/\epsilon)\cos\psi} = \sum_{m=-\infty}^{\infty} e^{im\psi} I_m(u/\epsilon)\,. \tag{A5}$$

Next, consider

$$\begin{aligned} e^{(u/\epsilon)(\cos\psi - c\psi^4)} &= [1 - (cu/\epsilon)\psi^4 + O(\psi^8)]\, e^{(u/\epsilon)\cos\psi} \\ &= [1 - 3c(u/\epsilon) + O(\epsilon^2)]\, e^{(u/\epsilon)\cos\psi} \\ &= e^{[(u/\epsilon)\cos\psi - 3c\epsilon/u]} + O(\epsilon^2)\,, \end{aligned} \tag{A6}$$

where in the second step we have used (A4). From Eq. (A6) it follows immediately by changing $\psi \to 3\psi$ and choosing $c = 1/27$, such that

$$e^{[(u/\epsilon)(\cos 3\psi - 3\psi^4)]} = e^{[(u/\epsilon)\cos 3\psi - \epsilon/9u]} + O(\epsilon^2) \tag{A7}$$

which is Eq. (23).

B. Derivation of Eq. (31)

Using the asymptotic expansion ($\epsilon \to 0$)

$$I_\nu\left(\frac{u}{\epsilon}\right) \simeq \left(\frac{\epsilon}{2\pi u}\right)^{1/2} \exp\left(\frac{u}{\epsilon} - \frac{\epsilon}{2u}(\nu^2 - \tfrac{1}{4}) + O(\epsilon^2)\right) \tag{A8}$$

we may write

$$\begin{aligned} &\exp\left(\frac{8u}{\epsilon} - \frac{\epsilon}{9u}\right) I_\nu\left(\frac{u}{\epsilon}\right) \\ &\quad \simeq 3\left(\frac{\epsilon}{18\pi u}\right)^{1/2} \exp\left[\frac{9u}{\epsilon} - \frac{\epsilon}{18u}\left((3\nu)^2 - \frac{1}{4}\right) + O(\epsilon^2)\right] \\ &\quad \simeq 3 I_{3\nu}\left(\frac{9u}{\epsilon}\right) + O(\epsilon^2) \end{aligned} \tag{A9}$$

which is the result of Eq. (31).

[1] F. Calogero, J. Math. Phys. **10**, 2191 (1969).
[2] D. C. Khandekar and S. V. Lawande, J. Phys. A: Gen. Phys. **5**, 812 (1972), hereafter referred to as I.
[3] D. C. Khandekar and S. V. Lawande, J. Phys. A: Gen. Phys. **5**, 157 (1972), hereafter referred to as II.
[4] J. Wolfes, J. Math. Phys. **15**, 1420 (1974).
[5] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965), Chapter 4.
[6] R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948).
[7] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics* (Pergamon, New York, 1958), Sec. 35.
[8] S. F. Edwards and Y. V. Gulyaev, Proc. R. Soc. London Ser. A **279** (1964).
[9] A. Erdelyi (Ed.), *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. 2, p. 194.
[10] As mentioned in Sec. 2, this range of θ corresponds to a particular ordering of the particles given by $\xi^1, \xi^2 > \xi^3$.
[11] Here the value of $B_N = (2\pi i\epsilon)^{-N}$ has been inserted.
[12] See Ref. 9, p. 189.
[13] See Ref. 9, p. 99.
[14] Here the range corresponds to the ordering $\xi^1 > \xi^2 > \xi^3$ and $\xi^2 < (\xi^1 + \xi^3)/2$.
[15] Note that $h(\theta) = 9g/\sin^2 3\theta + 9f/\cos^2 3\theta$.
[16] D. C. Khandekar and S. V. Lawande, J. Math. Phys. **16**, 384 (1975).

Generalized method of a resolvent operator expansion

M. Znojil

Institute of Nuclear Physics, Czechoslovakian Academy of Sciences, Řež, Czechoslovakia
(Received 9 October 1975)

The method of inversion of a quasitridiagonal matrix is suggested. The more general interpretation of the resulting expansion of the resolvent operator on a vector is given that does not use the concept of matrix elements nor inner product in a vector space. From this point of view, the linear operator $\hat{H}$ is employed recursively to define the basis starting with M initial vectors. The expansion of the action of the resolvent operator $\hat{R}(E)=(E-\hat{H})^{-1}$ on any basis vector is given in the same basis with simply calculated coefficients, thus generalizing the usual power series expansion.

1. INTRODUCTION

Matrices $\bar{H}$ that in some basis for some $p>0$ satisfy the condition

$$\langle n|\bar{H}|n+l+p\rangle=\langle n+l+p|\bar{H}|n\rangle=0, \quad l=1,2,\cdots, \tag{1}$$

are often useful in physics (e.g., kinetic energy operator or local potential operator of the form $\bar{U}(r)=\sum_{k=0}^{p}a_k r^{2k}$ in the harmonic oscillator basis.[1]) At the same time, $\bar{H}$ can appear in the form of the resolvent operator (e.g., propagator) $\bar{R}(E)=(E-\bar{H})^{-1}$. In the first part of the present paper (Sec. 2) we use a slight generalization of the condition (1) and suggest the method of calculation of the matrix elements of $\bar{R}(E)$.

In the second part of this paper (Sec. 3), we give another more general interpretation of the formulas. We take any linear operator $\hat{H}$ in a vector space V and define the basis in V by the repeated action of $\hat{H}$, starting from p initial vectors $|X_1^1\rangle,\ldots,|X_1^p\rangle$. In this basis we then expand the action of the resolvent operator $\hat{R}(E)=(E-\hat{H})^{-1}$ on any basis vector. The basis need not be orthonormal, but it will be complete in a subspace $\bar{V}\subset V$ invariant with respect to $\hat{H}$, $\hat{H}\bar{V}\subset\bar{V}$ (cf. Ref. 2, which we shall denote as I).

The results may be looked upon as a generalization of I, and therefore also as a generalization of the expansion of Haydock[3] or that of the power series expansion of $\hat{R}(E)$.

2. QUASITRIDIAGONAL MATRIX INVERSION

We call the matrix H tridiagonal, when it fulfills the condition (1) with $p=1$,

$$H_{k,k+l+1}=H_{k+l+1,k}=0, \quad k,l=1,2,\cdots.$$

Let us now have a N-dimensional tridiagonal matrix H, the elements of which are not numbers, but M-dimensional matrices $\mathsf{H}_{k,l}$, $k,l=1,2,\ldots,N$. We shall call the $(M\times N=N_D)$-dimensional matrix H quasitridiagonal. In the matrix notation, we have

$$\begin{aligned}&\mathsf{H}_{k,k+l+1}=\mathsf{H}_{k+l+1,k}=0,\\ &l=1,2,\ldots,N-k-1, \quad k=1,2,\ldots,N-2,\end{aligned} \tag{2}$$

and denote the transposed nonzero matrices as follows

$$\begin{aligned}&\mathsf{A}_k=\mathsf{H}_{k,k}^{T}, \quad k=1,2,\ldots,N,\\ &\mathsf{B}_k=\mathsf{H}_{k+1,k}^{T},\\ &\mathsf{C}_{k+1}=\mathsf{H}_{k,k+1}^{T}, \quad k=1,2,\ldots,N-1.\end{aligned} \tag{3}$$

We shall denote the M-dimensional vectors and matrices by the boldface and sans serif fonts, respectively, and write the indices of their elements on top; below we denote the different M-dimensional matrices while the number in the matrix H is identified by the pair of indices. We define the basis in the form of the column vectors X_k^r,

$$\begin{aligned}&(X_k^r)_i=\delta_{i,(k-1)M+r},\\ &i=1,2,\ldots,N_D,\ r=1,2,\ldots,M,\ k=1,2,\ldots,N,\end{aligned}$$

in the N_D-dimensional vector space V_{N_D}. The matrix H acts on the vector X_k^r in the space V_{N_D} according to the prescription

$$HX_k^r=\sum_{s=1}^{M}\left[C_k^{rs}X_{k-1}^s+A_k^{rs}X_k^s+B_k^{rs}X_{k+1}^s\right]. \tag{4}$$

We shall write (4) also in the abbreviated matrix notation as

$$H\mathbf{X}_k=\mathsf{C}_k\mathbf{X}_{k-1}+\mathsf{A}_k\mathbf{X}_k+\mathsf{B}_k\mathbf{X}_{k+1}.$$

Now, we intend to find the resolvent matrix $R(E)=(E-H)^{-1}$, where E is a variable not equal to any eigenvalue of H. Matrix elements of $R(E)$ can be defined in our basis as the projections of the vectors

$$Y_k^r=R(E)X_k^r. \tag{5}$$

If we use the expansion

$$Y_k^r=\sum_{l=1}^{N}\sum_{s=1}^{M}D_{kl}^{rs}X_l^s=\left(\sum_{l=1}^{N}\mathsf{D}_{kl}\mathbf{X}_l\right)^r, \tag{6}$$

then we have

$$\langle X_k^r|R(E)|X_l^s\rangle=D_{kl}^{rs} \tag{7}$$

due to the orthogonality of the basis vectors. Since

$$\begin{aligned}X_k^r&=(E-H)Y_k^r=E\sum_{ls}D_{kl}^{rs}X_l^s-\sum_{ls}D_{kl}^{rs}HX_l^s\\ &=\sum_{lst}\left[D_{kl}^{rt}(\delta_{st}E-A_l^{ts})-D_{kl+1}^{rt}C_{l+1}^{ts}-D_{kl-1}^{rt}B_{l-1}^{ts}\right]X_l^s,\end{aligned} \tag{8}$$

i.e.,

$$\mathbf{X}_k=\sum_{l}\left[\mathsf{D}_{kl}(E\mathsf{I}-\mathsf{A}_l)-\mathsf{D}_{kl+1}\mathsf{C}_{l+1}-\mathsf{D}_{kl-1}\mathsf{B}_{l-1}\right]\mathbf{X}_l,$$

we get in view of the independence of X_k^r the relation for D's,

$$\delta_{rs}\delta_{kl}=\sum_{t}\left[D_{kl}^{rt}(\delta_{st}E-A_l^{ts})-D_{kl+1}^{rt}C_{l+1}^{ts}-D_{kl-1}^{rt}B_{l-1}^{ts}\right], \tag{9}$$

where δ_{kl} is a Kronecker delta and $D_{k0}^{rs}=0$ is added for convenience. Since we have (4) and (6), we may expand

$$\begin{aligned}X_k^r&=\frac{1}{E-H}EX_k^r-\frac{1}{E-H}HX_k^r\\ &=EY_k^r-\sum_{s}\left[C_k^{rs}Y_{k-1}^s+A_k^{rs}Y_k^s+B_k^{rs}Y_{k+1}^s\right],\end{aligned} \tag{10}$$

Copyright © 1977 American Institute of Physics

i.e.,

$$\mathbf{X}_k = \sum_l [E\,\mathsf{D}_{kl} - \mathsf{A}_k\,\mathsf{D}_{kl} - \mathsf{B}_k\,\mathsf{D}_{k+1,l} - \mathsf{C}_k\,\mathsf{D}_{k-1,l}]\,\mathbf{X}_l$$

and get the second relation for D's,

$$\delta_{rs}\delta_{kl} = \sum_t [(\delta_{st}E - A_k^{rt})D_{kl}^{ts} - B_k^{rt}D_{k+1,l}^{ts} - C_k^{rt}D_{k-1,l}^{ts}], \tag{11}$$

which may be used in the recurrent determination of the values D_{kl}^{rs}. We put

$$D_{kl+1}^{rs} = \sum_t D_{kl}^{rt}\,\alpha_{l+1}^{ts}(k), \tag{12}$$

$$\alpha_l^{ts}(k) = \sum_{r=1}^{M} B_{l-1}^{tr}F_l^{rs}, \quad l \geq k > 0, \quad B_0^{ts} = \delta_{ts},$$

in complete analogy with I. The method of proving is that used in the paper I (mathematical induction), the only change being the use of the noncommuting matrices instead of the bare numbers. The insertion of (12) into (9) leads to the recurrent relation

$$\sum_s F_k^{rs}\left[\delta_{st}E - A_k^{st} - \sum_v \alpha_{k+1}^{sv}C_{k+1}^{vt}\right] = \delta_{rt} \tag{13}$$

i.e.,

$$\mathsf{F}_k = [E\mathsf{I} - \mathsf{A}_k - \boldsymbol{\alpha}_{k+1}\mathsf{C}_{k+1}]^{-1}$$

together with the boundary condition

$$\boldsymbol{\alpha}_{N+1} = 0 \tag{14}$$

which follows from the finite dimension of the matrix H. The evaluation of the sequence F_l, $l = 1, 2, \ldots, N$ is, according to (13), the matter of the M-dimensional matrix inversion, which is N times repeated. It follows from the boundary condition (14) that $\alpha_l^{rs}(k)$ does not depend on the index in the parentheses, $\alpha_l^{rs}(k) = \alpha_l^{rs}$, so that the sequence F_l is universal for all the vectors Y_k^r and is used repeatedly in formula (12). This is the main advantage of the calculation method and the reason for saving a number of operations.

The second advantage of the method manifests itself in approximate numerical calculations when the number N is only the cutoff parameter corresponding to the truncation of an infinite basis. Then the limit $N \to \infty$ should be taken in principle, while the number of the calculated $R(E)$ matrix elements may remain fixed. In practice, we always stop at some finite N. In our case, the amount of necessary work is proportional to N, while in the inversion of the general N_D-dimensional matrix $(E - H)$ the number of operations is proportional to N^3 (see Ref. 4). We are able to work with a higher N and therefore reach higher precision. At the end of this section, we add some other useful formulas. Since the ansatz

$$D_{k+1\,l}^{rs} = \sum_t \gamma_{k+1}^{rt}(l)D_{kl}^{ts} \tag{15}$$

leads to the relation

$$\gamma_{k+1}^{rt}(l) = \gamma_{k+1}^{rt} = \sum_s F_{k+1}^{rs}C_{k+1}^{st} \tag{16}$$

according to (11), we save further operations and conclude that only the calculation of the D_{kk}^{rs} matrices necessitates the use of the recurrence relations (9) or (11). The insertion of the relations (12) and (15) into (9) provides the relation

$$D_{kk}^{rs} = \sum_t \left[\delta_{rt} + \sum_{uvw} F_k^{ru}C_k^{uv}D_{k-1\,k-1}^{vw}B_{k-1}^{wt}\right]F_k^{ts} \tag{17}$$

i.e.,

$$\mathsf{D}_{kk} = [\mathsf{I} + \mathsf{F}_k\mathsf{C}_k\mathsf{D}_{k-1\,k-1}\mathsf{B}_{k-1}]\,\mathsf{F}_k.$$

Thus, the full solution of the problem of finding the matrices D_{kl}^{rs} in the expansion (6) is represented by the recurrence relations (13) together with the initialization (14) and the relations (17) and (12) and (15), (16).

3. THE CASE OF A GENERAL OPERATOR

Let us suppose that we have a linear operator $\hat{H}$ in a vector space V and we choose the set of M initial vectors $|X_1^r\rangle \in V$, $r = 1, 2, \ldots, M$. We then produce the sequence of the vectors $|X_k^r\rangle$, $r = 1, 2, \ldots, M$, $k = 2, 3, \cdots$ according to the definition

$$|X_{k+1}^r\rangle = \sum_s (\mathsf{B}_k^{-1})^{rs}\left\{\hat{H}|X_k^s\rangle - \sum_t [C_k^{st}|X_{k-1}^t\rangle + A_k^{st}|X_k^t\rangle]\right\}, \tag{18}$$

where the sequences of numbers

$$A_k^{rs}, B_k^{rs}, C_{k+1}^{rs}, \quad r, s = 1, 2, \ldots, M, \quad k = 1, 2, \cdots, \tag{19}$$

are chosen arbitrarily provided that the $M \times M$ matrices B_k can be inverted. The choice of the sequences (19) for fixed number M should make the structure of the new vectors (18) sufficiently simple for practical purposes.

The formalism of the previous section is applicable. The definition (18) is equivalent to the relation (4). For the operator $\hat{R}(E) = (E - \hat{H})^{-1}$ we may define the vectors (5), and the expansion (6) can then be written

$$|Y_k^r\rangle = \hat{R}(E)|X_k^r\rangle = \sum_{ls} D_{kl}^{rs}|X_l^s\rangle. \tag{20}$$

The new basis vectors need not be orthonormalized so that the relation (7) need not be valid, but the derivation of (9) and (11) may proceed in the same way. Therefore, the relations (12), (15), (16), and (17), together with (13) and (14), may be used for the calculation of the expansion coefficients D_{kl}^{rs} at some cutoff N. When the value of some $D_{k_0l_0}^{r_0s_0}$ becomes independent of the cutoff N, the exact result is obtained provided that the expansion (20) is convergent.

The general proof of the convergence of (20) is especially simple in the $M = 1$ case. It is thoroughly discussed in I. For $M > 1$, complications arise only in view of the matrix character of the expansion. E.g., orthogonality of $|X_k^r\rangle$, at least for different lower indices, can sometimes be introduced and exploited in a way analogous to I. The general formal discussion of the convergence properties is of less value here because of the great number of free parameters. It should be performed for every special case separately, along the same general lines as in I.

4. SUMMARY

The main result of this paper is the expansion (20) expressing the action of a resolvent operator $\hat{R}(E) = (E - \hat{H})^{-1}$ on a vector. This expansion is characterized by a set of free input parameters with twofold significance. First, they enable us to define (and simplify) the basis vectors in a subspace of the given vector space invariant with respect to the action of the operator $\hat{H}$. Second, they appear in calculation of the expansion coefficients.

There exist two extreme possibilities:

(1) We choose trivial input parameters $M = 1$, $B = 1$, $A = C = 0$. Then the expansion reduces to the usual power series expansion of the resolvent operator, i.e., the calculation of the expansion coefficients becomes trivial. The expansion vectors are not simplified at all.

(2) We choose some most elementary expansion vectors in advance. The input parameters must then be put equal to the matrix elements of the operator $\hat{H}$ the way it was done in Sec. 2. If there are no zero matrix elements, the formulas may be used only with $N = 2$. The calculation of the expansion coefficients is very involved in this case, being practically equivalent to the whole matrix inversion.

Therefore, the presented formalism constitutes an appropriate link between these two extremes. It is very flexible and enables us to find an optimal compromise between the simplicity of the expansion basis vectors and that of the expansion coefficients calculation. In practice, it seems to be most adequate for treating the operator $\hat{H}$ with the largest matrix elements lying near the main diagonal in some reasonable basis.

Note added in proof: The range of applicability will be enlarged when the dimensions $M = M_k$ of submatrices in the definition (3) are not equal to the same number and depend on the index k. The multiplication of (not always square) submatrices remains well defined for $M_1 \neq M_2 \neq \cdots$, and the whole formalism is valid even in this case. Further, the method may be adapted for the operator diagonalization (paper submitted to Can. J. Phys.).

[1]S. G. Nilsson, Kgl. Danske Vidensk Selsk., Mat.-Fys. Medd. **29**, 1 (1955).
[2]M. Znojil, J. Phys. A: Math. and Gen. **9**, 1 (1976).
[3]R. Haydock, J. Phys. A: Math., Nucl. and Gen. **7**, 2120 (1974).
[4]F. S. Acton, *Numerical Methods That Work* (Harper I. E., New York, 1970).

Charged particles in external fields. I. Classical theory

S. N. M. Ruijsenaars

Instituut-Lorentz, University of Leiden, Leiden, The Netherlands
(Received 15 July 1976)

A general perturbation theory is presented which can be used to study the interaction of relativistic charged particles with external fields. The theory is applied to the classical Dirac and Klein–Gordon equations. For external fields which are test functions on space–time retarded and advanced fundamental solutions are shown to exist, and their relation with the evolution operator is established. The S operator is shown to be Lorentz covariant and causal. The implementability in Fock space of the evolution is considered. It is shown that electric fields give rise to an implementable evolution while other kinds of fields in general do not.

1. INTRODUCTION

This is the first of two papers on the external field problem, i.e., the problem of the interaction of relativistic particles with electromagnetic or other fields which are prescribed functions on space–time. The present paper is mainly concerned with classical aspects. The results of this paper and of a paper on the Bogoliubov transformations which occur in the (second-)quantized theory of charged particles[1] are used in Ref. 2 to study the quantized Dirac and Klein–Gordon theories.

The paper has two parts. The first part (Sec. 2 and the Appendix) contains mathematical results of a perturbation-theoretic character. A number of these can be regarded as generalizations of recent results of Bellissard.[3,4] Assuming that certain distributions, which are connected with the perturbed Green's functions, upon restriction to the mass shell give rise to bounded operators on the classical Hilbert space, he obtains relations between these operators which imply the (pseudo-)unitarity of closely related operators. This can be regarded as a special case of a quite general perturbation-theoretic structure which is presented in Sec. 2.

In the second part (Secs. 3 and 4) these general results are applied to the classical Dirac and Klein–Gordon equations with external fields. For fields which are test functions on space–time two-sided tempered retarded and advanced fundamental solutions are shown to exist, and their relation with the evolution operator is established. Furthermore, the S operator is proved to be Lorentz covariant and causal. In the spin-$\frac{1}{2}$ case we show that the evolution is unitarily implementable in Fock space if only the timelike component of the vector field (i.e., the electric field) or the pseudovector field is nonzero. In the spin-0 case this holds true for electric or scalar fields. (Earlier results on implementability of the evolution can be found in Refs. 5–11.) It was recently proved by Hochstenbach[11] that in the spin-0 case a time-independent magnetic field gives rise to an evolution operator which is not implementable in Fock space. We show that in the spin-$\frac{1}{2}$ case the same Haag phenomenon occurs in general for the fourteen remaining kinds of fields, and rederive his result in a different way.

As a rule we do not state the weakest conditions under which our results apply. Thus, we usually assume that the fields are test functions although in particular the results on implementability hold true for more general functions. Similarly, the hypotheses of Sec. 2 could be relaxed and generalized. Here, however, we did try to accommodate the general features of theories of relativistic particles. In view of the resulting generality the reader is advised to skip this section on first reading and refer back to it when needed.

2. CLASSICAL PERTURBATION THEORY

In this section it is assumed that H is a separable Hilbert space which is the direct sum of two subspaces H_+ and H_- with corresponding projections P_+ and P_-. A bounded operator U on H is defined to be pseudo-unitary if

$$UqU^* = U^*qU = q\,, \tag{2.1}$$

where

$$q \equiv P_+ - P_-\,. \tag{2.2}$$

We define

$$U_{\epsilon\epsilon'} = P_\epsilon U P_{\epsilon'}, \quad \epsilon, \epsilon' = +, -\,, \tag{2.3}$$

and note that

$$U_{\epsilon\epsilon'}{}^* = U^*{}_{\epsilon'\epsilon}\,. \tag{2.4}$$

The set $R \cup \{-\infty, \infty\}$ is denoted by $\bar{R}$; continuity on $\bar{R}$ is defined in the obvious way. We also need $\bar{R}^2$, which by definition has the product topology. Let $O(t)$ be a function on R with values in the bounded operators on H. We assume that $O(t)$ is strongly continuous on R and that

$$\|O(\cdot)\| \in L^1(R)\,. \tag{2.5}$$

We define

$$K = \int dt\, \|O(t)\|, \quad l = K^{-1}. \tag{2.6}$$

We now introduce a number of operator-valued functions. The set of (λ, T_2, T_1) for which they are defined is delineated in Theorems 1 and 2 below. Their arguments will be suppressed whenever this does not give rise to confusion. In this section any operator integration or differentiation is in the strong sense.

$$\left.\begin{matrix} R_\lambda(T_2, T_1) \\ A_\lambda(T_2, T_1) \end{matrix}\right\} = \sum_{n=1}^{\infty} (i\lambda)^n \int_{T_1}^{T_2} dt_1 \cdots \int_{T_1}^{T_2} dt_n\, O(t_1) \times [\pm\hat{\theta}(\pm(t_1 - t_2))] O(t_2) \cdots [\pm\hat{\theta}(\pm(t_{n-1} - t_n))] O(t_n)\,, \tag{2.7}$$

 Copyright © 1977 American Institute of Physics

$$\left.\begin{matrix} F_\lambda(T_2,T_1) \\ \overline{F}_\lambda(T_2,T_1) \end{matrix}\right\} = \sum_{n=1}^{\infty}(i\lambda)^n \sum_{\epsilon_1,\dots,\epsilon_{n-1}} \int_{T_1}^{T_2} dt_1 \cdots \int_{T_1}^{T_2} dt_n\, O(t_1) \times [\pm\epsilon_1 P_{\epsilon_1}\hat{\theta}(\pm\epsilon_1(t_1-t_2))]\cdots \times[\pm\epsilon_{n-1}P_{\epsilon_{n-1}}\hat{\theta}(\pm\epsilon_{n-1}(t_{n-1}-t_n))]O(t_n)\,, \quad (2.8)$$

where

$$\hat{\theta}(t) \equiv \begin{cases} \theta(t), & T_2 > T_1\,, \\ \theta(-t), & T_2 < T_1\,, \end{cases} \quad (2.9)$$

$$U = 1+R, \quad Z = 1+qF\,, \qquad V = 1-A, \quad \overline{Z} = 1-q\overline{F}\,. \quad (2.10)$$

Finally,

$$R^{(0)}(T_2,T_1) = 1\,,$$

$$R^{(n)}(T_2,T_1) = i^n \int_{T_1}^{T_2} dt_1 \cdots \int_{T_1}^{t_{n-1}} dt_n\, O(t_1)\cdots O(t_n), \quad n \geq 1\,. \quad (2.11)$$

We have occasion to use the properties of these functions which are enumerated in the following theorems.

Theorem 2.1: For any $(T_2,T_1)\in \tilde{R}^2$, R, A, U, and V are $\|\cdot\|$-entire functions of λ. For any $\lambda\in C$ they are $\|\cdot\|$-continuous on $\tilde{R}^2$. On $C\times\tilde{R}^2$,

$$U_\lambda(T_2,T_1) = \sum_{n=0}^{\infty}\lambda^n R^{(n)}(T_2,T_1)\,, \quad (2.12)$$

$$\|U_\lambda(T_2,T_1)\| \leq \exp(|\lambda|K)\,. \quad (2.13)$$

Theorem 2.2: For any $(T_2,T_1)\in\tilde{R}^2$, F, $\overline{F}$, Z, and $\overline{Z}$ are $\|\cdot\|$-analytic functions of λ in the disc

$$D_l \equiv \{\lambda\in C \mid |\lambda| < l\}\,. \quad (2.14)$$

For any $\lambda\in D_l$ they are $\|\cdot\|$-continuous on $\tilde{R}^2$.

Theorem 2.3: On $C\times\tilde{R}^2$ [(2.15)], resp. $D_l\times\tilde{R}^2$ [(2.16)–(2.20)],

$$R-A-RA = R-A-AR = 0\,, \quad (2.15)$$

$$F-\overline{F}-Fq\overline{F} = F-\overline{F}-\overline{F}qF = 0\,, \quad (2.16)$$

$$R-F-RP_-F = R-F-FP_-R = 0\,, \quad (2.17)$$

$$R-\overline{F}-RP_+\overline{F} = R-\overline{F}-\overline{F}P_+R = 0\,, \quad (2.18)$$

$$A-F+AP_+F = A-F+FP_+A = 0\,, \quad (2.19)$$

$$A-\overline{F}+AP_-\overline{F} = A-\overline{F}+\overline{F}P_-A = 0\,. \quad (2.20)$$

Theorem 2.4 (the interaction picture evolution operator): On $C\times\tilde{R}^2$,

$$U(T,T) = 1\,, \qquad U(T_3,T_2)U(T_2,T_1) = U(T_3,T_1)\,. \quad (2.21)$$

For any $T_1(T_2)\in\tilde{R}^2$ and any $T_2(T_1)\in R$, $U(T_2,T_1)$ is differentiable w.r.t. $T_2(T_1)$. In these points

$$\partial_{T_2}U_\lambda(T_2,T_1) = i\lambda O(T_2)U_\lambda(T_2,T_1)\,, \quad (2.22)$$

$$\partial_{T_1}U_\lambda(T_2,T_1) = -i\lambda U_\lambda(T_2,T_1)O(T_1)\,. \quad (2.23)$$

If $\phi(T_2)\in H$ is strongly differentiable on R and

$$\varphi(T_1) = \phi, \quad T_1\in R\,, \qquad i\dot{\phi}(T_2) = -\lambda O(T_2)\phi(T_2), \quad \forall T_2\in R\,, \quad (2.24)$$

then

$$\phi(T_2) = U_\lambda(T_2,T_1)\phi\,. \quad (2.25)$$

Theorem 2.5 (the Schrödinger picture evolution operator): Let

$$O(t) = \exp(iH_0t)V(t)\exp(-iH_0t)\,, \quad (2.26)$$

where H_0 is a self-adjoint operator on H and $V(t)$ is continuously differentiable. Let

$$U^s_\lambda(T_2,T_1) = \exp(-iH_0T_2)U_\lambda(T_2,T_1)\exp(iH_0T_1)\,. \quad (2.27)$$

Then, on $C\times R^2$,

$$U^s(T,T) = 1\,, \qquad U^s(T_3,T_2)U^s(T_2,T_1) = U^s(T_3,T_1)\,, \quad (2.28)$$

$$U^sD(H_0) = D(H_0)\,. \quad (2.29)$$

For any $(T_2,T_1)\in R^2$ and $\phi\in D(H_0)$, $U^s(T_2,T_1)\phi$ is strongly differentiable w.r.t. T_2 and T_1, and

$$\partial_{T_2}U^s_\lambda(T_2,T_1)\varphi = -i[H_0-\lambda V(T_2)]U^s_\lambda(T_2,T_1)\phi\,, \quad (2.30)$$

$$\partial_{T_1}U^s_\lambda(T_2,T_1)\phi = iU^s_\lambda(T_2,T_1)[H_0-\lambda V(T_1)]\phi\,. \quad (2.31)$$

If $\varphi(T_2)\in D(H_0)$ is strongly differentiable on R and

$$\phi(T_1) = \phi\,, \quad T_1\in R\,, \qquad i\dot{\phi}(T_2) = [H_0-\lambda V(T_2)]\varphi(T_2), \quad \forall T_2\in R\,, \quad (2.32)$$

then

$$\varphi(T_2) = U^s_\lambda(T_2,T_1)\varphi\,. \quad (2.33)$$

Theorem 2.6 (rearrangement): Let $\tilde{O}(t)$ satisfy the same conditions as $O(t)$, and let

$$\tilde{U}_\lambda(T_2,T_1) = \sum_{n=0}^{\infty}\tilde{R}^{(n)}_\lambda(T_2,T_1)\,, \quad (2.34)$$

where

$$\tilde{R}^{(0)} \equiv 1\,, \quad (2.35)$$

$$\tilde{R}^{(1)}_\lambda(T_2,T_1) \equiv i\int_{T_1}^{T_2} dt[\lambda O(t)+\lambda^2\tilde{O}(t)]\,, \quad (2.36)$$

$$\tilde{R}^{(n)}_\lambda(T_2,T_1) \equiv i\int_{T_1}^{T_2} dt[\lambda O(t)+\lambda^2\tilde{O}(t)]\tilde{R}^{(n-1)}_\lambda(t,T_1) - i\int_{T_1}^{T_2} dt\,\lambda^2\tilde{O}(t)\tilde{R}^{(n-2)}_\lambda(t,T_1), \quad n\geq 2\,. \quad (2.37)$$

Then, on $C\times\tilde{R}^2$,

$$\tilde{U}_\lambda(T_2,T_1) = U_\lambda(T_2,T_1)\,. \quad (2.38)$$

Theorem 2.7: For any $(\lambda,T_2,T_1)\in C\times\tilde{R}^2$ (resp. $D_l\times\tilde{R}^2$), U and V (resp. Z and $\overline{Z}$) are invertible, and

$$U^{-1} = V\,, \quad (2.39)$$

$$Z^{-1} = \overline{Z}\,. \quad (2.40)$$

On $C\times\tilde{R}^2$,

$$U_{--}V_{--} = 1_{--}-U_{-+}V_{+-}\,, \qquad V_{--}U_{--} = 1_{--}-V_{-+}U_{+-}\,. \quad (2.41)$$

For any $(\lambda,T_2,T_1)\in D_l\times\tilde{R}^2$, U_{++}, V_{++} and U_{--}, V_{--} are invertible as operators on H_+, resp. H_-, and

$$F_{--}=1_{--}-U_{--}^{-1},\quad F_{+-}=U_{+-}U_{--}^{-1},$$
$$F_{-+}=U_{--}^{-1}U_{-+},\quad F_{++}=U_{++}-1_{++}-U_{+-}U_{--}^{-1}U_{-+}. \tag{2.42}$$

Theorem 2.8: Let $R^{(1)}(T_2,T_1)_{\pm\mp}$ be continuous on $\tilde{R}^2$ in the Hilbert–Schmidt (H.S.) norm and let

$$\|R^{(1)}(T_2,T_1)_{\pm\mp}\|_2 \leq C_1 < \infty \tag{2.43}$$

on $\tilde{R}^2$. Then $U_{\pm\mp}$ are $\|\cdot\|_2$-entire functions of λ and $\|\cdot\|_2$-continuous on $\tilde{R}^2$. Moreover, the set $E(T_2,T_1)\subset C$ where $U_\lambda(T_2,T_1)_{--}$ or $V_\lambda(T_2,T_1)_{--}$ are singular lies outside D_1 and has no limit points in C. $F_\lambda(T_2,T_1)$ has an $\|\cdot\|$-analytic continuation to $C\backslash E(T_2,T_1)$. The operators $F_{\pm\mp}$ are $\|\cdot\|_2$-analytic functions of λ on $C\backslash E$ and they are $\|\cdot\|_2$-continuous on $\tilde{R}^2$ for any $\lambda\in D_1$.

Theorem 2.9: Let $\lambda_0\in C\backslash\{0\}$ and let

$$\|U_{\lambda_0}(T_2,T_1)_{+-}\|_2 \leq C < \infty \tag{2.44}$$

for any $(T_2,T_1)\in Q$, where

$$Q\equiv[a,b]\times[a,b],\quad -\infty\leq a<b\leq\infty. \tag{2.45}$$

Then $R^{(1)}(T_2,T_1)_{+-}$ is H.S. on Q. Furthermore, U_{+-} is an $\|\cdot\|_2$-entire function of λ for any $(T_2,T_1)\in Q$.

If, in addition, $U_{\lambda_0}(T_2,T_1)_{+-}$ is $\|\cdot\|_2$-continuous on Q, then U_{+-} is $\|\cdot\|_2$-continuous on Q for any $\lambda\in C$.

The same statements hold true if $+-$ is replaced by $-+$.

Theorem 2.10 (bosons): Let

$$O(t)^*=qO(t)q,\quad \forall t\in R. \tag{2.46}$$

Then

$$R^*=-qAq,\quad \forall\lambda\in R, \tag{2.47}$$

$$F^*=-q\bar{F}q,\quad \forall\lambda\in(-l,l). \tag{2.48}$$

For any $\lambda\in R$, U and V are pseudo-unitary and, for any $\lambda\in(-l,l)$, Z and $\bar{Z}$ are unitary. Setting

$$\Lambda\equiv qF, \tag{2.49}$$

$$U'\equiv Z, \tag{2.50}$$

$$\Lambda'\equiv qR, \tag{2.51}$$

one has for any $\lambda\in(-l,l)$

$$(U-1)-q\Lambda+(U-1)P_-\Lambda=(U-1)-q\Lambda-q\Lambda P_-(U-1)=0, \tag{2.52}$$

$$(U'-1)-\Lambda'-(U'-1)P_-\Lambda'=(U'-1)-\Lambda'-\Lambda'P_-(U'-1)=0. \tag{2.53}$$

Theorem 2.11 (fermions): Let

$$O(t)^*=O(t),\quad \forall t\in R. \tag{2.54}$$

Then

$$R^*=-A,\quad \forall\lambda\in R, \tag{2.55}$$

$$F^*=-\bar{F},\quad \forall\lambda\in(-l,l). \tag{2.56}$$

For any $\lambda\in R$, U and V are unitary and, for any $\lambda\in(-l,l)$, Z and $\bar{Z}$ are pseudo-unitary. Setting

$$\Lambda\equiv F, \tag{2.57}$$

$$U'\equiv Z, \tag{2.58}$$

$$\Lambda'\equiv R, \tag{2.59}$$

one has for any $\lambda\in(-l,l)$

$$(U-1)-\Lambda-(U-1)P_-\Lambda=(U-1)-\Lambda-\Lambda P_-(U-1)=0, \tag{2.60}$$

$$(U'-1)-q\Lambda'+(U'-1)P_-\Lambda'=(U'-1)-q\Lambda'-q\Lambda'P_-(U'-1)=0. \tag{2.61}$$

Proofs of Theorems 1–11: 1: The relation (2.12) is obvious. For any $(T_2,T_1)\in\tilde{R}^2$ and $\phi\in H$

$$\|R^{(n)}(T_2,T_1)\varphi\| \leq \int_{-\infty}^{\infty}dt_1\cdots\int_{-\infty}^{t_{n-1}}dt_n\,\|O(t_1)\|\cdots\|O(t_n)\|\,\|\varphi\|$$
$$\leq (K^n/n!)\|\varphi\| \tag{2.62}$$

so that

$$\|R^{(n)}(T_2,T_1)\| \leq K^n/n!. \tag{2.63}$$

Thus, (2.13) holds. Since U is the sum of a series of entire functions which converges uniformly on any bounded subset of C, U and R are entire functions of λ for any $(T_2,T_1)\in\tilde{R}^2$. To prove the second statement, we observe that for any $(T_2,T_1)\in R^2$ and $\phi\in H$

$$\|R^{(n)}(T_2+\delta_2,T_1+\delta_1)\varphi-R^{(n)}(T_2,T_1)\varphi\|$$
$$\leq\Big\|\Big[\Big(\int_{T_1+\delta_1}^{T_1}+\int_{T_1}^{T_2}+\int_{T_2}^{T_2+\delta_2}\Big)dt_1$$
$$\times\cdots\Big(\int_{T_1+\delta_1}^{T_1}+\int_{T_1}^{T_2}+\int_{T_2}^{T_2+\delta_2}\Big)dt_n$$
$$-\int_{T_1}^{T_2}dt_1\cdots\int_{T_1}^{T_2}dt_n\Big]\cdots\varphi\Big\|. \tag{2.64}$$

Multiplying out the first term and subtracting the second one, one obtains a finite sum of terms each of which contains at least one δ_1 or δ_2. Pulling the norm through the integral in all terms and using (2.5), one concludes that $R^{(n)}$ is norm continuous on R^2. Replacing either $T_2+\delta_2$ or $T_1+\delta_1$ or both by ∞ or $-\infty$ one similarly infers continuity on $\tilde{R}^2$. Thus, in view of the uniform convergence on $\tilde{R}^2$ of their perturbation series, U and R are norm continuous on $\tilde{R}^2$. The proofs for V and A are analogous. ■

2: Two different choices for $\epsilon_1,\ldots,\epsilon_{n-1}$ in (2.8) correspond to different integration regions in R^n. The first statement should therefore be clear from (2.6) and the proof that U is entire. The second one follows in the same way as for U. ■

3: One easily sees that (2.18)–(2.20) follow from (2.15)–(2.17). Indeed, multiplying (2.17) with $q\bar{F}$ from the right, resp. $\bar{F}q$ from the left, and using (2.16) and (2.17), one, e.g., obtains (2.18). It remains to prove (2.15)–(2.17). Since R and A are analytic, (2.15) holds if

$$\theta(t_1-t_2)\cdots\theta(t_{N-1}-t_N)+(-)^N\theta(t_2-t_1)\cdots\theta(t_N-t_{N-1})$$
$$=-\sum_{n=1}^{N-1}(-)^{N-n}\theta(t_1-t_2)\cdots\theta(t_{n-1}-t_n)\theta(t_{n+2}-t_{n+1})$$
$$\times\cdots\theta(t_N-t_{N-1})$$
$$=-\sum_{n=1}^{N-1}(-)^n\theta(t_2-t_1)\cdots\theta(t_n-t_{n-1})\theta(t_{n+1}-t_{n+2})$$
$$\times\cdots\theta(t_{N-1}-t_N). \tag{2.65}$$

The verification of (2.65) is straightforward. Choosing $\epsilon_1,\ldots,\epsilon_{N-1}$, we note that (2.65) still holds if t_i-t_j is replaced by $\epsilon_{\min(i,j)}(t_i-t_j)$, which proves (2.16). Finally, (2.17) follows from the easily verified relations

$$\theta(t_1-t_2)\cdots\theta(t_{N-1}-t_N)O(t_1)\cdots O(t_N)$$
$$-\sum_{\epsilon_1,\ldots,\epsilon_{N-1}}\theta(\epsilon_1(t_1-t_2))\cdots\theta(\epsilon_{N-1}(t_{N-1}-t_N))O(t_1)$$
$$\times\epsilon_1P_{\epsilon_1}\cdots\epsilon_{N-1}P_{\epsilon_{N-1}}O(t_N)$$
$$=\sum_{\epsilon_1,\ldots,\epsilon_{N-1}}\sum_{n=1}^{N-1}\theta(t_1-t_2)\cdots\theta(t_{n-1}-t_n)$$
$$\times\theta(\epsilon_{n+1}(t_{n+1}-t_{n+2}))\cdots\theta(\epsilon_{N-1}(t_{N-1}-t_N))O(t_1)$$
$$\times P_{\epsilon_1}\cdots P_{\epsilon_{n-1}}O(t_n)\tfrac{1}{2}(1-\epsilon_n)P_{\epsilon_n}O(t_{n+1})\epsilon_{n+1}P_{\epsilon_{n+1}}\cdots$$
$$\times\epsilon_{N-1}P_{\epsilon_{N-1}}O(t_N)$$
$$=\sum_{\epsilon_1,\ldots,\epsilon_{N-1}}\sum_{n=1}^{N-1}\theta(\epsilon_1(t_1-t_2))\cdots\theta(\epsilon_{n-1}(t_{n+1}-t_n))$$
$$\times\theta(t_{n+1}-t_{n+2})\cdots\theta(t_{N-1}-t_N)O(t_1)\epsilon_1P_{\epsilon_1}\cdots\epsilon_{n-1}P_{\epsilon_{n-1}}$$
$$\times O(t_n)\tfrac{1}{2}(1-\epsilon_n)P_{\epsilon_n}O(t_{n+1})P_{\epsilon_{n+1}}\cdots P_{\epsilon_{N-1}}O(t_N).\blacksquare \tag{2.66}$$

4: To prove (2.21), it suffices to show that

$$\sum_{n=0}^{N}R^{(n)}(T_3,T_2)R^{(N-n)}(T_2,T_1)=R^{(N)}(T_3,T_1). \tag{2.67}$$

But (2.67) is equivalent to

$$\sum_{n=0}^{N}\int_{T_2}^{T_3}dt_1\cdots\int_{T_2}^{t_{n-1}}dt_n\int_{T_1}^{T_2}dt_{n+1}$$
$$\times\cdots\int_{T_1}^{t_{N-1}}dt_N\,O(t_1)\cdots O(t_N)$$
$$=\int_{T_1}^{T_3}dt_1\cdots\int_{T_1}^{t_{N-1}}dt_N\,O(t_1)\cdots O(t_N), \tag{2.68}$$

which is easily verified. To prove (2.22), we note that

$$U_\lambda(T_2,T_1)=\sum_{n=0}^{\infty}(i\lambda)^n\int_{T_1}^{T_2}dt_1\cdots\int_{T_1}^{t_{n-1}}dt_n\,O(t_1)\cdots O(t_n). \tag{2.69}$$

By the uniform boundedness principle each term is differentiable w.r.t. T_2. Since the resulting series strongly converges to $i\lambda O(T_2)U_\lambda(T_2,T_1)$, uniformly in T_2 on any bounded neighborhood of any $T_2\in R$, (2.22) follows. Transforming (2.69) into

$$U_\lambda(T_2,T_1)=\sum_{n=0}^{\infty}(i\lambda)^n\int_{T_1}^{T_2}dt_1\cdots\int_{t_{n-1}}^{T_2}dt_n\,O(t_n)\cdots O(t_1), \tag{2.70}$$

one obtains (2.23) by a similar argument. To prove the uniqueness of the solution of (2.24), we define $\psi(T_2)$ $=\phi(T_2)-U_\lambda(T_2,T_1)\varphi$. Then $\psi(T_1)=0$ and $i\dot\psi(T_2)$ $=-\lambda\,O(T_2)\psi(T_2)$, and so

$$\psi(T_2)=i\lambda\int_{T_1}^{T_2}dtO(t)\psi(t)$$
$$=(i\lambda)^n\int_{T_1}^{T_2}dt_1\cdots\int_{T_1}^{t_{n-1}}dt_n\,O(t_1)\cdots O(t_n)\psi(t_n). \tag{2.71}$$

Estimating in the obvious way it follows that $\psi(T_2)=0$.■

5: By virtue of Theorem 4 and Stone's theorem it suffices to prove that for any $(\lambda,T_2,T_1)\in C\times R^2$

$$U_\lambda(T_2,T_1)D(H_0)\subset D(H_0). \tag{2.72}$$

Defining

$$R_n^s(T_2,T_1)=\exp(-iH_0T_2)\,R^{(n)}(T_2,T_1)\exp(iH_0T_1), \tag{2.73}$$

we observe that

$$R_{n+1}^s(T_2,T_1)=i\int_{T_1}^{T_2}dt\exp[-iH_0(T_2-t)]V(t)R_n^s(t,T_1), \tag{2.74}$$

which can also be written as

$$R_{n+1}^s(T_2,T_1)$$
$$=i\int_{T_1}^{T_2}dt$$
$$\times\exp[-iH_0(t-T_1)]V(T_2+T_1-t)R_n^s(T_2+T_1-t,T_1). \tag{2.75}$$

On $D(H_0)$, $V(T_2+T_1-t)R_0^s(T_2+T_1-t,T_1)$ is continuously differentiable w.r.t. T_2, and so from (2.75) it follows, using dominated convergence and an induction argument, that, on $D(H_0)$, $R_n^s(T_2,T_1)$ is continuously differentiable w.r.t. T_2 for any $n\in N$. It then follows from (2.74) by Stone's theorem that

$$R_n^s(T_2,T_1)D(H_0)\subset D(H_0)\quad\forall n\in N. \tag{2.76}$$

Differentiating (2.74) and (2.75), one obtains after some straightforward manipulations

$$H_0R^{(n+1)}(T_2,T_1)$$
$$=R^{(n+1)}(T_2,T_1)H_0+O(T_2)R^{(n)}(T_2,T_1)-R^{(n)}(T_2,T_1)O(T_1)$$
$$-\sum_{k=0}^{n}i^k\int_{T_1}^{T_2}dt_0\cdots\int_{T_1}^{t_{k-1}}dt_k\,O(t_0)\cdots O(t_{k-1})$$
$$\times\exp(iH_0t_k)\dot V(t_k)\exp(-iH_0t_k)R^{(n-k)}(t_k,T_1), \tag{2.77}$$

which holds on $D(H_0)$. Using (2.63), we get for any $\varphi\in D(H_0)$

$$\|H_0R^{(n+1)}(T_2,T_1)\varphi\|$$
$$\leq[K^{n+1}/(n+1)!]\|H_0\phi\|+\{[\|V(T_2)\|+\|V(T_1)\|]K^n/n!$$
$$+L(2K)^n/n!\}\|\phi\|, \tag{2.78}$$

where

$$L\equiv|T_2-T_1|\sup_{t\in[\min(T_1,T_2),\max(T_1,T_2)]}\|V(t)\|. \tag{2.79}$$

It evidently follows from (2.78) that the sequence $H_0\lambda^nR^{(n)}(T_2,T_1)\phi$ is absolutely summable. Thus, as H_0 is closed, (2.72) holds true.■

6: If $n \geq 2$ and, e.g., $T_2 > T_1$,

$$\|\tilde{R}_\lambda^{(n)}(T_2, T_1)\| \leq \int_{T_1}^{T_2} dt\, F(t, |\lambda|)[\|\tilde{R}_\lambda^{(n-1)}(t, T_1)\| + \|\tilde{R}_\lambda^{(n-2)}(t, T_1)\|], \tag{2.80}$$

where

$$F(t, |\lambda|) \equiv |\lambda|\,\|O(t)\| + |\lambda|^2\|\tilde{O}(t)\|. \tag{2.81}$$

Furthermore,

$$\|R_\lambda^{(1)}(T_2, T_1)\| \leq \int_{T_1}^{T_2} dt\, F(t, |\lambda|). \tag{2.82}$$

Iterating (2.80) and then using (2.35) and (2.82), one obtains less than 2^n terms. For the generic term we write

$$M(m, T_2, T_1, |\lambda|) \equiv \int_{T_1}^{T_2} dt_1 \cdots \int_{T_1}^{t_{m-1}} dt_m\, F(t_1, |\lambda|)\cdots F(t_m, |\lambda|), \tag{2.83}$$

where

$$n/2 \leq m \leq n. \tag{2.84}$$

Clearly,

$$M(m, T_2, T_1, |\lambda|) \leq C(|\lambda|)^m/m!, \tag{2.85}$$

where

$$C(|\lambda|) \equiv |\lambda|K + |\lambda|^2\tilde{K}. \tag{2.86}$$

Thus,

$$\|\tilde{R}_\lambda^{(n)}(T_2, T_1)\| \leq [2C'(|\lambda|)]^n/[n/2]!, \tag{2.87}$$

where

$$C'(|\lambda|) \equiv \max(1, C(|\lambda|)) \tag{2.88}$$

and where $\lfloor n/2 \rfloor$ denotes the greatest integer less than or equal to $n/2$. From (2.87) it follows as before that $\tilde{U}_\lambda(T_2, T_1)$ is $\|\cdot\|$-entire in λ and $\|\cdot\|$-continuous on $\tilde{R}^2$. Moreover,

$$\begin{aligned}\tilde{U}_\lambda(T_2, T_1) &= 1 + \tilde{R}_\lambda^{(1)}(T_2, T_1) + \sum_{n=2}^{\infty} i\int_{T_1}^{T_2} dt\,[\lambda O(t) + \lambda^2\tilde{O}(t)]\tilde{R}_\lambda^{(n-1)}(t, T_1) \\ &\quad - \sum_{n=1}^{\infty} i\int_{T_1}^{T_2} dt\,\lambda^2\tilde{O}(t)\tilde{R}_\lambda^{(n-1)}(t, T_1) \\ &= 1 + i\lambda\int_{T_1}^{T_2} dt\, O(t)\tilde{U}_\lambda(t, T_1).\end{aligned} \tag{2.89}$$

By virtue of (2.22), $U_\lambda(T_2, T_1)$ satisfies the same Volterra type integral equation which is easily seen to have a unique norm continuous solution. Thus, (2.38) follows. ■

7: This theorem is an immediate consequence of (2.15)–(2.18). ■

8: We need the following lemma.

Lemma 2.1: Let $X(t)$ be a function on R with values in the H.S. operators on $\mathcal{H}$, which is strongly continuous on R and such that

$$\|X(t)\|_2 \leq \alpha(t), \quad \alpha(\cdot) \in L^1(R). \tag{2.90}$$

If $Y \equiv \int dt X(t)$, then

$$\|Y\|_2 \leq \int dt \|X(t)\|_2 < \infty. \tag{2.91}$$

Proof: Y is well defined since $\|X(t)\phi\| \leq \alpha(t)\|\phi\|$. If $\{e_n\}_{n=1}^{\infty}$ is an orthonormal basis of $\mathcal{H}$, then $\|X(t)\|_2^2 = \sum_{n=1}^{\infty}(X(t)e_n, X(t)e_n)$ and so $\|X(t)\|_2$ is measurable and, by (2.90), integrable on R. Moreover, $\sum_{n=1}^{N}(Ye_n, Ye_n) \leq \int dt_1 dt_2 \sum_{n=1}^{N}\|X(t_1)e_n\|\,\|X(t_2)e_n\| < [\int dt\,\|X(t)\|_2]^2$, so that (2.91) holds true. ■

We have ($n \geq 1$)

$$\begin{aligned}-iR^{(n+1)}(T_2, T_1)_{+-} &= \int_{T_1}^{T_2} dt\,\{O(t)_{++}R^{(n)}(t, T_1)_{+-} \\ &\quad - i[\partial_t R^{(1)}(t, T_1)_{+-}]R^{(n)}(t, T_1)_{--}\} \\ &= \int_{T_1}^{T_2} dt\,[O(t)_{++}R^{(n)}(t, T_1)_{+-} \\ &\quad - R^{(1)}(t, T_1)_{+-}(O(t)R^{(n-1)}(t, T_1))_{--}] \\ &\quad - iR^{(1)}(T_2, T_1)_{+-}R^{(n)}(T_2, T_1)_{--}.\end{aligned} \tag{2.92}$$

An induction argument using (2.92), (2.43) and the lemma now shows that

$$\|R^{(n)}(T_2, T_1)_{+-}\|_2 \leq C_n < \infty \tag{2.93}$$

on $\tilde{R}^2$. In fact we can choose

$$C_n = C_1 K^{n-1}[2^{n-1}/(n-1)! + 2^{n-2}/(n-2)!], \quad n > 1, \tag{2.94}$$

since by (2.92) and (2.63), if, e.g., $T_2 > T_1$,

$$\begin{aligned}&\|R^{(n+1)}(T_2, T_1)_{+-}\|_2 \\ &\leq C_1 K^n\left(\frac{1}{n!} + \frac{1}{(n-1)!}\right) + \int_{T_1}^{T_2} dt_1\,\|O(t_1)\|\,\|R^{(n)}(t_1, T_1)_{+-}\|_2 \\ &\leq C_1 K^n\left(\frac{1}{n!} + \frac{1}{(n-1)!}\right) + \int_{T_1}^{T_2} dt_1\,\|O(t_1)\| \\ &\quad \times\left[C_1 K^{n-1}\left(\frac{1}{(n-1)!} + \frac{1}{(n-2)!}\right) + \int_{T_1}^{t_1} dt_2\,\|O(t_2)\|\right. \\ &\quad \times\left[C_1 K^{n-2}\left(\frac{1}{(n-2)!} + \frac{1}{(n-3)!}\right) + \cdots + \int_{T_1}^{t_{n-1}} dt_n\,\|O(t_n)\|\right. \\ &\quad \left.\left. \times\,\|R^{(1)}(t_n, T_1)\|_2\right]\cdots\right] \\ &\leq C_1 K^n\left(\frac{1}{0!n!} + \frac{1}{0!(n-1)!} + \cdots + \frac{1}{(n-1)!1!} + \frac{1}{(n-1)!0!}\right. \\ &\quad \left. + \frac{1}{n!0!}\right) \\ &= C_1 K^n\left[\frac{2^n}{n!} + \frac{2^{n-1}}{(n-1)!}\right].\end{aligned} \tag{2.95}$$

Thus, U_{+-} is an $\|\cdot\|_2$-entire function of λ for any $(T_2, T_1) \in \tilde{R}^2$. Since $\tilde{R}^2$ is a compact topological space the topology of which is generated by the metric

$$d((x_1, x_2), (y_1, y_2)) \equiv \sum_{i=1}^{2} |\arctan x_i - \arctan y_i|, \tag{2.96}$$

we can conclude from (2.92), using induction and uniform continuity w.r.t. d, that $R^{(n)}{}_{+-}$ is $\|\cdot\|_2$-continuous on $\tilde{R}^2$. Hence, in view of (2.93)–(2.94), U_{+-} is $\|\cdot\|_2$-con-

tinuous on $\tilde{R}^2$. The statements regarding U_{-+} follow on replacing $+,-$ by $-,+$ in (2.92)–(2.95). The remaining statements follow from (2.41) and the analytic Fredholm theorem[12] and from (2.42).■

9: Since

$$U_{\lambda_0}(T_2, T_1)_{+-}$$
$$= \int_{T_1}^{T_2} dt\, \partial_t U_{\lambda_0}(t, T_1)_{+-} = i\lambda_0 \int_{T_1}^{T_2} dt\, O(t)_{++} U_{\lambda_0}(t, T_1)_{+-}$$
$$-\lambda_0 \int_{T_1}^{T_2} dt\, (\partial_t R^{(1)}(T_2, t)_{+-}) U_{\lambda_0}(t, T_1)_{--}$$
$$= i\lambda_0 \int_{T_1}^{T_2} dt\, O(t)_{++} U_{\lambda_0}(t, T_1)_{+-} + \lambda_0 R^{(1)}(T_2, T_1)_{+-}$$
$$+ i\lambda_0^2 \int_{T_1}^{T_2} dt\, R^{(1)}(T_2, t)_{+-} (O(t) U_{\lambda_0}(t, T_1))_{--}$$

[we used (2.22)],

$$R^{(1)}(T_2, T_1)_{+-}$$
$$= \lambda_0^{-1} U_{\lambda_0}(T_2, T_1)_{+-} - i \int_{T_1}^{T_2} dt\, O(t)_{++} U_{\lambda_0}(t, T_1)_{+-}$$
$$-i\lambda_0 \int_{T_1}^{T_2} dt\, R^{(1)}(T_2, t)_{+-} (O(t) U_{\lambda_0}(t, T_1))_{--}, \tag{2.97}$$

i.e., $R^{(1)}(T_2, T_1)_{+-}$ is a solution of the Volterra type integral equation

$$x(T_2, T_1) = f_1(T_2, T_1) + \int_{T_1}^{T_2} dt\, x(T_2, t) f_2(t, T_1), \tag{2.98}$$

where

$$f_1(T_2, T_1) \equiv \lambda_0^{-1} U_{\lambda_0}(T_2, T_1)_{+-} - i \int_{T_1}^{T_2} dt\, O(t)_{++} U_{\lambda_0}(t, T_1)_{+-}, \tag{2.99}$$

$$f_2(t, T_1) \equiv -i\lambda_0 (O(t) U_{\lambda_0}(t, T_1))_{--}. \tag{2.100}$$

Using by now familiar estimates, one concludes that the iteration solution of (2.98) is the unique solution which is $\|\cdot\|$-continuous on $\tilde{R}^2$. Thus,

$$R^{(1)}(T_2, T_1)_{+-} = f_1(T_2, T_1) + \sum_{n=1}^{\infty} \int_{T_1}^{T_2} dt_1 \cdots \int_{t_{n-1}}^{T_2} dt_n \times f_1(T_2, t_n) f_2(t_n, T_1) \cdots f_2(t_1, T_1). \tag{2.101}$$

In view of (2.44), (2.99) and Lemma 2.1 it follows from (2.101) that

$$\|R^{(1)}(T_2, T_1)_{+-}\|_2 \leq C_1 < \infty \tag{2.102}$$

for any $(T_2, T_1) \in Q$. The remaining statements follow from (2.102), the proof of Theorem 8 and (2.101).■

10,11: These theorems immediately follow from (2.7)–(2.8) and (2.15)–(2.17).■

The Dyson expansion is, of course, well known (cf., e.g., Ref. 13). Theorem 5 was inspired by results of Phillips.[14] The method of proof of part of Theorem 8, in particular the lemma and (2.92), are taken from a paper by Bongaarts.[7] The idea of using an integral equation to derive, as in the proof of Theorem 9, from properties of the evolution operator properties of its "Born approximation" is due to Hochstenbach.[11]

It should be noted that if one only assumes [cf. (2.5)]

$$\|O(\cdot)\| \in L^1([\alpha, \beta]), \quad -\infty < \alpha < \beta < \infty, \tag{2.103}$$

then all theorems still hold true with some obvious changes [e.g., $\tilde{R}^2 \to [\alpha, \beta] \times [\alpha, \beta]$, $K = \int_\alpha^\beta dt\, \|O(t)\|$]. This observation permits us to use the results for time-independent external fields as well. We conclude this section with some results which are useful in this connection.

We assume that H_0 and V are self-adjoint resp. bounded operators on H and that $P_\pm$ are spectral projections of H_0. Defining the closed operator

$$H(\lambda) = H_0 - \lambda V, \quad \lambda \in C, \tag{2.104}$$

one concludes, using the second Neumann series for its resolvent and the Hille–Yosida–Phillips theorem,[13] that it generates a strongly continuous group $\exp[-iH(\lambda)t]$. Defining

$$O(t) = \exp(iH_0 t) V \exp(-iH_0 t), \tag{2.105}$$

we are ready for the following theorems.

Theorem 2.12: The following relations hold:

$$\exp(-iH_0 t) U_\lambda(t, 0) = \exp[-iH(\lambda)t], \tag{2.106}$$

$$\|\exp[-iH(\lambda)t]\| \leq \exp(|\lambda| \, \|V\| \, |t|). \tag{2.107}$$

Proof: On $D(H_0)$:

$$\frac{d}{dt} \exp(iH_0 t) \exp[-iH(\lambda)t] = i\lambda\, O(t) \exp(iH_0 t) \exp[-iH(\lambda)t]. \tag{2.108}$$

Thus,

$$\exp(iH_0 t) \exp[-iH(\lambda)t]$$
$$= 1 + i\lambda \int_0^t dt' O(t') \exp(iH_0 t') \exp[-iH(\lambda)t'], \tag{2.109}$$

which holds on H by continuity. Since $U_\lambda(t, 0)$ is the unique norm continuous solution of this integral equation, (2.106) must hold. The inequality (2.107) follows from estimates analogous to (2.62) or alternatively from semigroup theory.■

Theorem 2.13: Let

$$\|\exp(-iH(\lambda_0)t)_{+-}\|_2 < \infty, \quad \forall t \in R, \tag{2.110}$$

where $\lambda_0 \in C \setminus \{0\}$. Then $R^{(1)}(T_2, T_1)_{+-}$ is H.S. for any $(T_2, T_1) \in R^2$. Moreover, $\exp(-iH(\lambda)t)_{+-}$ is an $\|\cdot\|_2$-entire function of λ for any $t \in R$. The same statements hold true if $+-$ is replaced by $-+$.

Proof: From (2.21) and (2.106) it is obvious that

$$U_{\lambda_0}(T_2, T_1) = \exp(iH_0 T_2) \exp(-iH(\lambda_0)(T_2 - T_1)) \exp(-iH_0 T_1). \tag{2.111}$$

Thus,

$$\|U_{\lambda_0}(T_2, T_1)_{+-}\|_2 = \|\exp(-iH(\lambda_0)(T_2 - T_1))_{+-}\|_2. \tag{2.112}$$

In view of Theorem 9 it therefore remains to show that

$$f(t) \equiv \| \exp(-iH(\lambda_0)t)_+ - \|_2 \tag{2.113}$$

is bounded on bounded sets of R.

From the group property and (2.107) it follows that

$$f(t_1 + t_2) \leq \exp(a|t_1|) f(t_2) + \exp(a|t_2|) f(t_1) \tag{2.114}$$

where

$$a \equiv |\lambda_0| \, \|V\| . \tag{2.115}$$

Thus, by induction,

$$f(nt) \leq n \exp[a(n-1)|t|] f(t) , \tag{2.116}$$

which implies that it suffices to prove that $f(t)$ is bounded on a neighborhood of the origin. We will derive a contradiction from the assumption that this is not true.

Indeed, assume there exists a sequence $t_i \to 0$ such that $f(t_i) \to \infty$. We assert that this implies that for any $n \in N$

$$V_n \equiv \{t \in R | f(t) > n\} \tag{2.117}$$

is dense in R. To see this, let $t_0 \in R$. Then, by (2.114),

$$\exp(a|t_0|) f(t_0 + t_i) \geq f(t_i) - \exp(a|t_0 + t_i|) f(-t_0) \tag{2.118}$$

so that $f(t_0 + t_i) \to \infty$, from which our assertion follows. On the other hand, $f_N(t) \uparrow f(t)$, where

$$f_N(t) \equiv \left(\sum_{i=1}^{N} \| \exp(-iH(\lambda_0)t)_+ - e_i \|^2 \right)^{1/2} \tag{2.119}$$

and $\{e_i\}_{i=1}^{\infty}$ is an orthonormal basis for H. Thus,

$$V_n = \bigcup_{N=1}^{\infty} f_N^{-1}(n, \infty) . \tag{2.120}$$

Since f_N is continuous, it follows from (2.120) that V_n is open. As V_n is open and dense, its complement V_n^c is nowhere dense. Thus, by the Baire category theorem, $\cup_{n=1}^{\infty} V_n^c \neq R$. Therefore, there exists a $t_0 \in R$ for which $f(t_0) > n$, $\forall n \in N$. This contradicts (2.110), and so the theorem is proved. ■

3. THE SPIN-½ CASE

Using Sec. 2 and the Appendix, we will now obtain various results in the context of the classical Dirac theory. It is assumed that the reader is familiar with the Dirac theory as treated in Refs. 15, 16. We choose the Dirac representation of the γ-algebra (Ref. 16, Chap. 2). It is easily seen that the results hold true in any representation.

A. Preliminaries

The free Dirac Hamiltonian $\check{H}_0$ is defined as an operator on the Hilbert space $\check{H} \equiv L^2(R^3, d\vec{x})^4$ by

$$\check{H}_0 = \frac{1}{i} \vec{\alpha} \cdot \vec{\nabla} + \beta m , \quad D(\check{H}_0) = W_1(R^3)^4 . \tag{3.1}$$

(For a definition of the Sobolev spaces W_m see Ref. 13.) $\check{H}_0$ is easily seen to be self-adjoint. We have occasion to use a simultaneous spectral representation of $\check{H}_0$ and the momentum operator, generated by the unitary operator

$$W : H \equiv L^2(R^3, d\vec{p})^4 \to \check{H} , \tag{3.2}$$

which satisfies

$$(Wg)(\vec{x}) \equiv \sum_{i,\epsilon} \int d\vec{p} \, W_{i,\epsilon}(\vec{x}, \vec{p}) g_\epsilon^i(\vec{p}) , \quad \epsilon = +, - , \quad i = 1, 2 , \tag{3.3}$$

$$(W^{-1} f)_\epsilon^i(\vec{p}) = \int d\vec{x} \, W_{i,\epsilon}^{-1}(\vec{x}, \vec{p}) \cdot f(\vec{x}) . \tag{3.4}$$

The integrals are limits in the mean and

$$\begin{aligned} &W_{i,\epsilon}(\vec{x}, \vec{p}) \equiv (2\pi)^{-3/2} (m/E_p)^{1/2} w_\epsilon^i(\vec{p}) \exp(i\epsilon \vec{p} \cdot \vec{x}) , \\ &W_{i,\epsilon}^{-1}(\vec{x}, \vec{p}) \equiv \bar{W}_{i,\epsilon}(\vec{x}, \vec{p}) , \\ &E_p \equiv (p^2 + m^2)^{1/2} , \quad w_+^i(\vec{p}) \equiv u_i(\vec{p}) , \quad w_-^i(\vec{p}) \equiv v_i(\vec{p}) . \end{aligned} \tag{3.5}$$

In (3.5), $u_i(\vec{p})$ and $v_i(\vec{p})$ are defined by

$$\begin{aligned} u_1(\vec{p}) &= \left(\frac{E_p + m}{2m} \right)^{1/2} \left(1, 0, \frac{p_3}{E_p + m} , \frac{p_1 + ip_2}{E_p + m} \right) \\ u_2(\vec{p}) &= \left(\frac{E_p + m}{2m} \right)^{1/2} \left(0, 1, \frac{p_1 - ip_2}{E_p + m} \frac{-p_3}{E_p + m} \right) \end{aligned} \tag{3.6}$$

and

$$v_i(\vec{p}) = \check{C} u_i(\vec{p}) , \tag{3.7}$$

where $\check{C}$ is the charge conjugation operator on $\check{H}$:

$$(\check{C} f)(\vec{x}) \equiv i\gamma^2 \bar{f}(\vec{x}) . \tag{3.8}$$

In the sequel we will use the convention

$$O \equiv W^{-1} \check{O} W \tag{3.9}$$

if $\check{O}$ is an operator on $\check{H}$. With this convention,

$$(H_0 f)_\epsilon^i(\vec{p}) = \epsilon E_p f_\epsilon^i(\vec{p}) \quad \forall f \in D(H_0) , \tag{3.10}$$

$$(Cf)_\epsilon^i(\vec{p}) = \bar{f}_{-\epsilon}^i(\vec{p}) \quad \forall f \in H . \tag{3.11}$$

We omit the easy proofs of the unitarity of W and of (3.4), (3.10)–(3.11).

We will frequently need the well-known commutator and Green's functions associated with the free Klein–Gordon and Dirac equations.[15] Our conventions for these distributions are

$$\begin{aligned} &\Delta_\epsilon(x) = -i\epsilon (2\pi)^{-3} \int d\vec{p} \exp(i\vec{p} \cdot \vec{x}) \frac{\exp(-i\epsilon E_p t)}{2E_p} , \\ &\Delta(x) = \Delta_+(x) + \Delta_-(x) , \\ &S_{(\epsilon)}(x) = (i\partial\!\!\!/ + m) \Delta_{(\epsilon)}(x) . \end{aligned} \tag{3.12}$$

Of course, the integral is a distributional Fourier transform.

$$\begin{aligned} &\Delta_R(x) = -\theta(t)\Delta(x), \quad \Delta_F(x) = -\theta(t)\Delta_+(x) + \theta(-t)\Delta_-(x) , \\ &\Delta_A(x) = \theta(-t)\Delta(x), \quad \Delta_{\bar{F}}(x) = -\theta(t)\Delta_-(x) + \theta(-t)\Delta_+(x) , \\ &S_I(x) = (i\partial\!\!\!/ + m)\Delta_I(x), \quad I = R, A, F, \bar{F} . \end{aligned} \tag{3.13}$$

As a consequence the propagators fulfill the equations

$$\begin{aligned} &(\vec{\Box}_x + m^2)\Delta_I(x - y) = \Delta_I(x - y)(\overleftarrow{\Box}_y + m^2) = \delta(x - y) , \\ &(-i\vec{\partial}\!\!\!/_x + m) S_I(x - y) = S_I(x - y)(i\overleftarrow{\partial}\!\!\!/_y + m) = \delta(x - y) . \end{aligned} \tag{3.14}$$

(Notice that the Feynman propagators differ by a constant from the usual ones.) We define the (partial) Fourier transforms of these distributions by

$$\hat{D}(t,\vec{p}) = \int d\vec{x}\, \exp(-i\vec{p}\cdot\vec{x}) D(t,\vec{x}),$$
$$\check{D}(p) = \int dx\, \exp(ipx) D(x). \tag{3.15}$$

One then has, e.g.,

$$\hat{\Delta}_R(t,\vec{p}) = \theta(t)\frac{\sin E_p t}{E_p}, \tag{3.16}$$

$$\check{\Delta}_\epsilon(p) = -2\pi i \epsilon\, \theta(\epsilon p^0)\delta(p^2 - m^2), \tag{3.17}$$

$$\check{\Delta}_I(p) = \lim_{\delta \downarrow 0} \Delta_I^\delta(p), \tag{3.18}$$

$$\Delta_{\substack{R\\A}}^\delta(p) \equiv (E_p^2 - (p_0 \pm i\delta)^2)^{-1}, \quad \Delta_{\substack{F\\\bar{F}}}^\delta(p) \equiv (m^2 - p^2 \mp i\delta)^{-1},$$

$$\check{S}_I(p) = \lim_{\delta \downarrow 0} S_I^\delta(p), \quad S_I^\delta(p) \equiv (\not{p} + m)\Delta_I^\delta(p). \tag{3.19}$$

It is convenient to define the (partial) Fourier transform of a function F in $S(R^4)$ by

$$\hat{F}(t,\vec{p}) = (2\pi)^{-3} \int d\vec{x}\, \exp(-i\vec{p}\cdot\vec{x}) F(t,\vec{x}),$$
$$\check{F}(p) = (2\pi)^{-4} \int dx\, \exp(ipx) F(x). \tag{3.20}$$

Let $K(T_2, T_1)$ be a function from R^2 to the bounded operators on H which is strongly continuous on R^2, except possibly on the set $T_2 = T_1$, and such that

$$\sup_{(T_1,T_2)\in R^2} \|K(T_2,T_1)\| < \infty. \tag{3.21}$$

We define for any $f, g \in S(R^3)^4$, resp. $F, G \in S(R^4)^4$,

$$[K]_{T_2,T_1}(f,g) = (\bar{\check{f}}, \check{K}(T_2,T_1)\check{g}), \tag{3.22}$$

$$[K](F,G) = \int dt\, dt' (\bar{\check{F}}(t,\cdot), \check{K}(t,t')\check{G}(t',\cdot)), \tag{3.23}$$

where, e.g., $\check{f}$ denotes the injection of f into $\check{H}$. In virtue of the nuclear theorem $[K]_{T_2,T_1}$ extends to a distribution in $S'(R^6)$[16] for any $(T_2, T_1) \in R^2$, while $[K]$ extends to a distribution in $S'(R^8)$[16].

The free Schrödinger picture evolution operator is defined by

$$U_0^s(T_2,T_1) = \exp[-iH_0(T_2 - T_1)]. \tag{3.24}$$

We denote the spectral projections of H_0 on $[m,\infty)$ and $(-\infty,-m]$ by P_+, resp. P_-. Using the operator W, the well-known relations

$$\sum_{i=1}^{2} w_\epsilon^i(\vec{p}) \bar{w}_\epsilon^i(\vec{p}) = (\not{p} + \epsilon m)\gamma^0/2m, \quad p \equiv (E_p, \vec{p}), \tag{3.25}$$

(3.10), and (3.12), one obtains

$$[P_\epsilon U_0^s]_{T_2,T_1}(\vec{x},\vec{y}) = iS_\epsilon(T_2 - T_1, \vec{x} - \vec{y})\gamma^0, \tag{3.26}$$

$$[P_\epsilon U_0^s](x,y) = iS_\epsilon(x-y)\gamma^0, \tag{3.27}$$

$$[\check{P}_\epsilon \exp(-i\check{H}_0 t)\check{f}](\vec{x}) = i\int d\vec{y} S_\epsilon(t, \vec{x} - \vec{y})\gamma^0 f(\vec{y}), \tag{3.28}$$

$$\left\{\int dt' \check{P}_\epsilon \exp[-i\check{H}_0(t-t')]\check{F}(t',\cdot)\right\}(\vec{x}) = i\int dy S_\epsilon(x-y)\gamma^0 F(y), \tag{3.29}$$

where f and F belong to $S(R^3)^4$, resp. $S(R^4)^4$. The integrals on the rhs stand for convolutions and the integral on the lhs is a strong Riemann integral in $\check{H}$. One easily sees that the lhs of (3.28) and (3.29) belong to $S(R^3)^4$ for any $t \in R$, and belong to $O_M(R^4)^4$ as functions of x.

B. The operators R, A, F, and $\bar{F}$

We now introduce the interaction with external fields. We shall consider functions $V(x)$ from R^4 to the Hermitean 4×4 matrices, the matrix elements of which belong to $S(R^4)$. At any time t such a function defines a bounded self-adjoint multiplication operator $\check{V}(t)$ on $\check{H}$. We define

$$H(t) = H_0 - V(t), \tag{3.30}$$

$$O(t) = \exp(iH_0 t)\, V(t) \exp(-iH_0 t). \tag{3.31}$$

It is obvious that $O(t)$ satisfies the assumptions of Sec. 2, and so we can use its results. [The coupling constant is absorbed in $V(t)$ since analyticity properties will only be used in the quantized theory.] In particular, by Theorems 2.4, 2.5, and 2.11, $\check{U}(T_2,T_1)$ $[\equiv \check{U}_1(T_2,T_1)]$ and $\check{U}^s(T_2,T_1)$ solve the interaction picture, resp. Schrödinger picture, Cauchy problems and are unitary.

We will now show that

$$[\check{I}(\infty,-\infty)_{\epsilon\epsilon'}\check{f}](\vec{x})$$
$$= -i\sum_{n=1}^{\infty}\int dx_1\cdots dx_n d\vec{x}' S_\epsilon(x-x_1)B(x_1)S_I(x_1-x_2)B(x_2)$$
$$\times\cdots S_I(x_{n-1}-x_n)B(x_n)S_{\epsilon'}(x_n-x')\gamma^0 f(\vec{x}'), \quad f\in S(R^3)^4, \tag{3.32}$$

where

$$I = R, A, F, \bar{F}, \quad x^{(\prime)} \equiv (0, \vec{x}^{(\prime)}) \tag{3.33}$$

and

$$B(x) \equiv \gamma^0 V(x). \tag{3.34}$$

[For F and $\bar{F}$ one should require $\int dt \|V(t)\| < 1$, cf. Theorem 2.2.] The integrals stand for convolutions which should be performed in the indicated order. The limit is a limit in the mean in $\check{H}$. It is easily seen that each term in the sum belongs to $S(R^3)^4$.

To prove that (3.32) holds, we observe that, e.g., $(g \in S(R^3)^4)$,

$$(\check{g}, \check{R}^{(n)}(\infty,-\infty)_{\epsilon\epsilon'}\check{f})$$
$$= i^n \int dt_1\cdots dt_n (W^{-1}\check{g}, P_\epsilon \exp(iH_0 t_1)V(t_1)\cdots$$
$$\times V(t_n)\exp(-iH_0 t_n)P_{\epsilon'}W^{-1}\check{f})$$
$$= i^n \int dt_1\cdots dt_n d\vec{p}\, d\vec{k}_1\cdots d\vec{k}_{n-1} d\vec{q} \sum_{i,i'} \overline{(W^{-1}\check{g})}_\epsilon^i(\epsilon\vec{p}) \left(\frac{m}{E_p}\right)^{1/2}$$
$$\times \bar{w}_\epsilon^i(\epsilon\vec{p})\exp(i\epsilon E_p t_1)\hat{V}(t_1, \vec{p} - \vec{k}_1)$$
$$\times\left[\sum_{\epsilon_1}(E_{k_1}\gamma^0 - \epsilon_1\vec{k}_1\cdot\vec{\gamma} + \epsilon_1 m)\right.$$
$$\left.\times\frac{\exp[-i\epsilon_1 E_{k_1}(t_1-t_2)]}{2E_{k_1}}\theta(t_1-t_2)\right]\gamma^0\hat{V}(t_2,\vec{k}_1-\vec{k}_2)\cdots$$
$$\times\gamma^0\hat{V}(t_n,\vec{k}_{n-1}-\vec{q})\exp(-i\epsilon' E_q t_n)w_{\epsilon'}^{i'}(\epsilon'\vec{q})\left(\frac{m}{E_q}\right)^{1/2}$$
$$\times(W^{-1}\check{f})_{\epsilon'}^{i'}(\epsilon'\vec{q}). \tag{3.35}$$

Notice that the factor in the brackets equals $-i\hat{S}_R(t_1-t_2, \vec{k}_1)$ and that the integral is absolutely convergent. We now choose a sequence T_N, the matrix elements of which belong to $S(R^4)$ and satisfy

$$(\hat{T}_N)_{ij}(t,\vec{k}) - (\hat{S}_R)_{ij}(t,\vec{k}) \quad \forall i,j\in\{1,\ldots,4\}, \quad \forall (t,\vec{k})\in R^4,$$
$$\sup_{t,\vec{k},i,j,N} |(\hat{T}_N)_{ij}(t,\vec{k})| < \infty . \tag{3.36}$$

Then, using dominated convergence and Fubini's theorem,

$$\begin{aligned}
&(\check{g}, \check{R}^{(n)}(\infty,-\infty)_{\epsilon\epsilon'}\check{f}) \\
&= i \lim_{N_{n-1}} \cdots \lim_{N_1} \int dt_1 \cdots d\vec{q} \cdots \hat{T}_{N_{n-1}}(t_1 - t_2, \vec{k}_1) \cdots \\
&\times \hat{T}_{N_1}(t_{n-1} - t_n, \vec{k}_{n-1}) \cdots \\
&= -i \lim_{N_{n-1}} \cdots \lim_{N_1} \int dx_1 \cdots dx_n \left[\int d\vec{x}\, \bar{g}(\vec{x}) S_\epsilon(x - x_1)\right] B(x_1) \\
&\times T_{N_{n-1}}(x_1 - x_2) \cdots T_{N_1}(x_{n-1} - x_n) B(x_n) \\
&\times \left[\int d\vec{x}'\, S_{\epsilon'}(x_n - x')\gamma^0 f(\vec{x}')\right] \\
&\equiv \lim_{N_{n-1}} \cdots \lim_{N_1} \int dx_{n-1}\, dx_n G_{N_{n-1},\ldots,N_2}(x_{n-1}) \\
&\times T_{N_1}(x_{n-1} - x_n) F(x_n),
\end{aligned} \tag{3.37}$$

where G and F are in $S(R^4)^4$. Since, by (3.36), $T_N \to S_R$ weakly, $\lim_{N_1} \int \cdots T_{N_1} \cdots$ equals $\int \cdots S_R \cdots$ by virtue of the weak continuity of the convolution.[13] In the same fashion one successively removes the other limits. Further details are left to the reader.

Similarly, one concludes that

$$\begin{aligned}
&(I(\infty,-\infty)f)^i_\epsilon(\vec{p}) \\
&= 2\pi i \sum_{n=1}^{\infty} \sum_{i',\epsilon'} \int dk_1 \cdots dk_{n-1}\, d\vec{q} \left(\frac{m}{E_p}\right)^{1/2} \\
&\times \bar{w}^i_\epsilon(\vec{p}) \tilde{B}(\epsilon p - k_1) \tilde{S}_I(k_1) \tilde{B}(k_1 - k_2) \cdots \tilde{S}_I(k_{n-1}) \\
&\times \tilde{B}(k_{n-1} - \epsilon' q) w^{i'}_{\epsilon'}(\vec{q}) \left(\frac{m}{E_q}\right)^{1/2} f^{i'}_{\epsilon'}(\vec{q}), \qquad f\in S(R^3)^4,
\end{aligned} \tag{3.38}$$

where

$$p \equiv (E_p, \vec{p}), \quad q \equiv (E_q, \vec{q}), \quad \bar{w} \equiv \bar{w}\gamma^0 . \tag{3.39}$$

The injection of f into H is denoted by f. The k_i-integrals stand again for convolutions and the limit is a limit in the mean in H. Note that each term in the sum belongs to $S(R^3)^4$ since W and W^{-1} map $S(R^3)^4$ onto $S(R^3)^4$. We can write (3.38) as

$$(I(\infty,-\infty)f)^i_\epsilon(\vec{p}) = \sum_{n=1}^{\infty} \sum_{i',\epsilon'} \int d\vec{q}\, I^{(n)}{}^{ii'}_{\epsilon\epsilon'}(\vec{p},\vec{q}) f^{i'}_{\epsilon'}(\vec{q}) . \tag{3.40}$$

By using properties of O'_c (Ref. 17, pp. 100, 124), it is easily seen that $I^{(n)ii'}_{\epsilon\epsilon'}(\vec{p},\vec{q})$ is in $S(R^3)$ in $\vec{p}$ and $\vec{q}$ separately. We further observe that

$$\begin{aligned}
&I^{(n)ii'}_{\epsilon\epsilon'}(\vec{p},\vec{q}) \\
&= 2\pi i \lim_{\delta_{n-1}\downarrow 0} \cdots \lim_{\delta_1 \downarrow 0} \int dk_1 \cdots dk_{n-1} (m/E_p)^{1/2} \\
&\times \bar{w}^i_\epsilon(\vec{p}) \tilde{B}(\epsilon p - k_1)\, S_I^{\delta_{n-1}}(k_1) \tilde{B}(k_1 - k_2) \cdots S_I^{\delta_1}(k_{n-1}) \\
&\times \tilde{B}(k_{n-1} - \epsilon' q) w^{i'}_{\epsilon'}(\vec{q}) (m/E_q)^{1/2} .
\end{aligned} \tag{3.41}$$

Evidently, the integral in (3.41) is absolutely convergent.

C. Lorentz covariance, causality

We will now show that the operators $I(\infty,-\infty)$ are Lorentz covariant in the sense that for, e.g., an electromagnetic field

$$U(a,\Lambda) I(A_\mu) U^*(a,\Lambda) = I(A_\mu{}^{a,\Lambda}), \tag{3.42}$$

where

$$A_\mu{}^{a,\Lambda}(x) \equiv \Lambda_\mu{}^\nu A_\nu(\Lambda^{-1}(x-a)) . \tag{3.43}$$

In (3.42), $U(a,\Lambda)$ is the representation of $i\mathrm{SL}(2,C)$ in H (we do not consider inversions):

$$\begin{aligned}
(U(a,\Lambda)g)^i_\epsilon(\vec{p}) &\equiv \exp(i\epsilon pa) \left(\frac{(\Lambda^{-1}p)_0}{p_0}\right)^{1/2} \\
&\times \sum_j \overline{\left[\left(\frac{\tilde{p}}{m}\right)^{1/2} A \left(\frac{\widetilde{\Lambda^{-1}p}}{m}\right)^{1/2}\right]}^{\epsilon}_{ij} g^j_\epsilon(\overrightarrow{\Lambda^{-1}p}),
\end{aligned} \tag{3.44}$$

where

$$\overline{f(\vec{p})}^\epsilon \equiv \begin{cases} f(\vec{p}), & \epsilon = +, \\ \overline{f(\vec{p})}, & \epsilon = -. \end{cases} \tag{3.45}$$

In (3.44) the notation of Ref. 18 is used. The validity of (3.42) follows from (3.38), (3.44), the relation

$$\sum_i \overline{\left[\left(\frac{\widetilde{\Lambda^{-1}p}}{m}\right)^{1/2} A^{-1} \left(\frac{\tilde{p}}{m}\right)^{1/2}\right]}^{\epsilon}_{ij} w^i_\epsilon(\overrightarrow{\Lambda^{-1}p}) = S(A^{-1}) w^j_\epsilon(\vec{p}), \tag{3.46}$$

and the well-known relations

$$\begin{aligned}
&S(A)^* \gamma^0 S(A)^{-1} , \\
&S(A)^{-1} \gamma^\mu S(A) = \Lambda^\mu{}_\nu \gamma^\nu .
\end{aligned} \tag{3.47}$$

[To see that (3.46) holds for u_i, use

$$(u_i(\vec{p}))_j = S\left(\left(\frac{\tilde{p}}{m}\right)^{1/2}\right)_{ji} \tag{3.48}$$

and

$$S(A) = \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix} \quad \forall A\in SU(2) . \tag{3.49}$$

To prove it for v_i, use

$$\bar{A} = \zeta A \zeta^{-1}, \quad \forall A\in SU(2) \tag{3.50}$$

where

$$\zeta \equiv \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{3.51}$$

and

$$(v_i'(\vec{p}))_j = S\left(\left(\frac{p}{m}\right)^{1/2}\right)_{j,i+2}, \tag{3.52}$$

where

$$v_i'(\vec{p}) \equiv \sum_j \xi^{-1}{}_{ij} v_j(\vec{p}). \tag{3.53}$$

Of course, (3.48), (3.49), and (3.52) only hold true in the Dirac representation of the γ-algebra. However, (3.46) holds in any representation.]

Introducing the classical S operator

$$S(V) \equiv U(\infty, -\infty; V) \tag{3.54}$$

we now have the following theorem.

Theorem 3.1: $S(V)$ is Lorentz covariant and causal in the sense that

$$S(V_1 + V_2) = S(V_1)S(V_2) \tag{3.55}$$

if the supports of V_1 and V_2 are spacelike separated.

Proof: The Lorentz covariance is obvious from (3.42). Defining

$$R(V) = R(\infty, -\infty; V), \tag{3.56}$$

we have [cf. (3.32) and its proof]

$$[\check{R}(V_1)\check{R}(V_2)\check{f}\,](\vec{x})$$
$$= \sum_{n,m=1}^{\infty} \int \cdots B_1(x_n) S(x_n - x_{n+1}) B_2(x_{n+1}) \cdots f(\vec{x}') = 0,$$
$$f \in S(R^3)^4 \tag{3.57}$$

so that

$$R(V_1)R(V_2) = 0. \tag{3.58}$$

Similarly, from (3.32),

$$R(V_1 + V_2) = R(V_1) + R(V_2). \tag{3.59}$$

(3.55) now follows from (3.58) and (3.59). ■

We now define

$$G_{\substack{R\\A}}(T_2, T_1) = \pm i U^s(T_2, T_1)\gamma^0 \theta(\pm(T_2 - T_1)). \tag{3.60}$$

Then

$$[G_I](F, G)$$
$$= \sum_{n=0}^{\infty} \int dx\, dx_1 \cdots dx_n\, dy\, F(x) S_I(x - x_1) B(x_1)$$
$$\times S_I(x_1 - x_2) \cdots B(x_n) S_I(x_n - y) G(y), \qquad I = R, A. \tag{3.61}$$

To see this, note that, e.g.,

$$\int dt\, dt' (\bar{\check{F}}(t, \cdot), i\check{U}^s(t, t')\gamma^0 \theta(t - t')\check{G}(t', \cdot))$$
$$= \sum_{n=0}^{\infty} i^{n+1} \int_{-\infty}^{\infty} dt_0 \int_{-\infty}^{t_0} dt_{n+1} \int_{t_{n+1}}^{t_0} dt_1 \cdots \int_{t_{n+1}}^{t_{n-1}} dt_n$$
$$\times (W^{-1}\bar{F}(t_0, \cdot), \exp[-iH_0(t_0 - t_1)]V(t_1) \cdots$$
$$\times V(t_n)\exp[-iH_0(t_n - t_{n+1})]W^{-1}\check{G}(t_{n+1}, \cdot))$$
$$= \sum_{n=0}^{\infty} i^{n+1} \int_{-\infty}^{\infty} dt_0 \cdots \int_{-\infty}^{t_n} dt_{n+1} (\cdots), \tag{3.62}$$

and then proceed as in the proof of (3.32). Since the arguments of the $S_R(S_A)$ in (3.61) must be all in the closed forward (backward) light cone, denoted by $\bar{V}_+(\bar{V}_-)$, to give a nonzero contribution, their sum, $x - y$, must be in $\bar{V}_+(\bar{V}_-)$. Hence,

$$\operatorname{supp}\left[G_{\substack{R\\A}}\right](x, y) \subset \{(x, y) \mid x - y \in \bar{V}_{\pm}\}. \tag{3.63}$$

Furthermore, using (3.14), we have

$$[-i\vec{\partial}_x + m - B(x)][G_I](x, y) = [G_I](x, y)[i\vec{\partial}_y + m - B(y)]$$
$$= \delta(x - y), \qquad I = R, A. \tag{3.64}$$

Thus, $[G_{R,A}]$ are two-sided retarded, resp. advanced, fundamental solutions of the perturbed Dirac equation. We note that the existence of retarded and advanced solutions, given by (3.61), also follows from Theorem A.1. This theorem, moreover, implies that they are regular in the sense defined by Wightman,[19] i.e., they define continuous maps from $S(R^4)^4$ into $O_M(R^4)^4$.

It should be noted that all results obtained thus far also hold for $V(x) \in S(R^4)^{16}$ since we did not use the self-adjointness of $V(t)$. In particular, if one defines G_I by (3.60) for these V, then (3.61), (3.63), and (3.64) again hold true. Thus, we have proved:

Theorem 3.2: For any $V \in S(R^4)^{16}$ the perturbed Dirac equation admits two-sided tempered retarded and advanced fundamental solutions, given by (3.61). They are connected with the evolution operator through (3.60) and (3.23), and are Wightman regular. ■

D. Implementability of the evolution in $\mathcal{F}_a(H)$

It is well known that V can be decomposed in 16 fields of five different tensor types: scalar, vector, tensor, pseudovector, and pseudoscalar. We will now show that the hypothesis of Theorem 2.8 is satisfied if only the timelike component of the vector field or the pseudovector field is nonzero. In the former case (only electric field)

$$[\check{V}_1(t)\phi](\vec{x}) \equiv F(t, \vec{x})\phi(\vec{x}) \quad \forall \phi \in H, \tag{3.65}$$

where F is a real-valued function in $S(R^4)$. In the latter case

$$[\check{V}_2(t)\phi](\vec{x}) \equiv F(t, \vec{x})(\gamma^5\phi)(\vec{x}), \tag{3.66}$$

where

$$\gamma^5 \equiv i\gamma^0\gamma^1\gamma^2\gamma^3. \tag{3.67}$$

Theorem 3.3: Let

$$O_k(t) \equiv \exp(iH_0 t)V_k(t)\exp(-iH_0 t), \quad k = 1, 2. \tag{3.68}$$

Then the operators $\int_{T_1}^{T_2} dt\, O_k(t)_{\pm\mp}$, $k = 1, 2$, are $\|\cdot\|_2$-continuous on $\bar{R}^2$ and

$$\left\| \int_{T_1}^{T_2} dt\, O_k(t)_{\pm\mp} \right\|_2 \leq C_1 < \infty \quad \forall (T_2, T_1) \in \bar{R}^2. \tag{3.69}$$

Proof: It is easily seen that it suffices to prove our statements for the integral operators on $L^2(R^3)^4$ with the kernels

$$P_\pm(\vec{p})\int_{T_1}^{T_2} dt \exp[\pm i(E_p+E_q)t]\hat{F}(t,\vec{p}-\vec{q})(\gamma^5)^{k-1}P_\mp(\vec{q}),$$
$$k=1,2, \tag{3.70}$$

where

$$P_\pm(\vec{k}) \equiv \frac{1}{2}\left(1 \pm \frac{\vec{\alpha}\cdot\vec{k}+\beta m}{E_k}\right). \tag{3.71}$$

Since, for any $a\in R$ and $(T_2,T_1)\in R^2$,

$$a\int_{T_1}^{T_2} dt \exp(iat)\hat{F}(t,\vec{k})$$
$$= -i\exp(iaT_2)\hat{F}(T_2,\vec{k}) + i\exp(iaT_1)\hat{F}(T_1,\vec{k})$$
$$+ i\int_{T_1}^{T_2} dt \exp(iat)\partial_t\hat{F}(t,\vec{k}), \tag{3.72}$$

there exists a $C>0$ such that for any $(T_2,T_1)\in R^2$, $a\in R$, $\vec{k}\in R^3$,

$$|a|\left|\int_{T_1}^{T_2} dt \exp(iat)\hat{F}(t,\vec{k})\right| \le \frac{C}{(1+k^2)^2}. \tag{3.73}$$

From well-known trace relations it follows that

$$\mathrm{Tr}P_\pm(\vec{p})(\gamma^5)^{k-1}P_\pm(\vec{q})(\gamma^5)^{k-1} = 1 - \frac{\vec{p}\cdot\vec{q}\pm m^2}{E_pE_q}, \quad k=\tfrac{1}{2}. \tag{3.74}$$

Thus, (3.69) holds if

$$\int d\vec{p}\,d\vec{q}\left(1-\frac{\vec{p}\cdot\vec{q}\pm m^2}{E_pE_q}\right)(E_p+E_q)^{-2}(1+|\vec{p}-\vec{q}|^2)^{-4}<\infty. \tag{3.75}$$

Taking the upper sign, one sees that the integral belongs to a class of integrals investigated in detail and proved to be convergent by Bongaarts.[7] Using his estimates, one easily sees that the lower sign leads to a convergent integral as well. The $\|\cdot\|_2$-continuity follows from dominated convergence. ■

Corollary 3.4: The evolution operator $U(T_2,T_1)$, corresponding to an electric field in $S(R^4)$, or to the time-like component [in $S(R^4)$] of a pseudovector field, or to both together, induces a unitarily implementable Bogoliubov transformation in $\mathfrak{I}_a(H)$ for any $(T_2,T_1)\in R^2$.

Proof: This follows from Theorems 3.3, 2.8, and Ref. 1. ■

We now define, for any $\phi\in\check{H}$,

$$(\check{V}_k\phi)(\vec{x}) = f(\vec{x})(\Gamma_k\phi)(\vec{x}), \quad k=1,\ldots,16, \tag{3.76}$$

where f is a real-valued function in $S(R^3)$ and where $\Gamma_1,\ldots,\Gamma_{16}$ are the 16 Hermitean matrices $1, \gamma^5, (i)\gamma^\mu, (i)\gamma^\mu\gamma^\nu, (i)\gamma^\mu\gamma^\nu\gamma^\rho$. Defining

$$O^k(t) = \exp(iH_0t)V_k\exp(-iH_0t), \quad k=1,\ldots,16, \tag{3.77}$$

one concludes, as in the proof of Theorem 3.3, that

$$\left\|\int_{T_1}^{T_2} dt O^k(t)_{\pm\mp}\right\|_2^2$$
$$= 4\int d\vec{p}\,d\vec{q}\,\mathrm{Tr}(P_\pm(\vec{p})\Gamma_kP_\mp(\vec{q})\Gamma_k)\sin^2[\alpha(E_p+E_q)](E_p+E_q)^{-2}$$
$$\times|\hat{f}(\vec{p}-\vec{q})|^2, \quad k=1,\ldots,16, \quad (T_2,T_1)\in R^2, \tag{3.78}$$

where

$$\alpha \equiv (T_2-T_1)/2. \tag{3.79}$$

It clearly follows from (3.75) that the integral converges for $k=1,2$. Defining

$$H_k = H_0 - V_k, \quad k=1,\ldots,16, \tag{3.80}$$

it then follows from Theorem 2.8, (2.106), and Ref. 1 that $\exp(-iH_kt)$ is implementable in $\mathfrak{I}_a(H)$ for any $t\in R$ and $k=1,2$. However, for $k>2$ we have

Theorem 3.5: Let $(T_2,T_1)\in R^2$ and $T_2\neq T_1$. Then the operators $\int_{T_1}^{T_2} dt O^k(t)_{\pm\mp}$ are not H.S. if $V_k\neq 0$, $k=3,\ldots,16$.

Proof: It suffices to show that the integral in (3.78) diverges for $k=3,\ldots,16$ if $\alpha f\neq 0$. Since, if $k>2$, Γ_k anticommutes with at least one α_i, $i=1,2,3$, one obtains for the trace one of the two expressions on the rhs of (3.74), but with at least one of the terms containing p_iq_i having the opposite sign. Thus, by subtracting a convergent integral the rhs of (3.78) becomes

$$8\int d\vec{p}\,d\vec{q}\sum_{i=1}^{3}\delta_i(k)p_iq_i\sin^2[\alpha(E_p+E_q)]$$
$$\times(E_pE_q)^{-1}(E_p+E_q)^{-2}|\hat{f}(\vec{p}-\vec{q})|^2, \tag{3.81}$$

where $\delta_i(k)=0$ or 1; for fixed $k>2$ at least one of the $\delta_i(k)$ is 1. Introducing variables $\vec{s}=\vec{p}+\vec{q}$, $\vec{t}=\vec{p}-\vec{q}$ and adding a convergent integral, we obtain

$$\frac{1}{4}\sum_{i=1}^{3}\delta_i(k)\int d\vec{s}\,d\vec{t}\,s_i^2\sin^2[\alpha(E_++E_-)]$$
$$\times(E_+E_-)^{-1}(E_++E_-)^{-2}|\hat{f}(\vec{t})|^2, \tag{3.82}$$

where

$$E_\pm \equiv (\tfrac{1}{4}|\vec{s}\pm\vec{t}|^2+m^2)^{1/2}. \tag{3.83}$$

We now assume that the integral

$$\int d\vec{s}\,d\vec{t}\,s_{i_0}^2\sin^2[\alpha(E_++E_-)](E_+E_-)^{-1}(E_++E_-)^{-2}|\hat{f}(\vec{t})|^2,$$
$$i_0\in\{1,2,3\} \tag{3.84}$$

converges. By introducing polar variables (s,θ,ϕ) it then follows from Fubini's theorem that there exist a $\vec{t}_0$ with $\hat{f}(\vec{t}_0)\neq 0$ and $\theta_0,\phi_0\neq 0,\pi/2,\pi,3\pi/2$ such that the integral over s converges. This integral is a nonzero multiple of

$$\int_0^\infty ds s^4\sin^2[\alpha(E_+^0+E_-^0)](E_+^0E_-^0)^{-1}(E_+^0+E_-^0)^{-2}, \tag{3.85}$$

where

$$E_\pm^0 = (\tfrac{1}{4}s^2+\tfrac{1}{4}t_0^2\pm\tfrac{1}{2}st_0c_0+m^2)^{1/2}, \tag{3.86}$$

in which c_0 is the cosine of the angle between (θ_0,ϕ_0) and $\vec{t}_0$. However, it is easily seen that (3.85) diverges. Thus, (3.84) diverges. Therefore, (3.78) diverges for $k>2$. ■

Corollary 3.6: Assume that $\exp(-iH_{k_0}t)$, $k_0\in\{3,\ldots,16\}$, is implementable in $\mathfrak{I}_a(H)$ for any $t\in R$. Then V_{k_0} is zero.

Proof: According to Ref. 1, Theorem 6.1, the assumption implies that the hypothesis of Theorem 2.13 is satisfied. Thus, $\int_{T_1}^{T_2} dt O^{k_0}(t)_{+-}$ is H.S. for any $(T_2,T_1)\in R^2$. In view of Theorem 3.5 this implies that V_{k_0} is zero. ■

Finally, we define

$$(\check{V}\phi)(\vec{x}) = \sum_{k=1}^{16} g_k f_k(\vec{x})(\Gamma_k\phi)(\vec{x}), \tag{3.87}$$

$$O'(t) = \exp(iH_0t)V\exp(-iH_0t), \tag{3.88}$$

$$H = H_0 - V, \tag{3.89}$$

where the g_k are real coupling constants and f_k are nonzero real-valued functions in $S(R^3)$. It is, of course, quite plausible that the operators $\int_{T_1}^{T_2} dt\, O'(t)_{\pm\mp}$ are only H.S. for all $(T_2, T_1)\in R^2$ if $g_k = 0$ for $k > 2$. If this is true, then the assumption that $\exp(-iHt)$ is implementable for any $t\in R$ clearly implies that $g_k = 0$ for $k > 2$. However, it does not seem easy to prove this, as cancellations may occur. In any case, these operators are in general not H.S., in the sense that if they are H.S. for some set of $\{g_k\}$, then they are not H.S. if one changes one of the g_k with $k > 2$. Thus, if some g_k with $k > 2$ are nonzero, then $\exp(-iHt)$ is in general not implementable. Clearly, if one replaces the f_k in (3.87) by real-valued F_k in $S(R^4)$, similar remarks can be made regarding $U(T_2, T_1)$. In particular, we conjecture that Theorem 3.3 will not hold for any $V(x)$ having nonzero components for $k > 2$.

4. THE SPIN-0 CASE

In this section we treat the Klein–Gordon case along the same lines as the Dirac case. To stress the analogy the same symbols are used for corresponding operators and distributions.

A. Preliminaries

We write the free Klein–Gordon equation

$$(\Box + m^2)\phi(x) = 0 \tag{4.1}$$

in the two-component form

$$\check{H}_0\begin{pmatrix}\phi_1(x)\\ \phi_2(x)\end{pmatrix} = i\partial_t\begin{pmatrix}\phi_1(x)\\ \phi_2(x)\end{pmatrix}, \tag{4.2}$$

where

$$\check{H}_0 \equiv \begin{pmatrix}0 & 1\\ -\Delta + m^2 & 0\end{pmatrix}. \tag{4.3}$$

Obviously, the upper component of a weak solution of (4.2) satisfies (4.1). We define a multiplication operator $\hat{H}_0$ on the Hilbert space $\hat{H} \equiv L^2(R^3, E_p\,d\vec{p})\oplus L^2(R^3, E_p^{-1}\,d\vec{p})$ by

$$\hat{H}_0 = \begin{pmatrix}0 & 1\\ E_p^2 & 0\end{pmatrix}. \tag{4.4}$$

$\hat{H}_0$ is clearly self-adjoint on

$$D(\hat{H}_0) \equiv L^2(R^3, E_p^3\,d\vec{p})\oplus L^2(R^3, E_p\,d\vec{p}). \tag{4.5}$$

Since the functions in $\hat{H}$ are tempered distributions, we can introduce the space $\check{H}$ of distributional Fourier transforms, which becomes a Hilbert space by defining as inner product

$$(\check{f}, \check{g}) = (\hat{f}, \hat{g}) \quad \forall \check{f}, \check{g}\in\check{H}. \tag{4.6}$$

Evidently,

$$\check{H} = W_{1/2}(R^3)\oplus W_{-1/2}(R^3). \tag{4.7}$$

We can now regard (4.2) as an equation in $\check{H}$: The action of $\check{H}_0$ is defined by

$$\check{H}_0\check{f} = \hat{H}_0\hat{f} \quad \forall \hat{f}\in D(\hat{H}_0); \tag{4.8}$$

$\check{H}_0$ is then self-adjoint on

$$D(\check{H}_0) = W_{3/2}(R^3)\oplus W_{1/2}(R^3). \tag{4.9}$$

Observe that the distributions in $D(\check{H}_0)$ are square-integrable functions.

We introduce a unitary operator

$$W : H \equiv L^2(R^3, d\vec{p})^2 \to \check{H}, \tag{4.10}$$

which gives rise to a simultaneous spectral representation of $\check{H}_0$ and the momentum operator (defined on $\hat{H}$ as multiplication by $\vec{p}$):

$$(Wg)(\vec{x}) \equiv \sum_\epsilon \int d\vec{p}\, W_\epsilon(\vec{x}, \vec{p})g_\epsilon(\vec{p}), \quad \epsilon = +, -, \tag{4.11}$$

$$(W^{-1}f)_\epsilon(\vec{p}) = \int d\vec{x}\, W_\epsilon^{-1}(\vec{x}, \vec{p})\cdot f(\vec{x}), \tag{4.12}$$

where

$$\begin{aligned} W_\epsilon(\vec{x}, \vec{p}) &\equiv (2\pi)^{-3/2} w_\epsilon(\vec{p})\exp(i\epsilon\vec{p}\cdot\vec{x}),\\ W_\epsilon^{-1}(\vec{x}, \vec{p}) &\equiv (2\pi)^{-3/2} w_\epsilon^*(\vec{p})\exp(-i\epsilon\vec{p}\cdot\vec{x}), \end{aligned} \tag{4.13}$$

in which

$$\begin{aligned} w_+(\vec{p}) &\equiv 2^{-1/2}\begin{pmatrix}E_p^{-1/2}\\ E_p^{1/2}\end{pmatrix}, \quad w_+^*(\vec{p}) \equiv 2^{-1/2}\begin{pmatrix}E_p^{1/2}\\ E_p^{-1/2}\end{pmatrix},\\ w_-(\vec{p}) &\equiv 2^{-1/2}\begin{pmatrix}E_p^{-1/2}\\ -E_p^{1/2}\end{pmatrix}, \quad w_-^*(\vec{p}) \equiv 2^{-1/2}\begin{pmatrix}E_p^{1/2}\\ -E_p^{-1/2}\end{pmatrix}. \end{aligned} \tag{4.14}$$

The integrals are distributional Fourier transforms. We define the charge conjugation operator $\check{C}$ on $\check{H}$ by

$$(\check{C}f)_i(\vec{x}) = (-)^{i+1}\bar{f}_i(\vec{x}), \quad i = 1, 2. \tag{4.15}$$

Then the following relations hold [cf. (3.9)]:

$$(H_0f)_\epsilon(\vec{p}) = \epsilon E_p f_\epsilon(\vec{p}) \quad \forall f\in D(H_0), \tag{4.16}$$

$$(Cf)_\epsilon(\vec{p}) = \bar{f}_{-\epsilon}(\vec{p}) \quad \forall f\in H. \tag{4.17}$$

The straightforward proofs of the unitarity of W and of the relations (4.12), (4.16)–(4.17) are omitted.

Denoting the spectral projections of H_0 on $[m, \infty)$ and $(-\infty, -m]$ by P_+, resp. P_-, we have [cf. (2.2)]

$$(qf)_\epsilon(\vec{p}) = \epsilon f_\epsilon(\vec{p}) \quad \forall f\in H. \tag{4.18}$$

It easily follows that for vectors $\phi, \psi\in\check{H}$ which are square-integrable functions:

$$(\psi, \check{q}\phi) = \int d\vec{x}\,[\bar{\psi}_1(\vec{x})\phi_2(\vec{x}) + \bar{\psi}_2(\vec{x})\phi_1(\vec{x})]. \tag{4.19}$$

Thus, $\check{q}$ corresponds to the well-known indefinite metric of the Klein–Gordon theory.

Proceeding now as in Sec. 3, but replacing (3.22)–(3.23) by

$$[K]_{T_2, T_1}(f, g) \equiv (\check{L}\bar{\check{f}}, \check{q}\check{K}(T_2, T_1)\check{g}), \tag{4.20}$$

$$[K](F, G) \equiv \int dt\, dt'(\check{L}\bar{\check{F}}(t, \cdot), \check{q}\check{K}(t, t')\check{G}(t', \cdot)), \tag{4.21}$$

where

$$\check{L}\equiv\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix},\quad D(L)\equiv D(H_0), \tag{4.22}$$

one obtains (3.26)–(3.29) with $S_\epsilon\gamma^0\to S_\epsilon$, where

$$S_\epsilon(x)\equiv\begin{pmatrix} i\partial_t & 1\\ -\Delta+m^2 & i\partial_t\end{pmatrix}\Delta_\epsilon(x). \tag{4.23}$$

In particular, for $f\in S(R^3)^2$,

$$[\check{P}_\epsilon\exp(-i\check{H}_0t)\check{f}](\vec{x})=i\int d\vec{y}\,S_\epsilon(t,\vec{x}-\vec{y})\cdot f(\vec{y}). \tag{4.24}$$

Again one concludes that the lhs of (4.24) belongs to $S(R^3)^2$ for any $t\in R$, and belongs to $O_M(R^4)^2$ as a function of x. We further note that

$$\overrightarrow{D_x^0}S_I(x-y)=S_I(x-y)\overleftarrow{D_y^0}=\delta(x-y),\quad I=R,A,F,\bar{F}, \tag{4.25}$$

where

$$D^0\equiv\begin{pmatrix}-i\partial_t & 1\\ -\Delta+m^2 & -i\partial_t\end{pmatrix}. \tag{4.26}$$

B. The operators R, A, F, and $\bar{F}$

If electromagnetic and scalar fields are present [which we assume to be real-valued functions in $S(R^4)$], the Klein–Gordon equation reads

$$\{[\partial_\mu-iA_\mu(x)][\partial^\mu-iA^\mu(x)]+m^2-A_4(x)\}\phi(x)=0. \tag{4.27}$$

We write it in the following two-component form:

$$\check{H}(t)\begin{pmatrix}\phi_1(x)\\ \phi_2(x)\end{pmatrix}=i\partial_t\begin{pmatrix}\phi_1(x)\\ \phi_2(x)\end{pmatrix}, \tag{4.28}$$

where

$$\check{H}(t)\equiv\begin{pmatrix}-A_0(x) & 1\\ -\Delta+m^2-i\vec{\nabla}\cdot\vec{A}(x)-i\vec{A}(x)\cdot\vec{\nabla}+A^2(x)-A_4(x) & -A_0(x)\end{pmatrix}. \tag{4.29}$$

Clearly, the upper component of a weak solution of (4.28) satisfies (4.27). One easily sees that, formally,

$$(H(t)g)_\epsilon(\vec{p})=\epsilon E_p g_\epsilon(\vec{p})-\epsilon\sum_{\epsilon'}\int d\vec{q}\,(4E_pE_q)^{-1/2}[\hat{A}_0(t,\epsilon\vec{p}-\epsilon'\vec{q})(\epsilon E_p+\epsilon' E_q)-\vec{A}(t,\epsilon\vec{p}-\epsilon'\vec{q})\cdot(\epsilon\vec{p}+\epsilon'\vec{q})$$
$$-\widehat{A^2}(t,\epsilon\vec{p}-\epsilon'\vec{q})+\hat{A}_4(t,\epsilon\vec{p}-\epsilon'\vec{q})]\,g_{\epsilon'}(\vec{q}). \tag{4.30}$$

Since convolution with an L^1 function is a bounded operation on L^2, the interaction term defines a bounded operator on H if

$$E_p^{1/2}\hat{A}_\mu(t,\vec{p}),\hat{A}_4(t,\vec{p})\in L^1(R^3,d\vec{p}). \tag{4.31}$$

Thus, $H(t)$ is actually a well-defined closed operator on $D(H_0)$ for any $t\in R$. We write (4.30), with obvious notation, as

$$H(t)=H_0-V(t), \tag{4.32}$$

$$V(t)=\sum_{k=e,m,s}V_k(t). \tag{4.33}$$

From (4.29) and (4.15) it follows that if $A_4(x)=0$ and the dependence on the electric charge e (which is absorbed in A_μ) is explicitly indicated,

$$CH_e(t)=-H_{-e}(t)C, \tag{4.34}$$

which motivates the name charge conjugation operator for C.

Defining

$$O(t)=\exp(iH_0t)V(t)\exp(-iH_0t), \tag{4.35}$$

it is clear that $O(t)$ satisfies the conditions of Sec. 2. However, λ can only be interpreted as the coupling constant if there is no magnetic field present. If the magnetic field is nonzero one should replace $\lambda\check{O}(t)$ by

$$\check{O}(t,\lambda)\equiv\lambda\exp(i\check{H}_0t)\times\begin{pmatrix}A_0(t,\cdot) & 0\\ i\vec{\nabla}\cdot\vec{A}(t,\cdot)+i\vec{A}(t,\cdot)\cdot\vec{\nabla}+A_4(t,\cdot) & A_0(t,\cdot)\end{pmatrix}\times\exp(-i\check{H}_0t)-\lambda^2\exp(i\check{H}_0t)\begin{pmatrix}0 & 0\\ A^2(t,\cdot) & 0\end{pmatrix}\exp(-i\check{H}_0t). \tag{4.36}$$

The results of Sec. 2 have straightforward analogs for this case. In particular, one should now define l [cf. (2.6)] as the supremum of the numbers $\alpha>0$ such that

$$\int dt\,\|O(t,\lambda)\|<1\quad\forall\lambda\in D_\alpha \tag{4.37}$$

[cf. (2.14)].

It follows from (4.18) and (4.30) that

$$V^*(t)=qV(t)q, \tag{4.38}$$

i.e., $V(t)$ is pseudo-self-adjoint (w.r.t. q). Theorem 2.10 then implies that U and U^s are pseudo-unitary. In view of Theorems 2.4 and 2.5 these operators solve the interaction picture, resp. Schrödinger picture, Cauchy problems.

Before we obtain the analog of (3.32), we make some remarks. Clearly, (4.38) implies that $qH(t)$ is self-adjoint. Moreover, using (4.19) and (4.29), one infers that for $f\in D(\check{H}_0)$

$$(f, q\check{H}(t)f)$$

$$= \int d\vec{x}\{|[\vec{\nabla} + i\vec{A}(t,\vec{x})]f_1(\vec{x})|^2 + [m^2 - A_0^2(t,\vec{x}) - A_4(t,\vec{x})]|f_1(\vec{x})|^2 + |f_2(\vec{x}) - A_0(t,\vec{x})f_1(\vec{x})|^2\}. \quad (4.39)$$

Thus, $qH(t)$ is strictly positive if

$$\sup_{\vec{x}\in R^3} |A_0^2(t,\vec{x}) + A_4(t,\vec{x})| < m^2. \quad (4.40)$$

Assuming now that the external fields are time-independent real-valued functions satisfying (4.31), it follows from the pseudo-unitarity of $\exp(-iHt)$ (cf. Theorems 2.10, 2.12) that on $D(\check{H})$

$$(\exp(-i\check{H}t)f, q\check{H}\exp(-i\check{H}t)f) = (f, q\check{H}f) \quad \forall t\in R, \quad (4.41)$$

i.e., the "energy" of solutions is time-independent. If the fields, moreover, satisfy (4.40), one can make $W_1(R^3)\oplus W_0(R^3)$ $[=D((q\check{H})^{1/2})]$ into a Hilbert space $\mathcal{H}_v$ by setting

$$(f,g)_{\mathcal{H}_v} \equiv ((q\check{H})^{1/2}f, (q\check{H})^{1/2}g) \quad \forall f,g\in\mathcal{H}_v. \quad (4.42)$$

In view of (4.41), $\exp(-i\check{H}t)\restriction D(\check{H})$ extends to a $\|\cdot\|_{\mathcal{H}_v}$-strongly continuous one-parameter unitary group in $\mathcal{H}_v$, the generator of which is a restriction of $\check{H}$.

In Ref. 20, Lundberg presents a scattering theory for the time-independent case, which is based on a different two-component equation [if $A_0(\vec{x})\neq 0$]. It would be worthwhile to investigate to what extent the two approaches are equivalent. Of course, on the level of solutions to the partial differential equation (4.27), one obtains the same results, like the time independence of the "charge" and the "energy." However, it is, e.g., not clear whether the spectrum of the operator $\check{H}$ on $\mathcal{H}_v$ and the spectrum of his operator A on $\mathcal{H}$ are equal, or whether the two approaches lead to the same quantized theory in the time-dependent case (cf. Refs. 2 and 10). Similar remarks can be made regarding the Pétiau–Duffin–Kemmer approach to the Klein–Gordon equation (cf. Ref. 19).

Proceeding now again as in Sec. 3, one obtains, using (4.24) [cf. (3.32) and its proof]:

$$[\check{I}(\infty,-\infty)_{\epsilon\epsilon'}\check{f}](\vec{x})$$

$$= -i\sum_{n=1}^{\infty}\int dx_1\cdots dx_n d\vec{x}' S_\epsilon(x-x_1)B(x_1)S_I(x_1-x_2)B(x_2)\times\cdots S_I(x_{n-1}-x_n)B(x_n)S_{\epsilon'}(x_n-x')f(\vec{x}'), \quad f\in S(R^3)^2, \quad (4.43)$$

where

$$B(x) \equiv \begin{pmatrix} A_0(x) & 0 \\ i\vec{\nabla}\cdot\vec{A}(x) + i\vec{A}(x)\cdot\vec{\nabla} - A^2(x) + A_4(x) & A_0(x) \end{pmatrix}. \quad (4.44)$$

For F and $\overline{F}$ it is assumed that $l>1$. As in Sec. 3, each term belongs to $S(R^3)^2$.

The analog of (3.38) is

$$(I(\infty,-\infty)f)_\epsilon(\vec{p})$$

$$= 2\pi i\epsilon\sum_{n=1}^{\infty}\sum_{\epsilon'}\int dk_1\cdots dk_{n-1}d\vec{q}(2E_p)^{-1/2}\times[\tilde{V}_1(\epsilon p,k_1)\tilde{\Delta}_I(k_1)\tilde{V}_1(k_1,k_2)\cdots\tilde{\Delta}_I(k_{n-1})\tilde{V}_1(k_{n-1},\epsilon' q) + \text{all } \tilde{A}_0\tilde{A}^0\text{-contractions}](2E_q)^{-1/2}f_{\epsilon'}(\vec{q}), \quad f\in S(R^3)^2, \quad (4.45)$$

where

$$\tilde{V}_1(k,k') \equiv \tilde{A}_\mu(k-k')(k^\mu + k'^\mu) - \widetilde{A^2}(k-k') + \tilde{A}_4(k-k'). \quad (4.46)$$

An $\tilde{A}_0\tilde{A}^0$-contraction of the term in brackets is by definition the same term where one or several different triplets $\tilde{V}_1(k_i,k_{i+1})\tilde{\Delta}_I(k_{i+1})\tilde{V}_1(k_{i+1},k_{i+2})$ are replaced by $\tilde{A}_0(k_i-k_{i+1})\tilde{A}^0(k_{i+1}-k_{i+2})$ $(i=0,\ldots,n-2,\ k_0\equiv\epsilon p,\ k_n\equiv\epsilon' q)$.

To see that (4.45) holds, obtain the analog of (3.35) (with, of course, $\check{f}\to Wf, \check{g}\to Wg$) and note that the factor $\epsilon_i E_{k_i}$ [cf. (4.30)] amounts to a distributional time derivative of $\hat{\Delta}_R(t_i - t_{i+1}, \vec{k}_i)$ in those terms where it occurs once. Choose then uniformly bounded sequences $\hat{F}_N, \hat{G}_N \in S(R^4)$ such that $\lim_{N\to\infty}\hat{F}_N, \hat{G}_N = \hat{\Delta}_R, \hat{\Delta}_R$ pointwise and proceed as in Sec. 3. Finally, replace $E_{k_i}^2$ by $m^2 - k_i^2 + k_i^{02}$ in those terms where $\epsilon_i E_{k_i}$ occurred twice and use (3.18).

Since I is analytic in the coupling constant, we can rearrange (4.45):

$$(I(\infty,-\infty)f)_\epsilon(\vec{p})$$

$$= 2\pi i\epsilon\sum_{n=1}^{\infty}\sum_{\epsilon'}\int dk_1\cdots dk_{n-1}d\vec{q}(2E_p)^{-1/2}\times[\tilde{V}_2(\epsilon p,k_1)\tilde{\Delta}_I(k_1)\tilde{V}_2(k_1,k_2)\cdots\tilde{\Delta}_I(k_{n-1})\tilde{V}_2(k_{n-1},\epsilon' q) + \text{all } \tilde{A}_\mu\tilde{A}^\mu\text{-contractions}](2E_q)^{-1/2}f_{\epsilon'}(\vec{q}), \quad (4.47)$$

where

$$\tilde{V}_2(k,k') \equiv \tilde{A}_\mu(k-k')(k^\mu + k'^\mu) + \tilde{A}_4(k-k'). \quad (4.48)$$

We can write (4.47) as

$$(I(\infty,-\infty)f)_\epsilon(\vec{p}) = \sum_{n=1}^{\infty}\sum_{\epsilon'}\int d\vec{q}\, I^{(n)}_{\epsilon\epsilon'}(\vec{p},\vec{q})f_{\epsilon'}(\vec{q}) \quad (4.49)$$

and observe that $I^{(n)}_{\epsilon\epsilon'}(\vec{p},\vec{q})$ is in $S(R^3)$ in $\vec{p}$ and $\vec{q}$ separately. Furthermore,

$$I^{(n)}_{\epsilon\epsilon'}(\vec{p},\vec{q}) = 2\pi i\epsilon\lim_{\delta_{n-1}\downarrow 0}\cdots\lim_{\delta_1\downarrow 0}\int dk_1\cdots dk_{n-1}(2E_p)^{-1/2}\times[\tilde{V}_2(\epsilon p,k_1)\Delta_I^{\delta_{n-1}}(k_1)\tilde{V}_2(k_1,k_2)\cdots\Delta_I^{\delta_1}(k_{n-1})\tilde{V}_2(k_{n-1},\epsilon' q) + \text{all } \tilde{A}_\mu\tilde{A}^\mu\text{-contractions}](2E_q)^{-1/2}. \quad (4.50)$$

Formally rearranging (4.45), we can also write

$$(I(\infty,-\infty)f)_\epsilon(\vec{p}) = 2\pi i\epsilon\sum_{n=1}^{\infty}\sum_{\epsilon'}\int dk_1\cdots dk_{n-1}d\vec{q}(2E_p)^{-1/2}\times\tilde{V}_3(\epsilon p,k_1)\tilde{\Delta}_I(k_1)\tilde{V}_3(k_1,k_2)\cdots\tilde{\Delta}_I(k_{n-1})\tilde{V}_3(k_{n-1},\epsilon' q)\times(2E_q)^{-1/2}f_{\epsilon'}(\vec{q}), \quad (4.51)$$

where

$$\tilde{V}_3(k,k') \equiv \tilde{A}_\mu(k-k')(k^\mu + k'^\mu) + \widetilde{A_\mu A^\mu}(k-k') + \tilde{A}_4(k-k'). \tag{4.52}$$

In fact, this rearrangement is allowed if $I = R, A$. Indeed, defining

$$\check{\tilde{O}}(t) = \exp(i\check{H}_0 t)\begin{pmatrix} 0 & 0 \\ A_0^2(t,\cdot) & 0 \end{pmatrix}\exp(-i\check{H}_0 t), \tag{4.53}$$

it follows from Theorem 2.6 that

$$R(\infty,-\infty) = \sum_{n=1}^{\infty} \tilde{R}_1^{(n)}(\infty,-\infty). \tag{4.54}$$

However, it can be seen that $(\tilde{R}_1^{(n)}(\infty,-\infty)f)_\epsilon(\vec{p})$ is equal to the nth term in (4.51). Thus, (4.51) holds for $I = R$, and, similarly, for $I = A$.

[Provided the coupling constant is small enough, the rearrangement is also permitted if $I = F, \bar{F}$. This is a consequence of work of Bellissard.[3,4] It follows from his results (in particular Ref. 3, Lemmas A 5.4, A 5.6) and (Ref. 4, Theorem II 1.1) that the rhs of (4.51) defines a bounded operator on H if the coupling constant is small enough. Since, moreover, this operator is analytic in a neighborhood of the origin, we can rearrange (4.51) and obtain (4.47).]

C. Lorentz covariance, causality

The representation of the Poincaré group in H is defined by

$$(U(a,\Lambda)g)_\epsilon(\vec{p}) = \exp(i\epsilon pa)\left(\frac{(\Lambda^{-1}p)_0}{p_0}\right)^{1/2} g_\epsilon(\overrightarrow{\Lambda^{-1}p}). \tag{4.55}$$

It easily follows from (4.47) and (4.55) that

$$U(a,\Lambda)I(A_I)U^*(a,\Lambda) = I(A_I^{a,\Lambda}), \tag{4.56}$$

where $A_\mu^{a,\Lambda}$ is defined by (3.43), and

$$A_4^{a,\Lambda}(x) \equiv A_4(\Lambda^{-1}(x-a)). \tag{4.57}$$

Thus, the operators $I(\infty,-\infty)$ are Lorentz covariant. Defining the classical S operator by (3.54) we have

Theorem 4.1: $S(V)$ is Lorentz covariant and causal.

Proof: This follows as in Sec. 3. ■

Defining

$$\underset{A}{G_R}(T_2,T_1) = \pm iU^s(T_2,T_1)\theta(\pm(T_2-T_1)), \tag{4.58}$$

it follows as in Sec. 3 from (4.21) that (3.61) and (3.63) hold true. Hence, using (4.25), we have

$$\vec{D}_x[G_I](x,y) = [G_I](x,y)\overleftarrow{D}_y = \delta(x-y), \quad I = R, A, \tag{4.59}$$

where

$$D_x \equiv D_x^0 - B(x). \tag{4.60}$$

Thus, $[G_{R,A}]$ are retarded resp. advanced fundamental solutions of the perturbed two-component Klein–Gordon equation (4.28). It should be noted that their existence does not follow from Theorem A.1 if the electromagnetic field is nonzero, since in this case second order derivatives occur in $B * S_R$.

As in Sec. 3 we did not use the assumption that the fields are real-valued in obtaining (3.61) and (3.63). Consequently, these relations also hold if the fields are only assumed to be in $S(\mathbf{R}^4)$. Clearly, (4.59) holds as well if one does not complex conjugate the fields in the expression $\overleftarrow{D}_y$. Thus, we have proved:

Theorem 4.2: For any $A_l \in S(\mathbf{R}^4)$, $l = 0,\ldots,4$, the perturbed two-component Klein–Gordon equation (4.28) admits two-sided tempered retarded and advanced fundamental solutions given by (3.61). They are connected with the evolution operator through (4.58) and (4.21), and are Wightman regular if the electromagnetic field vanishes. ■

Setting, for $F, G, A_0, \ldots, A_4 \in S(\mathbf{R}^4)$,

$$\overset{\circ}{G}_I(F,G) \equiv \sum_{n=0}^{\infty} \int dx\, dx_1 \cdots dx_n\, dy\, F(x)\Delta_I(x-x_1)K(x_1)\Delta_I(x_1-x_2) \times \cdots K(x_n)\Delta_I(x_n-y)G(y), \quad I = R, A, \tag{4.61}$$

where

$$K \equiv iA_\mu\partial^\mu + i\partial_\mu A^\mu + A_\mu A^\mu + A_4, \tag{4.62}$$

it is clear that $\overset{\circ}{G}_{R,A}$ are, formally, two-sided retarded resp. advanced fundamental solutions of the perturbed Klein–Gordon equation (4.27):

$$[\Box_x + m^2 - \vec{K}(x)]\,\overset{\circ}{G}_I(x,y) = \overset{\circ}{G}_I(x,y)(\overleftarrow{\Box}_y + m^2 - \overleftarrow{K}(y)) = \delta(x-y). \tag{4.63}$$

(Again, the complex conjugation in $\overleftarrow{K}$ is supposed not to act on the fields.) In fact, it easily follows from Theorem A.1 that $\overset{\circ}{G}_{R,A}$ are Wightman regular distributions in $S'(\mathbf{R}^8)$.

Defining now

$$[G_I] = \begin{pmatrix} G_I^{11} & G_I^{12} \\ G_I^{21} & G_I^{22} \end{pmatrix}, \tag{4.64}$$

it follows from (4.59) that $G_{R,A}^{12}$ are retarded resp. advanced fundamental solutions of (4.27). By using (4.23) and (4.44), it, moreover, follows from (3.61) that, if $A_0(x) = 0$,

$$G_I^{12} = \overset{\circ}{G}_I. \tag{4.65}$$

Actually, (4.65) holds as well if $A_0(x) \neq 0$, which can be seen if one uses Theorem 2.6. Therefore, we have proved:

Theorem 4.3: For any $A_l \in S(\mathbf{R}^4)$, $l = 0,\ldots,4$, the perturbed Klein–Gordon equation (4.27) admits two-sided tempered retarded and advanced fundamental solutions, given by (4.61). They are connected with the evolution operator through (4.65) and are Wightman regular. ■

D. Implementability of the evolution in $\mathcal{J}_s(H)$

In this subsection we again assume that the fields are real-valued since we need the pseudo-unitarity of $U(T_2,T_1)$. We first show that the hypothesis of Theorem 2.8 is satisfied if the magnetic field vanishes.

Theorem 4.4: Let

$$O_k(t) \equiv \exp(iH_0 t) V_k(t) \exp(-iH_0 t), \quad k = e, s. \tag{4.66}$$

Then the operators $\int_{T_1}^{T_2} dt\, O_k(t)_{\pm\mp}$, $k = e, s$, are $\|\cdot\|_2$-continuous on $\tilde{R}^2$ and

$$\left\| \int_{T_1}^{T_2} dt\, O_k(t)_{\pm\mp} \right\|_2 \leq C_1 < \infty \quad \forall (T_2, T_1) \in \tilde{R}^2. \tag{4.67}$$

Proof: It follows from (4.30) that it suffices to prove these statements for the integral operators on $L^2(R^3)^2$ with the kernels

$$\int_{T_1}^{T_2} dt \exp[\pm i(E_p + E_q)t] \hat{A}_0(t, \vec{p} - \vec{q})(E_p - E_q)(E_p E_q)^{-1/2}, \tag{4.68}$$

$$\int_{T_1}^{T_2} dt \exp[\pm i(E_p + E_q)t] \hat{A}_4(t, \vec{p} - \vec{q})(E_p E_q)^{-1/2}. \tag{4.69}$$

Using (3.73), we conclude that it is sufficient to show that

$$\int d\vec{p}\, d\vec{q}\, (E_p - E_q)^2 (E_p E_q)^{-1} (E_p + E_q)^{-2} (1 + |\vec{p} - \vec{q}|^2)^{-4} < \infty, \tag{4.70}$$

$$\int d\vec{p}\, d\vec{q}\, (E_p E_q)^{-1} (E_p + E_q)^{-2} (1 + |\vec{p} - \vec{q}|^2)^{-4} < \infty. \tag{4.71}$$

However, (4.70) and (4.71) are easily seen to hold true (cf. Ref. 7). ■

Corollary 4.5: The evolution operator $U(T_2, T_1)$, corresponding to an electric or scalar field in $S(R^4)$, or to both together, gives rise to a unitarily implementable Bogoliubov transformation in $\mathcal{J}_s(\mathcal{H})$ for any $(T_2, T_1) \in \tilde{R}^2$.

Proof: This follows as in Sec. 3. ■

We now assume that the fields are real-valued functions in $S(R^3)$. Defining

$$O^k(t) = \exp(iH_0 t) V_k \exp(-iH_0 t), \quad k = e, m, s, \tag{4.72}$$

one obtains

$$\left\| \int_{T_1}^{T_2} dt\, O^k(t)_{\pm\mp} \right\|_2^2 = \int d\vec{p}\, d\vec{q}\, |F_k(\vec{p}, \vec{q})|^2 \sin^2[\alpha(E_p + E_q)]\, (E_p E_q)^{-1} (E_p + E_q)^{-2}, \tag{4.73}$$

where

$$F_e(\vec{p}, \vec{q}) \equiv \hat{A}_0(\vec{p} - \vec{q})(E_p - E_q), \tag{4.74}$$

$$F_m(\vec{p}, \vec{q}) \equiv \hat{\vec{A}}(\vec{p} - \vec{q}) \cdot (\vec{p} + \vec{q}) - \widehat{A^2}(\vec{p} - \vec{q}), \tag{4.75}$$

$$F_s(\vec{p}, \vec{q}) \equiv \hat{A}_4(\vec{p} - \vec{q}). \tag{4.76}$$

Defining

$$H = H_0 - \sum_{k=e,m,s} V_k \tag{4.77}$$

it follows as in Sec. 3 that $\exp(-iHt)$ is implementable in $\mathcal{J}_s(\mathcal{H})$ for any $t \in R$ if $V_m = 0$. However, if a magnetic field is present, the situation is different:

Theorem 4.6: Let $(T_2, T_1) \in R^2$ and $T_2 \neq T_1$. Then the operators $\int_{T_1}^{T_2} dt\, O^m(t)_{\pm\mp}$ are not H.S. if $V_m \neq 0$.

Proof: It suffices to show that the integral on the rhs of (4.73) diverges if $F_k(\vec{p}, \vec{q})$ is replaced by $\hat{\vec{A}}(\vec{p} - \vec{q}) \cdot (\vec{p} + \vec{q})$, and if $\alpha \vec{A} \neq 0$. We introduce variables s, θ, ϕ, and $\vec{t}$ as in the proof of Theorem 3.5 and assume the integral converges. It then follows from Fubini's theorem that there exist a $\vec{t}_0$ and θ_0, ϕ_0 with $\vec{n}_{\theta_0, \phi_0} \cdot \vec{A}(\vec{t}_0) \neq 0$ such that the integral over s converges ($\vec{n}_{\theta,\phi}$ is the unit vector in the direction θ, ϕ). However, the latter integral is a nonzero multiple of (3.85) so that it actually diverges. This contradiction completes the proof. ■

Corollary 4.7: Assume that $\exp(-iHt)$ is implementable in $\mathcal{J}_s(\mathcal{H})$ for any $t \in R$. Then the magnetic field vanishes.

Proof: This follows as in Sec. 3. ■

This corollary was first proved by Hochstenbach[11] using results of Shale.[21] Again, we expect that an analogous result holds for time-dependent fields.

APPENDIX

We consider two transformations T_R and Γ_A of $S(R^4)^K$. Let $V_l^{\sigma\sigma'}(x)$ $(\sigma, \sigma' = 1, \ldots, K,\ l = 0, \ldots, 4)$ be functions in $S(R^4)$ and let

$$A_{\sigma\sigma'}(x) \equiv V_0^{\sigma\sigma'}(x)\partial_t + \sum_{j=1}^{3} V_j^{\sigma\sigma'}(x)\partial_{x_j} + V_4^{\sigma\sigma'}(x). \tag{A1}$$

For any $F \in S(R^4)^K$ we define

$$T_{\substack{A \\ R}} F = \left(1 - J_{\substack{R \\ A}}\right) F, \tag{A2}$$

where

$$(J_I F)_\sigma(x) \equiv A_{\sigma\sigma'}(x) \int dy \Delta_I(x - y) F_{\sigma'}(y), \quad I = R, A. \tag{A3}$$

In (A3), as in the sequel, the summation convention is used. The following theorem is a generalization of a result of Bellissard.[4]

Theorem A.1: The transformations T_R and T_A are linear bicontinuous bijections of $S(R^4)^K$ onto $S(R^4)^K$ with inverses

$$T_{\substack{R \\ A}}^{-1} = \sum_{N=0}^{\infty} J_{\substack{A \\ R}}^N. \tag{A4}$$

Proof: It is easily seen that T_R and Γ_A are linear continuous injections. Since $S(R^4)^K$ is a Fréchet space, they are bicontinuous if they are also surjections.[12] Thus, in view of (A2), it suffices to prove that

$$\lim_{M \to \infty} \sum_{N=0}^{M} J_I^N F$$

exists in $S(R^4)^K$. We will show this for $I = R$. The proof for $I = A$ is similar. We first note that [cf. (3.16)]

$$\begin{aligned} &\hat{\Delta}_R(t, \vec{p}) = \theta(t) \sin(E_p t)/E_p, \\ &\widehat{(\partial_t \Delta_R)}(t, \vec{p}) = \theta(t) \cos E_p t, \\ &\widehat{(\partial_{x_j} \Delta_R)}(t, \vec{p}) = \theta(t) i p_j \sin(E_p t)/E_p, \quad j = 1, 2, 3. \end{aligned} \tag{A5}$$

We now define for any $F \in S(R^4)^K$

$$\|F\|_{\tau,\alpha,\sigma} = \sup_{(t,\vec{p})\in R^4} |(1+t^2+p^2)^\tau D^\alpha \hat{F}_\sigma(t,\vec{p})|\,, \tag{A6}$$

where $\alpha \equiv (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ is a multi-index, $\tau = 0, 1, \ldots,$ and $\sigma = 1, \ldots, K$. Clearly,

$$\lim_{M\to\infty} \sum_{N=0}^{M} J_R^N F$$

exists if the sequence $J_R^N F$ is absolutely summable in all $\|\cdot\|_{\tau,\alpha,\sigma}$. In order to show this, we shall majorize the function

$$H_{N\tau\alpha\sigma}(t,\vec{p}) \equiv (1+t^2+p^2)^\tau D^\alpha \widehat{(J_R^N F)}_\sigma(t,\vec{p})\,. \tag{A7}$$

It is easy to see that

$$\begin{aligned}
&\widehat{(J_R^N F)}_\sigma(t,\vec{p}) \\
&= \int_{-\infty}^{t} dt_1 \cdots \int_{-\infty}^{t_{N-1}} dt_N \int d\vec{q}_1 \cdots d\vec{q}_N \hat{V}_{t_1}^{\sigma\sigma_1}(t,\vec{p}-\vec{q}_1) \\
&\quad \times G_{t_1}(t-t_1,\vec{q}_1)\cdots \hat{V}_{t_N}^{\sigma_{N-1}\sigma_N}(t_{N-1},\vec{q}_{N-1}-\vec{q}_N) \\
&\quad \times G_{t_N}(t_{N-1}-t_N,\vec{q}_N)\hat{F}_{\sigma_N}(t_N,\vec{q}_N)\,,
\end{aligned} \tag{A8}$$

where

$$\begin{aligned}
&G_0(t,\vec{p}) \equiv \cos E_p t\,, \\
&G_j(t,\vec{p}) \equiv i p_j \sin(E_p t)/E_p\,, \quad j=1,2,3\,, \\
&G_4(t,\vec{p}) \equiv \sin(E_p t)/E_p\,.
\end{aligned} \tag{A9}$$

Obviously, for any $l = 0, \ldots, 4$, $(t,\vec{p}) \in R^4$, $\beta = 1, \ldots, \alpha_0$:

$$|G_l(t,\vec{p})| \le \delta \tag{A10}$$

and

$$|\partial_t^\beta G_l(t,\vec{p})| \le \delta^{\alpha_0} E_p^{2\alpha_0}\,, \tag{A11}$$

where

$$\delta \equiv \max(m^{-2}, 1)\,. \tag{A12}$$

Assuming from now on that $N > \alpha_0$, we conclude that $H_{N\tau\alpha\sigma}$ is equal to a sum of terms the number of which is bounded by $(5K)^N L$, where L only depends on α_0. Denoting the generic term by H_i, we have, suppressing the indices,

$$\begin{aligned}
&H_i(t,\vec{p}) \\
&= \int_{-\infty}^{t} dt_n \cdots \int_{-\infty}^{t_{N-1}} dt_N \int d\vec{q}_1 \cdots d\vec{q}_N (1+t^2+p^2)^\tau \\
&\quad \times \hat{V}_{\beta_1,\vec{\alpha}}(t,\vec{p}-\vec{q}_1)(\partial_t^{\beta_2} G)(0,\vec{q}_1)\cdots(\partial_t^{\beta_{2n-1}}\hat{V})(t,\vec{q}_{n-1}-\vec{q}_n) \\
&\quad \times (\partial_t^{\beta_{2n}} G)(t-t_n,\vec{q}_n)\hat{V}(t_n,\vec{q}_n-\vec{q}_{n+1})\cdots \hat{F}(t_N,\vec{q}_N)\,,
\end{aligned} \tag{A13}$$

where

$$\hat{V}_{\beta_1,\vec{\alpha}}(t,\vec{k}) \equiv (\partial_t^{\beta_1}\partial_{k_1}^{\alpha_1}\partial_{k_2}^{\alpha_2}\partial_{k_3}^{\alpha_3}\hat{V})(t,\vec{k}) \tag{A14}$$

and

$$0 \le n-1, \beta_1, \ldots, \beta_{2n} \le \alpha_0\,. \tag{A15}$$

Thus, for any H_i with a fixed n,

$$\begin{aligned}
|H_i(t,\vec{p})| &\le \delta^{\alpha_0^2+\alpha_0+N} \int_{-\infty}^{\infty} dt_n \cdots \int_{-\infty}^{t_{N-1}} dt_N \\
&\quad \times \int d\vec{k}_1 \cdots d\vec{k}_N (1+t^2)^\tau (1+p^2)^\tau \\
&\quad \times \left|\hat{V}_{\beta_1,\vec{\alpha}}\left(t,\vec{p}-\sum_{j=1}^{N}\vec{k}_j\right)\right| \left(\left|\sum_{j=1}^{N}\vec{k}_j\right|^2+m^2\right)^{\alpha_0} \\
&\quad \times \cdots |(\partial_t^{\beta_{2n-1}}\hat{V})(t,\vec{k}_{n-1})| \left(\left|\sum_{j=n}^{N}\vec{k}_j\right|^2+m^2\right)^{\alpha_0} \\
&\quad \times \prod_{j=n}^{N-1} |\hat{V}(t_j,\vec{k}_j)|\,|\hat{F}(t_N,\vec{k}_N)|\,.
\end{aligned} \tag{A16}$$

Using the estimate

$$\left|\sum_{j=r}^{N}\vec{k}_j\right|^2 + m^2 \le \gamma N \prod_{j=r}^{N}(1+k_j^2), \quad 1 \le r \le N\,, \tag{A17}$$

$$\gamma \equiv \max(m^2, 1)\,,$$

we conclude that

$$\begin{aligned}
&|H_i(t,\vec{p})| \\
&\le (\delta\gamma N)^{\alpha_0^2+\alpha_0}\delta^N \int_{-\infty}^{\infty} dt_1 \cdots \int_{-\infty}^{t_{N-1}} dt_N \\
&\quad \times \int d\vec{k}_1 \cdots d\vec{k}_N (1+t^2)^\tau (1+p^2)^\tau \\
&\quad \times \left|\hat{V}_{\beta_1,\vec{\alpha}}\left(t,\vec{p}-\sum_{j=1}^{N}\vec{k}_j\right)\right| \prod_{j=1}^{n-1}(1+k_j^2)^{\alpha_0^2+\alpha_0} |(\partial_t^{\beta_{2j+1}}\hat{V})(t,\vec{k}_j)| \\
&\quad \times \prod_{j=n}^{N-1}(1+k_j^2)^{\alpha_0^2+\alpha_0}|\hat{V}(t_j,\vec{k}_j)|\,(1+k_N^2)^{\alpha_0^2+\alpha_0}|\hat{F}(t_N,\vec{k}_N)|\,.
\end{aligned} \tag{A18}$$

Since $(\hat{V}_l^{\sigma\sigma'})_{\beta,\vec{\alpha}}$, $\partial_t^\beta \hat{V}_l^{\sigma\sigma'}$, $\hat{F}_\sigma$ are test functions ($\beta = 0, \ldots, \alpha_0$, $l = 0, \ldots, 4$, $\sigma, \sigma' = 1, \ldots, K$) there exists a $C > 0$, only depending on V, F, α, τ, such that the sum of all H_i with the same n, denoted by S_n, satisfies

$$\begin{aligned}
&|S_n(t,\vec{p})| \le (5K)^N L(\delta\gamma N)^{\alpha_0^2+\alpha_0}\delta^N C^{N+1}\pi^{N-n+1}[(N-n+1)!\,]^{-1} \\
&\quad \times \int d\vec{k}_1 \cdots d\vec{k}_N (1+p^2)^\tau \\
&\quad \times \left(1+\left|\vec{p}-\sum_{j=1}^{N}\vec{k}_j\right|^2\right)^{-\tau}\prod_{j=1}^{N}(1+k_j^2)^{-\tau-2}\,.
\end{aligned} \tag{A19}$$

Using the estimate

$$(1+p^2)(1+q^2)^{-1} \le 2(1+|\vec{p}-\vec{q}|^2) \tag{A20}$$

and (A17), we obtain

$$|S_n(t,\vec{p})| \le C_1 C_2^N N^{\alpha_0^2+\alpha_0+\tau}/(N-n+1)!\,, \tag{A21}$$

where C_1, C_2 do not depend on N and n. Thus,

$$|H_{N\tau\alpha\sigma}(t,\vec{p})| \le (\alpha_0+1)C_1 C_2^N N^{\alpha_0^2+\alpha_0+\tau}/(N-\alpha_0)!\,. \tag{A22}$$

It easily follows from (A22) that the sequence $J_R^N F$ is absolutely summable in all $\|\cdot\|_{\tau,\alpha,\sigma}$, and so the theorem is proved. ■

It can be seen that the theorem still holds true if R^4 is replaced by R^{1+n}, with $n \ge 1$, and Δ_R, Δ_A by the (unique) retarded and advanced fundamental solutions of the partial differential operator $\partial_t^2 - \sum_{j=1}^{n}\partial_{x_j}^2 + m^2$ [the retarded one satisfies (A5) with $j = 1, \ldots, n$]. We note further that we have made essential use of the boundedness of the G_l,

i.e., of the fact that the derivatives of Δ_R in T_A are only first order. The theorem can therefore in general not be applied to higher spin external field theories (see, however, Ref. 4). This is in agreement with the nonexistence of retarded solutions in most of these theories (cf. Ref. 19).

ACKNOWLEDGMENT

This investigation is part of the research program of the "Stichting voor Fundamenteel Onderzoek der Materie (F.O.M.)," which is financially supported by the "Organisatie voor Zuiver Wetenschappelijk Onderzoek (Z.W.O.)."

[1]S. N. M. Ruijsenaars, "On Bogoliubov transformations for systems of relativistic charged particles," to be published in J. Math. Phys.
[2]S. N. M. Ruijsenaars, "Charged particles in external fields. II. The quantized Dirac and Klein–Gordon theories," to be published in Commun. Math. Phys.
[3]J. Bellissard, Commun. Math. Phys. **41**, 235–66 (1975).
[4]J. Bellissard, "Quantized fields in external field. II. Existence theorems," Marseille preprint, 1974.
[5]H. E. Moses, Phys. Rev. **95**, 237–42 (1954).
[6]J. M. Chadam, J. Math. Phys. **9**, 386–96 (1968).
[7]P. J. M. Bongaarts, Ann. Phys. (N.Y.) **56**, 108–39 (1970).
[8]B. Schroer, R. Seiler, and J. A. Swieca, Phys. Rev. D **2**, 2927–37 (1970).
[9]R. Seiler, Commun. Math. Phys. **25**, 127–51 (1972).
[10]L.-E. Lundberg, Commun. Math. Phys. **31**, 295–316 (1973).
[11]W. Hochstenbach, thesis (Hamburg, 1975).
[12]M. Reed and B. Simon, *Functional Analysis* (Academic, New York, 1972).
[13]M. Reed and B. Simon, *Fourier Analysis, Self-Adjointness* (Academic, New York, 1975).
[14]R. S. Phillips, Trans. Amer. Math. Soc. **74**, 199–221 (1953).
[15]S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Row, Peterson, Evanston, Ill., 1961).
[16]J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics* (McGraw-Hill, New York, 1964).
[17]L. Schwartz, *Théorie des distributions. II* (Hermann, Paris, 1951).
[18]R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (Benjamin, New York, 1964).
[19]A. S. Wightman, "Relativistic wave equations as singular hyperbolic systems," in *Partial Differential Equations*, edited by D. Spencer (Symp. Pure Math. **23**, pp. 441–77) (Amer. Math. Soc., Providence, R. I., 1973).
[20]L.-E. Lundberg, Commun. Math. Phys. **31**, 243–57 (1973).
[21]D. Shale, Trans. Amer. Math. Soc. **103**, 149–67 (1962).

Collineations and motions in self-gravitating magnetofluids

G. G. Asgekar

Department of Mathematics, A. S. C. College, Shahunagar-416 211, India

T. H. Date

Department of Mathematics, Shivaji University, Kolhapur 416 004, India
(Received 31 October 1975; revised manuscript received 3 May 1976)

For the self-gravitating magnetofluid admitting a Ricci collineation with respect to the flow vector, it is proved that (i) the stream lines are expansion-free if and only if the heat-flux vector is divergence-free; (ii) the stream lines are geodesic if and only if the heat-flux vector remains invariant along the system of stream lines. The affine collineation with respect to the geodesic and expansion-free flow of the magnetofluid is a motion. Conformal motions with respect to the flow vector, the magnetic field vector, and the heat flux vector are investigated.

1. INTRODUCTION

The quest for geometrical symmetries like motions and collineations is motivated by the necessity of discovering conservation laws in the theory of relativity. The conservation laws of energy, linear momentum, and angular momentum are easily formulated in the special theory of relativity.[1] The difficulties encountered in discovering such conserved quantities in general relativity are well known.[2,3] This has prompted research workers in this field to find special geometrical symmetries which generate "weak" conservation laws. The successes in this direction are enumerated in a picturesque way by Davis.[4] An important symmetry is that due to the vanishing of the Lie derivative of the curvature tensor which is popularly known as curvature collineation (CC).[5-9] The existence of proper CC has been demonstrated by Aichelberg[10] for *PP*-waves whereas Collinson[11] has shown that for the empty space–time not of Petrov type *N*, the only CC admitted are the conformal motions. Collineations in nonempty space–times have been considered by Collinson,[12] who has obtained the conservation law $(R_a{}^b\,\xi^a)_{;b}=0$, when ${}_\xi\pounds\, R_{ab}=0$ and the law $(Mf(\xi_a p^a)p^b)_{;b}=0$, when ${}_\xi\pounds\, F_a{}^b=0$, where f is an arbitrary function, M is a scalar with $p^a p_a=0$, R_{ab} is the Ricci tensor, and $F_a{}^b$ is the electromagnetic field tensor. An elegant account of groups of motions and collineations in the space–time filled with a perfect fluid allowing infinite electrical conductivity and constant magnetic permeability (perfect magnetofluid) is given by Khade.[13] Following imperfect fluid schemes initiated by Lichnerowicz,[14] Shaha[15] has developed a definite magnetofluid scheme characterized by a stress–energy tensor with timelike eigenvalue as the energy density and the three spacelike eigenvalues as the partial pressures. For such schemes he has obtained a set of necessary and sufficient conditions for the Ricci collineation with respect to (w.r.t.) the world line to imply the motion. An "Einstein collineation," viz. ${}_\xi\pounds\,(R_{ab}-\frac{1}{2}Rg_{ab})=0$, in the space–time of the perfect magnetofluid w.r.t. the flow vector or the magnetic field vector has been investigated by the authors.[16] Optical collineations and Einstein collineations w.r.t. propagation or polarization vector in a self-gravitating null electromagnetic field have been examined by Khade and Radhakrishna.[17]

The object of this paper is to study the weak conservation laws based upon the existence of collineations and motions in a self-gravitating universe filled with magnetofluid. The distribution of matter considered in this paper is more general than Ellis,[18] Greenberg,[19] and Kato *et al.*[20] All of them have studied a thermally conducting, viscous, compressible fluid. Date[21] has extended their works to fluids with infinite electrical conductivity and constant magnetic permeability. He[22] has calculated the expressions for the refractive index and the ray shear of a null gravitational field due to the magnetofluid. It is proved by the authors[23] that for an essentially expanding flow of the magnetofluid the gravitational tidal force is due to the magnetic field only.

Throughout the investigation, arbitrary coordinates x^a are used to describe the four-dimensional Riemannian space–time with the metric g_{ab} of signature -2. Latin indices $a, b, c, \cdots$ range and sum over 1–4.

The field equations, the kinematical parameters, the thermodynamical relations, and the equations of motion are delineated in Sec. 2. The consequences of the Ricci collineation w.r.t. the flow vector or the magnetic field vector are discussed in Sec. 3, whereas an affine collineation w.r.t. the flow vector or the magnetic field vector is considered in Sec. 4. The next section deals with the conformal motion w.r.t. the magnetic field vector, the flow vector, and heat-flux vector.

2. MAGNETOHYDRODYNAMIC FIELD EQUATIONS

Magnetohydrodynamics is the study of the properties of a fluid with infinite electrical conductivity. We consider a viscous, compressible, thermally conducting and self-gravitating fluid with infinite electrical conductivity and constant magnetic permeability. We call it a magnetofluid and the corresponding stress–energy tensor is chosen[22] as

$$T_{ab}=(\rho+p+\mu\,|\bar{h}|^2)\,u_a u_b-(p+\tfrac{1}{2}\,\mu\,|\bar{h}|^2)\,g_{ab}+\nu\sigma_{ab}+u_a q_b+u_b q_a-\mu h_a h_b\,, \tag{2.1}$$

where ρ is the matter energy density, p is the isotropic pressure, $\nu(\geq 0)$ is the coefficient of shear viscosity, u^a is the 4-velocity, h^a is the magnetic field vector, and q^a is the heat-flux vector satisfying the relations

$$u^a u_a=1,\quad u^a q_a=0,\quad u^a h_a=0\,,$$
$$-h^a h_a=|\bar{h}|^2,\quad -q^a q_a=|\bar{q}|^2\,.$$

 Copyright © 1977 American Institute of Physics

Here the bulk viscosity has been ignored since it is negligible.[18]

The kinematical parameters associated with the time-like congruence u^a are the expansion $\theta = u^a{}_{;a}$, the shear tensor $\sigma_{ab} = u_{(a;b)} - \dot{u}_{(a} u_{b)} - \frac{1}{3}\theta P_{ab}$, and the rotation tensor $\omega_{ab} = u_{[a;b]} - \dot{u}_{[a} u_{b]}$, where $P_{ab} = g_{ab} - u_a u_b$ is the space-projection operator and $\dot{u}^a = u^a{}_{;b} u^b$ is the 4-acceleration. We define $2\sigma^2 = \sigma^{ab}\sigma_{ab}$ and $2\omega^2 = \omega^{ab}\omega_{ab}$. Here and in what follows, the semicolon indicates covariant differentiation, the round brackets, and the square brackets around the suffixes denote symmetrization and anti-symmetrization, respectively.

The expression for the heat-flux vector is given by[24]

$$q^a = -K(T_{0;b} - T_0 \dot{u}_b) P^{ab}, \tag{2.2}$$

where K is the coefficient of heat conduction and T_0 is the rest temperature. The expression (2.2) is the generalization of the classical law of heat conduction and incorporates the inertia of flowing heat.

The field equations[25] of relativistic magnetohydrodynamics are the Einstein's equations

$$R_{ab} - \tfrac{1}{2} R g_{ab} = -T_{ab}\ , \tag{2.3}$$

and the Maxwell equations

$$(u^a h^b - u^b h^a)_{;b} = 0\ . \tag{2.4}$$

The relations connecting the thermodynamical variables in the analysis are[25]

$$T_0\, dS = d\epsilon + p\, d(1/\rho_0)\ , \tag{2.5}$$

$$\rho = \rho_0(1+\epsilon)\ . \tag{2.6}$$

Here ρ_0 is the matter density, ϵ is the internal energy density, and S is the specific entropy.

The identities $(R^{ab} - \frac{1}{2} R g^{ab})_{;b} = 0$ for the fluid under investigation yield the local "energy balance" equation in the form

$$\begin{aligned}
&(\rho + p + \mu|\bar{h}|^2)_{,c}\, u^c u^a \\
&\quad + (\rho + p + \mu|\bar{h}|^2)(\dot{u}^a + \theta u^a) - (p + \tfrac{1}{2}\mu|\bar{h}|^2)_{,b} g^{ab} \\
&\quad + \nu\sigma^{ab}{}_{;b} + \tfrac{4}{3}\theta q^a + q^b{}_{;b} u^a + \dot{q}^a + (\sigma^{ab} + \omega^{ab}) q_b \\
&\quad - \mu|\bar{h}|^2{}_{,b} n^b n^a - \mu|\bar{h}|^2 (n^a{}_{;b} n^b + n^b{}_{;b} n^a) = 0\ ,
\end{aligned} \tag{2.7}$$

where $n^a = h^a/|\bar{h}|$, $(n^a n_a = -1,\ n^a u_a = 0)$. On transvecting (2.7) with u_a and P_{ac} and employing the Maxwell equations (2.4) in the resulting expressions, we have the equations of motion:

$$\dot{\rho} + (\rho + p)\theta - 2\nu\sigma^2 + q^a{}_{;a} - q^a \dot{u}_a = 0\ , \tag{2.8}$$

$$\begin{aligned}
&(\rho + p + \mu|\bar{h}|^2)\dot{u}_c - (p + \tfrac{1}{2}\mu|\bar{h}|^2)_{,b} P^b{}_c + \nu\sigma^b{}_{c;b} + 2\nu\sigma^2 u_c \\
&\quad + \tfrac{4}{3}\theta q_c + q^a{}_{;b} u^b P_{ac} + (\sigma_{cb} + \omega_{cb}) q^b - \mu|\bar{h}|^2{}_{,b} n^b n_c \\
&\quad - \mu|\bar{h}|^2 (n_{c;b} n^b - n_{a;b} u^b u^a u_c + n^b{}_{;b} n_c) = 0\ .
\end{aligned} \tag{2.9}$$

Equation (2.8) is identified as the equation of continuity for the magnetofluid. Note that the magnetic field is not explicitly present in this equation.

From Eqs. (2.5), (2.6), and (2.7), we obtain the heat-transfer equation in the magnetofluid as

$$\chi(\rho_0 u^a)_{;a} + T_0 \rho_0 S_{,a} u^a = \dot{u}^a q_a - q^a{}_{;a} + 2\nu\sigma^2\ , \tag{2.10}$$

where $\chi = 1 + \epsilon + p/\rho_0$ is the generalized density.

3. RICCI COLLINEATIONS

This section deals with the space-time of the magnetofluid admitting the Ricci collineation w.r.t. the flow vector or the magnetic field vector.

Theorem 1: For a magnetofluid admitting a Ricci collineation w.r.t. the flow vector, (i) the stream lines are expansion-free if and only if the heat-flux vector is divergence-free, (ii) the stream lines are geodesic if and only if the heat-flux vector remains invariant along the system of the stream lines.

Proof: We consider

$${}_u\pounds\, R_{ab} = 0\ , \tag{3.1}$$

where ${}_u\pounds$ denotes the Lie derivative w.r.t. u^a.

(i) For the magnetofluid distribution characterized by (2.1), Eq. (3.1) gives

$$\begin{aligned}
&2A\, \dot{u}_{(a} u_{b)} + A_{,k} u^k u_a u_b + B_{,k} u^k g_{ab} + 2B u_{(a;b)} - 2\dot{q}_{(a} u_{b)} \\
&\quad - 2q_{(a} \dot{u}_{b)} - 2u_{k;(b} u_{a)} q^k - \nu(\sigma_{ab;k} u^k + 2u^k{}_{;(b}\, \sigma_{a)k}) \\
&\quad + 2\mu h_a u^k h_{[b;k]} + 2\mu h_b u^k h_{[a;k]} = 0\ ,
\end{aligned} \tag{3.2}$$

where $A = -(\rho + p + \mu|\bar{h}|^2)$, $B = \frac{1}{2}(\rho - p + \mu|\bar{h}|^2)$.

On innermultiplying (3.2) with $u^a u^b$, we have

$$(A + B)_{,a} u^a = 0\ . \tag{3.3}$$

The conservation law generator[12] $(R_a{}^b u^a)_{;b} = 0$ corresponding to the Ricci collineation with the result (3.3) yields

$$\begin{aligned}
&(A + B)\theta - q^a{}_{;a} = 0\ , \\
&\text{i.e.,}\quad \tfrac{1}{2}(\rho + 3p + \mu|\bar{h}|^2)\theta + q^a{}_{;a} = 0\ .
\end{aligned} \tag{3.4}$$

Consequently, we have

$$\theta = 0 \leftrightarrow q^a{}_{;a} = 0,\ \text{as}\ \rho + 3p + \mu|\bar{h}|^2 \neq 0\ . \tag{3.5}$$

(ii) From Eqs. (3.2) and (3.3) we derive

$$(A + B)\dot{u}_a - (q_{a;k} - q_{k;a}) u^k = 0\ . \tag{3.6}$$

If the heat-flux vector is invariant by the system of the stream lines, then we have

$$\begin{aligned}
&{}_u\pounds\, q_a = 0\ , \\
&\text{i.e.,}\quad (q_{a;k} - q_{k;a}) u^k = 0,\ \text{as}\ u^a q_a = 0\ .
\end{aligned} \tag{3.7}$$

Hence the proof of (ii) follows from (3.6) and (3.7).

Corollary 1: Let the space-time of the magnetofluid admit Ricci collineations w.r.t. the flow vector in addition to a harmonic heat-flux vector. If the isotropic pressure and the magnitude of the magnetic field vector are conserved w.r.t. the world line, the stream lines are shear-free.

On using the defining conditions of the harmonic heat-flux vector $q^a{}_{;a} = 0$, $q_{[a;b]} = 0$ in Theorem 1 and the equation of continuity (2.8), we obtain the shear-free stream lines.

Theorem 2: For the magnetofluid with the equation of

state $3p+\mu|\bar{h}|^2=\text{const}$, if the heat-flux vector is harmonic,

$$ {}_u\pounds R_{ab}=0 \leftrightarrow {}_u\pounds g_{ab}=0 . $$

Proof: The result ${}_u\pounds\, g_{ab}=0 \Rightarrow {}_u\pounds\, R_{ab}=0$ follows trivially. By Theorem 1 and the conditions $q^a{}_{;a}=0=q_{[a;b]}$ we get

$$ \theta=0, \quad \dot{u}_a=0 . \tag{3.8} $$

On using the equation of state $3p+\mu|\bar{h}|^2=\text{const}$ in Eq. (3.3), we have $\dot{\rho}=0$. Consequently, the equation of continuity (2.8) yields

$$ \sigma_{ab}=0 . \tag{3.9} $$

Thus Eqs. (3.8) and (3.9) lead to

$$ {}_u\pounds\, g_{ab}=0 . $$

Hence the proof is complete.

Theorem 3: If the magnetofluid with $|q|_{,a}u^a=0$, admits a Ricci collineation w.r.t. the geodesic stream lines, then

$$ \theta=0 \leftrightarrow \theta^*=0 , $$

where θ^* is the expansion parameter of the spacelike congruence q^a.

Proof: We have the expression for the expansion parameter[26] of q^a as

$$ \theta^*=\frac{1}{2}\left\{\left(\frac{q^a}{|\bar{q}|}\right)_{;a}-\left(\frac{q_a}{|\bar{q}|}\right)_{;b}u^a u^b\right\} . \tag{3.10} $$

For $\dot{u}^a=0$ and $|\bar{q}|_{,a}u^a=0$, (3.10) yields

$$ \theta^*=\frac{1}{2}\left(\frac{q^a{}_{;a}}{|\bar{q}|}\right). \tag{3.11} $$

Therefore, by (3.5) we write

$$ \theta=0 \leftrightarrow \theta^*=0 . \tag{3.12} $$

This completes the proof.

Theorem 4: If the space-time of magnetofluid with $\rho+3p+\mu|\bar{h}|^2=\text{const}$, admits a Ricci collineation with respect to the magnetic field vector, the 4-acceleration is orthogonal to the magnetic field vector if the stream lines are expansion-free.

Proof: A Ricci collineation with respect to magnetic field vector h^a gives

$$ \begin{aligned} &A_{,c}h^c u_a u_b + B_{,c}h^c g_{ab} + 2Au_{[a;c]}h^c u_b + 2Au_{[b;c]}h^c u_a \\ &+ 2Bh_{(a;b)} - \nu(\sigma_{ab;c}h^c + \sigma_{ac}h^c{}_{;b} + \sigma_{bc}h^c{}_{;a}) - 2u_{[b;c]}h^c q_a \\ &- 2u_{[a;c]}h^c q_b - 2u_{(a}q_{b)\,;c}h^c - 2h_{c\,;(b}u_{a)}q^c + 2\mu h_{(b;c)}h^c h_a \\ &+ 2\mu h_{(a;c)}h_b h^c = 0 . \end{aligned} \tag{3.13} $$

The corresponding conservation law generator furnishes

$$ \left(\tfrac{1}{2}(\rho-p-\mu|\bar{h}|^2)h^b - q_a h^a u^b - \nu\sigma_a{}^b h^a\right)_{;b}=0 . \tag{3.14} $$

By equations of motion (2.8) and (2.9), we deduce

$$ (\rho+p)\dot{u}_a h^a - p_{,a}h^a + \dot{q}_a h^a + q^a h_a \theta + u_{a;b}h^a q^b + \nu\sigma^{ab}{}_{;b}h_a=0 . \tag{3.15} $$

Equations (3.14) and (3.15) are coupled with the result (3.13) to obtain

$$ -\tfrac{1}{4}(\rho+3p+\mu|\bar{h}|^2)_{,a}h^a + \tfrac{1}{2}(\rho+3p+\mu|\bar{h}|^2)\dot{u}_a h^a + \theta h^a q_a=0 . \tag{3.16} $$

Consequently, $\rho+3p+\mu|\bar{h}|^2=\text{const}$ yields

$$ \dot{u}_a h^a=0 \leftrightarrow \theta=0 , \quad \text{as } \rho+3p+\mu|\bar{h}|^2\neq 0, \ h^a q_a\neq 0 . \tag{3.17} $$

Hence the proof is complete.

Remark: If $h^a q_a=0$ and $\rho+3p+\mu|\bar{h}|^2=\text{const}$, then it follows from (2.2) and (3.16) that

$$ {}_h\pounds R_{ab}=0 \Rightarrow T_{0,b}h^b=0 . $$

4. AFFINE COLLINEATION WITH RESPECT TO THE FLOW

An affine collineation w.r.t. the arbitrary vector field is defined[5] as

$$ {}_\xi\pounds\Gamma^a_{bc}=\xi^a{}_{;cb}+R^a{}_{cmb}\xi^m=0 , \tag{4.1} $$

where Γ^a_{bc} are the Christoffel symbols.

Theorem 5: If the magnetofluid admits an affine collineation w.r.t. the geodesic and expansion-free flow vector, then the gradient of the energy density along the flow is balanced by the heat influx and the affine collineation is a motion.

Proof: For the affine collineation w.r.t. u^a, (4.1) gives

$$ u^a{}_{;cb}+R^a{}_{cmb}u^m=0 . \tag{4.2} $$

On transvecting this result with u_a and using the normality condition $u^a u_a=1$ in the resulting equation, we have

$$ u^{a;c}u_{a;c}-R_{ac}u^a u^c=0 , \tag{4.3} $$

$$ \text{i.e.,} \quad 2\omega^2+2\sigma^2+\tfrac{1}{3}\theta^2+\dot{u}^a\dot{u}_a-R_{ac}u^a u^c=0 . \tag{4.4} $$

By combining (4.4) with Raychaudhuri's[27] equation

$$ \dot{\theta}+\tfrac{1}{3}\theta^2-\dot{u}^a{}_{;a}+2(\sigma^2-\omega^2)+R_{ac}u^a u^c=0 , \tag{4.5} $$

we get

$$ \dot{\theta}+\tfrac{2}{3}\theta^2+4\sigma^2-\dot{u}^a{}_{;a}+\dot{u}^a\dot{u}_a=0 . \tag{4.6} $$

Accordingly, the equation of continuity (2.8) yields

$$ \dot{\rho}+(\rho+p)\theta+q^a{}_{;a}-\tfrac{1}{2}\nu(\dot{u}^a{}_{;a}-\dot{u}^a\dot{u}_a-\dot{\theta}-\tfrac{2}{3}\theta^2)-\dot{u}_a q^a=0 . \tag{4.7} $$

$$ \text{When } \dot{u}^a=0 , \quad \theta=0 , \tag{4.8} $$

we have from (4.7)

$$ \dot{\rho}=-q^a{}_{;a} , \tag{4.9} $$

and by (4.6), we get $\sigma^2=0$. Consequently, we have

$$ u_{a;b}+u_{b;a}=0 . \tag{4.10} $$

This completes the proof.

Remark 1: For the magnetofluid with uniform matter density and admitting an affine collineation w.r.t. the flow vector, the specific entropy is preserved along the expansion-free and geodesic stream lines if and only if the heat flux vector is divergence-free.

The statement follows directly from (2.10), (4.6), and (4.8).

Remark 2: Equations (4.6) and (4.7) give

$$\dot{\theta} - \dot{u}^c{}_{;c} - 4\omega^2 - \dot{u}^a\dot{u}_a - (\rho + 3p + \mu|\bar{h}|^2) = 0. \quad (4.11)$$

According to Ellis[18] the active gravitational mass density for the magnetofluid is $-\frac{1}{2}(\rho + 3p + \mu|\bar{h}|^2)$. Hence we conclude by (4.11) that the active gravitational mass density of the magnetofluid is independent of the shear.

5. THE MAGNETOFLUID AND THE CONFORMAL MOTIONS

The infinitesimal transformation $\bar{x}^a = x^a + \xi^a \delta t$ (δt is an infinitesimal parameter) is defined[5] as

(i) conformal motion if

$$_{\xi}\pounds g_{ab} = \lambda g_{ab}, \quad (5.1)$$

where λ is a scalar function,

(ii) motion if

$$_{\xi}\pounds g_{ab} = 0. \quad (5.2)$$

(a) The conformal motion w.r.t. the magnetic field vector. Equation (5.1) gives

$$h_{a;b} + h_{b;a} = \lambda g_{ab}. \quad (5.3)$$

Consequently, we have

$$\lambda = \tfrac{1}{2}h^a{}_{;a} = -2\dot{u}_a h^a. \quad (5.4)$$

The Maxwell equations (2.4) with (5.4) yield

$$\lambda = 0. \quad (5.5)$$

Accordingly (5.1) gives

$$_{h}\pounds g_{ab} = 0. \quad (5.6)$$

Hence in the space-time V_4, permeated by the magnetofluid, a conformal motion w.r.t. the magnetic field vector degenerates into a motion.

(b) The conformal motion w.r.t. the flow vector reduces to a motion. It can be proved trivially.

(c) The conformal motion w.r.t. the heat-flux vector. Equation (5.1) gives

$$q_{a;b} + q_{b;a} = \lambda g_{ab}. \quad (5.7)$$

Consequently, we have

$$q^a{}_{;a} = -4\dot{u}^a q_a, \quad (5.8)$$

$$\dot{u}^a q_a = -\frac{|\bar{q}|_{,a}q^a}{|\bar{q}|}, \quad (5.9)$$

$$u_{a;b}q^a q^b = -\tfrac{1}{2}|\bar{q}|^2{}_{,a}u^a. \quad (5.10)$$

The equation $T^{ab}{}_{;b}q_a = 0$ with (2.1) and (5.10) produces

$$(\rho + p + \mu|\bar{h}|^2)\dot{u}^a q_a - p_{,b}q^b - |\bar{q}|^2\theta + 2u_{a;b}q^a q^b + \nu\sigma^{ab}{}_{;b}q_a - \tfrac{1}{2}\mu|\bar{h}|^2{}_{,b}q^b - \mu(h^a h^b)_{;b}q_a = 0. \quad (5.11)$$

When the magnetic field is uniform (i.e., $h_{a;b} = 0$), the Maxwell equations (2.4) imply

$$|\bar{h}|_{,a} = 0, \quad \theta = 0. \quad (5.12)$$

Consequently, (5.11) reduces to

$$(\rho + p + \mu|\bar{h}|^2)\dot{u}^a q_a - p_{,a}q^a + 2u_{a;b}q^a q^b + \nu\sigma^{ab}{}_{;b}q_a = 0. \quad (5.13)$$

Moreover, if the magnetofluid is shear-free ($\sigma_{ab} = 0$), Eq. (5.13) becomes

$$(\rho + p + \mu|\bar{h}|^2)\dot{u}^a q_a - p_{,a}q^a = 0. \quad (5.14)$$

Further, on using (5.8), (5.12), and (5.14) in (2.10), we have

$$\chi\rho_{0,a}u^a + T_0\rho_0 S_{,a}u^a = 5\dot{u}^a q_a. \quad (5.15)$$

From (5.14) and (5.15), we get

$$\chi(\rho + p + \mu|\bar{h}|^2)\rho_{0,a}u^a + (\rho + p + \mu|\bar{h}|^2)T_0\rho_0 S_{,a}u^a = 5p_{,a}q^a. \quad (5.16)$$

For the magnetofluid obeying mass conservation law,[19] i.e., $(\rho_0 u^a)_{;a} = 0$, (5.16) produces

$$S_{,a}u^a = 0 \Leftrightarrow P_{,a}q^a = 0. \quad (5.17)$$

Thus, the conformal motion w.r.t. the heat-flux vector implies that the shear-free flow of the magnetofluid satisfying the conditions $h_{a;b} = 0$ and $(\rho_0 u^a)_{;a} = 0$ is adiabatic if and only if the isotropic pressure is conserved in the direction of the heat-flux vector.

ACKNOWLEDGMENTS

The authors are grateful to Dr. L. Radhakrishna for valuable discussions. The authors' thanks are due to the referee for remarks on the original version of the paper. This work has been made possible by the Grant from Shivaji University to one of us (T.H.D.).

[1]W. R. Davis and G. H. Katzin, Amer. J. Phys. **30**, 750 (1962).
[2]W. R. Davis and M. K. Moss, Nuovo Cimento, Ser. X **38**, 1531 (1965).
[3]W. R. Davis and M. K. Moss, Nuovo Cimento, Ser. X **38**, 1558 (1965).
[4]W. R. Davis, *Lanczos Festschrift* (Academic, New York, 1974), Sec. X, p. 29.
[5]G. H. Katzin, J. Levine, and W. R. Davis, J. Math. Phys. **10**, 617 (1969).
[6]G. H. Katzin, J. Levine, and W. R. Davis, J. Math. Phys. **11**, 1578 (1970).
[7]G. H. Katzin, J. Levine, and W. R. Davis, Tensor, N.S. **21**, 51 (1970).
[8]G. H. Katzin and J. Levine, Colloq. Math. **26**, 21 (1972).
[9]W. R. Davis, in *Symmetry, Similarity and Group Theoretic Methods in Mechanics*, edited by P. G. Glocknar and M. C. Sing (University of Calgary Press, Canada, 1974).
[10]P. C. Aichelberg, J. Math. Phys. **11**, 2458 (1970).
[11]C. D. Collinson, J. Math. Phys. **11**, 818 (1970).
[12]C. D. Collinson, Gen. Rel. Grav. **1**, 137 (1970).
[13]V. D. Khade, Ph.D. thesis (Shivaji Univ., Kolhapur, India, 1974).
[14]A. Lichnerowicz, *Théories relativities de la gravitation et de l'electromagnetism* (Masson, Paris, 1955), Chap. I.
[15]R. R. Shaha, Ann. Inst. H. Poincaré **XX** (2), 189 (1974).
[16]G. G. Asgekar and T. H. Date, Proceedings, International Conference on Unified Field Theory and General Relativity, S. N. B. Inst. of Physical Sciences, Calcutta University (to be published), (1976).
[17]V. D. Khade and L. Radhakrishna, J. Shivaji Univ. **7** (14), 173 (1974).
[18]G. F. R. Ellis, *General Relativity and Cosmology*, edited by R. K. Sachs (Academic, New York, 1971), p. 104.
[19]P. J. Greenberg, Astrophys. J. **164**, 569 (1971).
[20]S. Kato, H. Nariai, and K. Tomita, Publ. Astron. Soc. Japan

19 (2), 130 (1967).
[21]T. H. Date, Ph.D. thesis (Shivaji Univ., Kolhapur, India, 1973).
[22]T. H. Date, Ann. Inst. H. Poincaré **XXIV** (4), 417 (1976).
[23]G. G. Asgekar and T. H. Date, J. Shiv. Univ. **7**, 151 (1974).
[24]C. Eckart, Phys. Rev. **58**, 919 (1940).
[25]A. Lichnerowicz, *Relativistic Hydrodynamics and Magnetohydrodynamics* (Benjamin, New York, 1967).
[26]P. J. Greenberg, J. Math. Anal. Appl. **30**, 128 (1970).
[27]A. K. Raychaudhuri, Phys. Rev. **98**, 1123 (1955).

Note on the representation spaces of O(2,1)

Debabrata Basu

Department of Physics, Indian Institute of Technology, Kharagpur, India
(Received 25 February 1976; revised manuscript received 14 July 1976)

We use Gel'fand's theory of invariant bilinear functionals to study a class of realizations of the representation spaces of O(2,1). We construct suitable Hermitian, invariant bilinear forms for the principal and supplementary series of representations. For the principal series the construction yields essentially the Bargmann realization of the Hilbert space while for the supplementary series we obtain other realizations in addition to the Bargmann space with nonlocal metric.

1. INTRODUCTION

The unitary irreducible representations (UIR) of the group O(2, 1) and the Hilbert spaces for the realization of these UIR were investigated by Bargmann[1] and by Barut and Fronsdal.[2] The main results that have emerged from these investigations are the following:

(a) The UIR of the group are of four types: (i) a principal series ($j=-\frac{1}{2}+i\lambda$), (ii) a supplementary series ($j=-\frac{1}{2}-\sigma$, $0<\sigma<\frac{1}{2}$), (iii) a positive discrete series ($2j=-ve$ integer, $m=-j+n$), and (iv) a negative discrete series ($2j=-ve$ integer, $m=j-n$).

(b) The principal series of representations can be realized in the space $L^2(\Phi)$ of square integrable Bargmann functions defined on the unit circle.

(c) The Hilbert space $L_\sigma(\Phi)$ of the Bargmann functions for the supplementary series involves a nonlocal metric.

(d) The discrete series is realized in the space $H^\pm$ of functions analytic inside the open unit circle.

More recently the extensive series of works of Moshinsky,[3] Wolf,[4] Mukunda,[5] and their collaborators have established several other realizations of the representation spaces of the group. As a result of these investigations many important results have been obtained and the theory has reached satisfactory stage.

The object of this short note is, therefore, not so much to derive new results as to reexamine certain aspects pertaining to the Hilbert spaces for the UIR from a different angle and to effect certain simplifications. A similar problem was treated in a previous paper[6] in the construction of the representation spaces of O(3, 1) in the O(2, 1) basis. There is, however, one important point where this note differs from the previous paper or from those of our predecessors. Here we work with the finite group element rather than with the algebra and have no occasion to deal with a specific subgroup in the reduced form.

We shall be concerned here with a class of realizations of the representation spaces which plays a basic role. These will be the representations on the spaces called D_x, the natural generalization of the spaces of homogeneous polynomials. Of course, the realizations of D_x which we consider do not exhaust all possible realizations of the representation spaces of O(2, 1) and there are many other realizations which frequently occur.[3-5]

Our analysis is based on Gel'fand's method[7] involving the study of invariant bilinear functionals in D_x. In this realization, the Hilbert space for the principal series of representations is seen to be isomorphic to the Bargmann space $L^2(\Phi)$ (Theorem 1). On the other hand, for the supplementary series, there exists other realizations of the representation spaces in addition to Bargmann's nonlocal realization $L_\sigma(\Phi)$ (Theorem 2).

2. INVARIANT BILINEAR FUNCTIONALS IN D_x

The UIR will be considered in the space D_x of homogeneous functions of two complex variables (ξ_1, ξ_2) transforming according to the fundamental representation of SU(1, 1) [covering group of O(2, 1)]. When D_x is realized as the space of functions of a single complex variable $z=\xi_1/\xi_2$, the operator T_u of the representation is given by

$$T_u f(z)=(\beta z+\bar{\alpha})^{2j} f\left(\frac{\alpha z+\bar{\beta}}{\beta z+\bar{\alpha}}\right). \tag{1}$$

Under certain conditions there may exist in D_x an invariant bilinear functional satisfying

$$(f_1, f_2)=(T_u f_1, T_u f_2). \tag{2}$$

The representations will then be unitary. Such a bilinear form, when it exists, may be taken as the scalar product in D_x. We may then complete D_x with respect to the norm $\|f\|$ to obtain a Hilbert space H in which D_x forms an everywhere dense subset. It has also been proved by Gel'fand that the existence of an invariant bilinear functional guarantees the irreducibility of the representations.

It is quite well known that under transformations of SU(1, 1) the unit circle is mapped onto itself. It is, therefore, natural to seek realizations of the UIR in the form of Eq. (1) for functions $f(z)$ defined in the open unit circle K_1 ($z\bar{z}<1$) or in the boundary K_2 ($z\bar{z}=1$) or in the annulus K_3 ($1<z\bar{z}<\infty$).

Let H be the space of functions regular in K. The completion of H with respect to the norm induced by the inner product,

$$(f_1, f_2)=\int_K f_1(z)\bar{f}_2(z)\,d\lambda(z), \tag{3}$$

yields the Hilbert space $L^2(K)$, provided a suitable measure $d\lambda(z)$ consistent with the unitarity of the representations exists in K. Here, the measure $d\lambda$ is defined by

Copyright © 1977 American Institute of Physics

$$d\lambda(z) = \omega(z)\, d\tau, \tag{4}$$

where $d\tau$ is the Euclidean measure on K.

Theorem 1: For the principal series of representations an invariant bilinear functional (f_1, f_2) does not exist unless f has bounded support on K_2.

Proof: If possible, let us suppose $K = K_1$ in Eq. (3). The invariant density $\omega(z)$, if it exists, is then determined by the unitarity condition and satisfies the functional equation,

$$\tilde{\omega}(z') = \left(\frac{\beta z + \bar{\alpha}}{\bar{\beta}\bar{z} + \alpha}\right)^{2j+1} \tilde{\omega}(z), \tag{5}$$

where

$$z' = \frac{\alpha z + \bar{\beta}}{\beta z + \bar{\alpha}}, \quad \tilde{\omega}(z) = (1 - |z|^2)\omega(z). \tag{6}$$

Setting $z = 0$, we obtain

$$\tilde{\omega}(\bar{\beta}\bar{t}) = \left(\frac{t}{\bar{t}}\right)^{2j+1} \tilde{\omega}(0). \tag{7}$$

Since the variable β does not appear in the rhs of the above equation, this solution is obviously unacceptable unless $\tilde{\omega}(0) = \omega(0) = 0$, which leads to

$$\omega(z) = 0.$$

There is, therefore, no invariant bilinear functional in K_1 for the principal series of representations. A similar result holds good for functions regular in K_3.

We next consider the space H of functions f having bounded support on K_2. The bilinear functional (f_1, f_2) is now concentrated on the boundary of the unit circle.[8] The Euclidean measure $d\tau$ then degenerates into the Euclidean line element $d\varphi$ on K_2 and $\omega(z)$ is determined by

$$\omega(z') = \left(\frac{\beta z + \bar{\alpha}}{\bar{\beta}\bar{z} + \alpha}\right)^{2j+1} \omega(z) \tag{8}$$

whence

$$\omega(z) = C z^{-(2j+1)}.$$

Setting the arbitrary constant $C = 1$, we have

$$(f_1, f_2) = \int_{K_2} f_1(z) \bar{f}_2(z) z^{-(2j+1)}\, d\varphi. \tag{9}$$

The Hilbert space $L^2(K_2)$ of functions defined on K_2 is therefore isomorphic to the Bargmann space $L^2(\Phi)$.

We next consider the supplementary series.

Theorem 2: For the supplementary series of representations an invariant bilinear form (f_1, f_2) with a nonlocal metric exists in K_2, and without nonlocality in K_1 and K_3.

Proof: We first consider the space H' of functions defined on K_2.

Since in Eq. (9) the condition of positive definiteness cannot be satisfied unless $(2j+1)$ is pure imaginary, for the supplementary series we seek a bilinear functional of the form

$$(f_1, f_2) = \int_{K_2} f_1(z_1) \bar{f}_2(z_2) \omega(z_1, z_2)\, d\varphi_1\, d\varphi_2 \tag{10}$$

which is a natural generalization of the bilinear form defined by Eq. (3). The kernel $\omega(z_1, z_2)$ is now determined by

$$\frac{\omega(z_1', z_2')}{z_1' z_2'} = (\beta z_1 + \bar{\alpha})^{1-2\sigma} (\bar{\beta}\bar{z}_2 + \alpha)^{1-2\sigma} \times \left(\frac{\beta z_2 + \bar{\alpha}}{\bar{\beta}\bar{z}_2 + \alpha}\right)^2 \frac{\omega(z_1, z_2)}{z_1 z_2}. \tag{11}$$

We now define $\tilde{\omega}(z_1, z_2) = (z_1 z_2)^{-1} \omega(z_1, z_2)$ and set $\beta = 0$, $\alpha = e^{i\varphi/2}$. Thus

$$\tilde{\omega}(z_1 e^{i\varphi}, z_2 e^{i\varphi}) = e^{-2i\varphi} \tilde{\omega}(z_1, z_2). \tag{12}$$

Since $f(z)$ has support in K_2, setting $z_2 = e^{-i\varphi}$ we have

$$\tilde{\omega}(z_1, z_2) = \frac{1}{z_2^2} \tilde{\omega}\left(\frac{z_1}{z_2}, 1\right) = \frac{1}{z_2^2} \tilde{\omega}_1\left(\frac{z_1}{z_2}\right) \text{ (say)}. \tag{13}$$

Further setting $z_1 = -z_2 = -1$ and combining Eqs. (11) and (13) we obtain

$$\tilde{\omega}_1\left(\frac{\bar{\alpha}\bar{\beta} - \alpha\beta - 1}{\bar{\alpha}\bar{\beta} - \alpha\beta + 1}\right) = \tilde{\omega}_1(-1)(1 + \bar{\alpha}\bar{\beta} - \alpha\beta)^{1-2\sigma}. \tag{14}$$

We now choose

$$\alpha = \frac{e^{i\pi/4}}{\sqrt{2}} (a+1)^{1/2} \sqrt{z_1},$$

$$\beta = \frac{e^{i\pi/4}}{\sqrt{2}} (a-1)^{1/2} (\bar{z}_2)^{1/2},$$

where a is a real number and is defined by

$$a = \left(1 - \frac{4 z_1 z_2}{(z_2 - z_1)^2}\right)^{1/2}.$$

With this parametrization Eq. (14) now reduces to

$$\tilde{\omega}_1\left(\frac{z_1}{z_2}\right) = C\left(1 - \frac{z_1}{z_2}\right)^{2\sigma - 1}, \tag{15}$$

where C is an arbitrary constant.

Setting $z_1 = e^{i\varphi}$, $z_2 = e^{i\psi}$ we may now immediately verify that this is equivalent to the Bargmann kernel. In this realization, the Hilbert space $L'(K_2)$ for the supplementary series is, therefore, isomorphic to the Bargmann space $L_\sigma(\Phi)$.

We next consider the space H of functions f regular in the open unit circle K_1. For the supplementary series, the existence of an invariant bilinear form in K_1 is guaranteed if the weight function $\omega(z)$ satisfies

$$\omega(z')\left(\frac{1 - |z'|^2}{1 - |z|^2}\right)^2 = |\beta z + \bar{\alpha}|^{-2-4\sigma} \omega(z) \tag{16}$$

whence

$$\omega(z) = C(1 - |z|^2)^{2\sigma - 1}, \tag{17}$$

where C is an arbitrary constant which we normalize as $C = 1/\Gamma(2\sigma)$.

We, therefore, complete H with respect to the norm $\|f\|$ to obtain a Hilbert space $L^2(K_1)$ in which the scalar

product is given by

$$(f_1, f_2) = \frac{1}{\Gamma(2\sigma)} \int_{K_1} f_1(z) \bar{f}_2(z) (1 - |z|^2)^{2\sigma - 1} d\tau. \tag{18}$$

A similar result holds good for functions regular in K_3.

We shall use Eq. (18) to calculate the normalizer for the supplementary series in the compact basis which was originally obtained by Bargmann[1] from the nonlocal form (10). In the O(2) basis $f(z)$ are the monomials

$$f(z) = f_m(z) = z^{j+m}.$$

Transforming to polar coordinates, we easily obtain the Bargmann normalizer,

$$(f_m, f_m) = \frac{1}{\Gamma(2\sigma)} \int_0^1 t^{m-\sigma-1/2} (1-t)^{2\sigma-1} dt$$

$$= \frac{\Gamma(m - \sigma + \frac{1}{2})}{\Gamma(m + \sigma + \frac{1}{2})}. \tag{19}$$

This expression is obviously valid for $m \geq 1$. However, since the line integral in the above equation can be converted into a contour integral of the Pochhammer type,

$$\int dt(\,) \Rightarrow \oint_p dt(\,),$$

Eq. (19) admits unique analytic continuation for all m.

We omit a detailed discussion of the positive and negative discrete series and conlcude this short note by pointing out that a scalar product similar to Eq. (18) holds also for the discrete series.

ACKNOWLEDGMENTS

The author would like to thank Professor N. Mukunda for helpful criticism and stimulating discussions.

[1] V. Bargmann, Ann. Math. **48**, 568 (1947).
[2] A. O. Barut and C. Fronsdal, Proc. R. Soc. A **287**, 532 (1965).
[3] P. Kramer, M. Moshinsky, and T. H. Seligman, in *Group Theory and Its Applications*, edited by E. Loebl (Academic, New York, 1975), Vol. III; see also M. Moshinsky and J. Patera, J. Math. Phys. **16**, 1866 (1975).
[4] K. B. Wolf, in *Group Theory and Its Applications*, edited by E. Loebl (Academic, New York, 1975), Vol. III; see also C. P. Boyer and K. B. Wolf, J. Math. Phys. **16**, 2215 (1975).
[5] N. Mukunda and B. Radhakrishnan, J. Math. Phys. **14**, 254 (1973).
[6] D. Basu and S. Datta Majumdar, J. Math. Phys. **17**, 185 (1976).
[7] I. M. Gel'fand, M. I. Graev, and N. Ya. Vilenkin, *Generalized Functions* (Academic, New York, 1966), Vol. 5.
[8] I. M. Gel'fand and G. E. Shilov, *Generalized Functions* (Academic, New York, 1964), Vol. I.

Monopole theory without string

Kuo-Shung Cheng*

Department of Applied Mathematics, National Chiao-Tung University, Hsinchu, Taiwan 300, Republic of China
(Received 27 April 1976)

We propose a Schrödinger equation for the nonrelativistic charge-monopole system without using strings and demonstrate that it is a correct equation for this system.

1. INTRODUCTION

It is well known that to quantize a classical system, i.e., to obtain a quantum mechanical system from a classical system, usually takes the following steps:

Classical equations of motion $\xrightarrow{\text{Step I}}$ Lagrangian of the classical system $\xrightarrow{\text{Step II}}$ Hamiltonian of the classical system $\xrightarrow{\text{Step III}}$ Quantum mechanical system (Schrödinger equation).

Many physically interesting quantum mechanical systems can be obtained from the corresponding classical system by using the above standard method. In the first step, there are three possible classes of systems as follows:

Class (1): systems with no Lagrangian whose Euler–Lagrange equation is the classical equation of motion.

Class (2): systems with no singularity-free Lagrangian whose Euler–Lagrange equation is the classical equation of motion. But, if we divide the configuration space into many overlapping regions, then, for each region, we can find a singularity-free Lagrangian whose Euler–Lagrange equation is the classical equation of motion.

Class (3): systems with Lagrangian whose Euler–Lagrange equation is the classical equation of motion.

In this paper, we are interested in considering the quantization of the classical systems which belong to class (2). It is a very interesting problem, at least theoretically, and it is not treated completely enough in the past. The classical electric charge-magnetic monopole system is a good example.[1]

We give the proposition to quantize this kind of classical system in Sec. 2, that is, we propose a Schrödinger equation with boundary conditions for this kind of classical system. In Sec. 3, we use the results in Sec. 2 to treat the monopole-charge case in detail, and in Sec. 4 we make some remarks.

2. QUANTIZATION

We consider a classical system whose equation of motion cannot be obtained from a singularity-free Lagrangian. But if we divide the configuration space into, say, two overlapped regions R_1 and R_2 (we can easily extend the results to the case of more than two overlapped regions), then we can find singularity-free Lagrangians L_1 in region R_1 and L_2 in region R_2, such that the equation of motion can be obtained from these Lagrangians. In the region $R_1 \cap R_2$, since the equation of motion obtained from Lagrangian L_1 and L_2 are the same, Lagrangian L_1 can differ from Lagrangian L_2 only up to a total time derivative of some function $f(q,t)$,[2] that is,

$$L_1 = L_2 + \frac{df(q,t)}{dt} \quad \text{in region } R_1 \cap R_2 . \tag{1}$$

Now, if we follow the usual procedure to construct the Hamiltonian, we get Hamiltonian H_1 in region R_1 and Hamiltonian H_2 in region R_2. In the region $R_1 \cap R_2$, we have

$$H_2(p,q,t) = H_1\left(p - \frac{\partial f}{\partial q}, q, t\right) - \frac{\partial f}{\partial t} . \tag{2}$$

If we follow the usual procedure, we would get the following Schrödinger equation for the wavefunction $\psi(q,t)$:

$$i\hbar \frac{\partial \psi(q,t)}{\partial t} = H_1 \psi(q,t) \quad \text{in region } R_1 , \tag{3}$$

$$i\hbar \frac{\partial \psi(q,t)}{\partial t} = H_2 \psi(q,t) \quad \text{in region } R_2 . \tag{4}$$

It is inconsistent in the overlapped region $R_1 \cap R_2$ due to Eq. (2), that is, in this region, if we use H_1 as the Hamiltonian, we get Eq. (3), and if we use H_2 as the Hamiltonian, we get Eq. (4), or, using Eq. (2), we can convert it into

$$i\hbar \frac{\partial}{\partial t}\left(e^{i[-f(q,t)/\hbar]}\psi(q,t)\right) = H_1\left(e^{-if(q,t)/\hbar}\psi(q,t)\right) . \tag{5}$$

It is interesting to note that the wavefunction in Eq. (5) differs from the wavefunction in Eq. (2) only up to a *phase* factor, or a gauge transformation of the second kind. Due to this fact, it is very natural to propose the following Schrödinger equation for this classical system. First we choose a surface σ in the overlapped region $R_1 \cap R_2$. The surface σ divides the configuration space into regions R_1' and R_2', where $R_1' \subset R_1$ and $R_2' \subset R_2$. Now the wavefunction $\psi(q,t)$ in region R_1', called $\psi_1(q,t)$, satisfies the following equation,

$$i\hbar \frac{\partial \psi_1}{\partial t} = H_1 \psi_1(q,t) , \quad q \in R_1' , \tag{6}$$

and the wavefunction in region R_2', called $\psi_2(q,t)$, satisfies the following equation,

$$i\hbar \frac{\partial \psi_2}{\partial t} = H_2 \psi_2(q,t) , \quad q \in R_2' . \tag{7}$$

On the surface σ, we assign the following boundary conditions

$$\psi_1(q,t)|_{q\in\sigma} = e^{if(q,t)/\hbar}\psi_2(q,t)|_{q\in\sigma} , \tag{8}$$

$$\frac{\partial}{\partial n}\left(\psi_1(q,t)\right)\Big|_{q\in\sigma} = \frac{\partial}{\partial n}\left(e^{if(q,t)/\hbar}\psi_2(q,t)\right)\Big|_{q\in\sigma} . \tag{9}$$

These boundary conditions are not the usual "wavefunc-

 Copyright © 1977 American Institute of Physics

tion continuity conditions," that is, we "twist" the wavefunction $\psi_2(q,t)$ and then connect it to the wavefunction $\psi_1(q,t)$.

From the proposed Schrödinger equation for the wavefunction, i.e., Eqs. (6), (7), (8), and (9), we know that the wavefunction $\psi(q,t)$ surely depends on the choice of surface σ. But it must be noted that the operators which represent physical observables also depend on the choice of surface σ.[2] For example, the operator representing the velocity $\dot{q}$ will be different in region R_1' and R_2' even in the overlapped region $R_1 \cap R_2$. The over-all effect of the choice of surface σ is that the expectation value of any observable does not depend on the choice of surface σ.[2]

We conclude that the proposed Schrödinger equations are well defined and physically sensible.

3. THE MONOPOLE-CHARGE CASE

We consider a particle of mass m with electric charge e moving in a magnetic field which is generated by a magnetic monopole of magnetic charge g resting at the coordinate origin.[3,4] The classical equations of motion are

$$m\ddot{\mathbf{r}} = \frac{eg}{c}\,\dot{\mathbf{r}} \times \frac{\mathbf{r}}{r^3}. \tag{10}$$

It is easy to see that the Euler-Lagrange equation of the following Lagrangian L,

$$L = \tfrac{1}{2}m\dot{\mathbf{r}}^2 - (eg/c)\dot{\mathbf{r}}\cdot\mathbf{A}, \tag{11}$$

are

$$m\ddot{\mathbf{r}} = (eg/c)\mathbf{r} \times (\nabla \times \mathbf{A}). \tag{12}$$

Thus, if we can find a vector potential $\mathbf{A}$ such that

$$\nabla \times \mathbf{A} = \mathbf{r}/r^3 = -\nabla(1/r), \tag{13}$$

then we can use this vector potential $\mathbf{A}$ to construct the Lagrangian in Eq. (11), and in turn we can obtain Eq. (10). But it is well known that we cannot have a singularity-free vector potential $\mathbf{A}$ which satisfies Eq. (13) over the space with the origin removed. Thus it is impossible to construct a singularity-free Lagrangian of the form as in Eq. (11). As pointed out by Wu and Yang,[5] we can divide the space into two overlapping regions R_1 and R_2. In each region we can obtain a singularity-free vector potential $\mathbf{A}$ which satisfies Eq. (13). In fact, we choose

$$\begin{aligned} &R_1\colon\ 0<r,\quad 0\le\theta<\pi/2+\delta,\quad 0\le\phi<2\pi,\\ &\qquad A_{1r}=A_{1\theta}=0,\quad A_{1\phi}=(1-\cos\theta)/r\sin\theta,\\ &R_2\colon\ 0<r,\quad \pi/2-\delta<\theta\le\pi,\quad 0\le\phi<2\pi,\\ &\qquad A_{2r}=A_{2\theta}=0,\quad A_{2\phi}=(-1-\cos\theta)/r\sin\theta,\quad 0<\delta<\pi/2. \end{aligned} \tag{14}$$

Now we can write down the Lagrangian

$$L_1 = \frac{m}{2}\left[\dot{r}^2 + r^2\dot{\theta}^2 + r^2\sin^2\theta\,\dot{\phi}^2\right] - \frac{eg}{c}\,r\sin\theta\,\dot{\phi}\left(\frac{1-\cos\theta}{r\sin\theta}\right), \tag{15}$$

$$L_2 = \frac{m}{2}\left[\dot{r}^2 + r^2\dot{\theta}^2 + r^2\sin^2\theta\,\dot{\phi}^2\right] - \frac{eg}{c}\,r\sin\theta\,\dot{\phi}\left(\frac{-1-\cos\theta}{r\sin\theta}\right). \tag{16}$$

It is easy to see that

$$L_1 = L_2 - \frac{2eg}{c}\,r\sin\theta\,\dot{\phi}\,\frac{1}{r\sin\theta} = L_2 - \frac{2eg}{c}\,\frac{d}{dt}\,\phi. \tag{17}$$

The Hamiltonians are

$$H_1 = \frac{1}{2m}\left[P_r^2 + \frac{P_\theta^2}{r^2} + \frac{1}{r^2\sin^2\theta}\left(P_\phi + \frac{eg}{c}(1-\cos\theta)\right)^2\right], \tag{18}$$

$$H_2 = \frac{1}{2m}\left[P_r^2 + \frac{P_\theta^2}{r^2} + \frac{1}{r^2\sin^2\theta}\left(P_\phi + \frac{eg}{c}(-1-\cos\theta)\right)^2\right]. \tag{19}$$

Now we choose the surface σ as $\theta=\pi/2$ and obtain the proposed Schrödinger equation from Sec. 2:

Region R_1': $0<r,\quad 0\le\theta<\pi/2,\quad 0\le\phi<2\pi$,

$$i\hbar\frac{\partial\psi_1}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi_1 + \frac{(eg/c)(1-\cos\theta)}{mr^2\sin^2\theta}\,\frac{\hbar}{i}\,\frac{\partial\psi_1}{\partial\phi} + \frac{[(eg/c)(1-\cos\theta)]^2}{2mr^2\sin^2\theta}\,\psi_1; \tag{20}$$

Region R_2': $0<r,\quad \pi/2<\theta\le\pi,\quad 0\le\phi<2\pi$,

$$i\hbar\frac{\partial\psi_2}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi_2 + \frac{(eg/c)(-1-\cos\theta)}{mr^2\sin^2\theta}\,\frac{\hbar}{i}\,\frac{\partial\psi_2}{\partial\phi} + \frac{[(eg/c)(-1-\cos\theta)]^2}{2mr^2\sin^2\theta}\,\psi_2. \tag{21}$$

The boundary conditions at the surface $\theta=\pi/2$ are

$$\psi_1(r,\theta,\phi,t)\big|_{\theta=\pi/2} = e^{-i(2eg/\hbar c)\phi}\psi_2(r,\theta,\phi,t)\big|_{\theta=\pi/2}, \tag{22}$$

$$\frac{\partial}{\partial\theta}\psi_1(r,\theta,\phi,t)\Big|_{\theta=\pi/2} = \frac{\partial}{\partial\theta}\left[e^{-i(2eg/\hbar c)\phi}\psi_2(r,\theta,\phi,t)\right]\Big|_{\theta=\pi/2} \tag{23}$$

If we assume $eg/\hbar c=\mu$, then, from the above equations, we can see that (single-valued wavefunction)

$$2\mu = n,\quad n \text{ is an integer}, \tag{24}$$

which is Dirac's charge quantization condition.

Now let us try to solve Eqs. (20) and (21) subjecting Eqs. (22) and (23) to the boundary conditions. We set

$$\psi_1(r,\theta,\phi,t) = \exp\left(-\frac{iE_1t}{\hbar}\right)f_1(r)g_1(\theta)h_1(\phi), \tag{25}$$

$$\psi_2(r,\theta,\phi,t) = \exp\left(-\frac{iE_2t}{\hbar}\right)f_2(r)g_2(\theta)h_2(\phi), \tag{26}$$

where

$$h_1(\phi) = e^{im_1\phi},\quad h_2(\phi) = e^{im_2\phi},\quad m_1 \text{ and } m_2 \text{ are integers}, \tag{27}$$

$$\frac{1}{\sin\theta}\,\frac{d}{d\theta}\left(\sin\theta\frac{dg_1(\theta)}{d\theta}\right) - \frac{1}{\sin^2\theta}\left[(m_1+\mu(1-\cos\theta)\right]^2 g_1(\theta) + \lambda_1 g_1(\theta) = 0, \tag{28}$$

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dg_2(\theta)}{d\theta}\right)$$

$$-\frac{1}{\sin^2\theta}[m_2+\mu(-1-\cos\theta)]^2 g_2(\theta)+\lambda_2 g_2(\theta)=0, \tag{29}$$

and

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{df_1(r)}{dr}\right)-\frac{\lambda_1}{r^2}f_1(r)+\frac{2mE_1}{\hbar^2}f_1(r)=0, \tag{30}$$

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{df_2(r)}{dr}\right)-\frac{\lambda_2}{r^2}f_2(r)+\frac{2mE_2}{\hbar^2}f_2(r)=0. \tag{31}$$

We can write down solutions for Eqs. (28) and (29) (without loss of generality, we will assume $\mu>0$). We also set $z=(1-\cos\theta)/2$ and $z'=(1+\cos\theta)/2$. We have:

Case (1): $0\leq m_1$:

$$g_1(\theta)=z^{m_1/2}(z-1)^{m_1/2+\mu}$$
$$\times{}_2F_1[(\mu+m_1)+\tfrac{1}{2}+\tfrac{1}{2}(1+4\lambda_1+4\mu^2)^{1/2},$$
$$(\mu+m_1)+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda_1+4\mu^2)^{1/2};\ m_1+1;z];$$

Case (2): $-2\mu\leq m_1<0$:

$$g_1(\theta)=z^{-m_1/2}(z-1)^{m_1/2+\mu}$$
$$\times{}_2F_1[\mu+\tfrac{1}{2}+\tfrac{1}{2}(1+4\lambda_1+4\mu^2)^{1/2},$$
$$\mu+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda_1+4\mu^2)^{1/2};\ -m_1+1;z];$$

Case (3): $m_1<-2\mu$:

$$g_1(\theta)=z^{-m_1/2}(z-1)^{-(m_1/2+\mu)}$$
$$\times{}_2F_1[-(\mu+m_1)+\tfrac{1}{2}+\tfrac{1}{2}(1+4\lambda_1+4\mu^2)^{1/2},$$
$$-(\mu+m_1)+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda_1+4\mu^2)^{1/2};\ -m_1+1;z];$$

and for $g_2(\theta)$:

Case (1): $2\mu\leq m_2$:

$$g_2(\theta)=(z')^{m_2/2}(z'-1)^{m_2/2-\mu}$$
$$\times{}_2F_1[(m_2-\mu)+\tfrac{1}{2}+\tfrac{1}{2}(1+4\lambda_2+4\mu^2)^{1/2},$$
$$(m_2-\mu)+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda_2+4\mu^2)^{1/2};\ m_2+1;z'];$$

Case (2): $0\leq m_2<2\mu$:

$$g_2(\theta)=(z')^{m_2/2}(z'-1)^{\mu-m_2/2}$$
$$\times{}_2F_1[\mu+\tfrac{1}{2}+\tfrac{1}{2}(1+4\lambda_2+4\mu^2)^{1/2},$$
$$\mu+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda_2+4\mu^2)^{1/2};\ m_2+1;z'];$$

Case (3): $m_2<0$:

$$g_2(\theta)=(z')^{-m_2/2}(z'-1)^{\mu-m_2/2}$$
$$\times{}_2F_1[(\mu-m_2)+\tfrac{1}{2}+\tfrac{1}{2}(1+4\lambda_2+4\mu^2)^{1/2},$$
$$(\mu-m_2)+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda_2+4\mu^2)^{1/2};\ 1-m_2;z'].$$

Thus the general solution for the wavefunctions ψ_1 and ψ_2 are

$$\psi_1(r,\theta,\phi,t)=\sum_{E_1,\lambda_1,m_1}a_{E_1\lambda_1 m_1}e^{-iE_1t/\hbar}$$
$$\times f_1(E_1,\lambda_1;r)g_1(\lambda_1,m_1;\theta)h_1(m_1;\phi), \tag{32}$$

$$\psi_2(r,\theta,\phi,t)=\sum_{E_2,\lambda_2,m_2}a_{E_2\lambda_2 m_2}e^{-iE_2t/\hbar}$$
$$\times f_2(E_2,\lambda_2;r)g_2(\lambda_2,m_2;\theta)h_2(m_2;\phi). \tag{33}$$

Substituting Eqs. (32) and (33) into the boundary conditions, Eqs. (22) and (23), we get

$$m_1=m_2-2\mu=m,\quad E_1=E_2=E,\quad \lambda_1=\lambda_2=\lambda, \tag{34}$$

and the following important relations:

Case (1): $m_1=m_2-2\mu=m\geq 0$:

$$(\mu+m)+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda_1+4\mu^2)^{1/2}=-n_1,\quad n_1=0,1,2,\ldots,$$
$$\text{or}\quad \lambda=(\mu+m+n_1)(\mu+m+n_1+1)-\mu^2; \tag{35}$$

Case (2): $0>m=m_1=m_2-2\mu\geq -2\mu$:

$$(\mu+\tfrac{1}{2})-\tfrac{1}{2}(1+4\lambda_1+4\mu^2)^{1/2}=-n_2,\quad n_2=0,1,2,\ldots,$$
$$\text{or}\quad \lambda=(\mu+n_2)(\mu+n_2+1)-\mu^2; \tag{36}$$

Case (3): $m_1=m_2=2\mu=m<-2\mu$:

$$-(\mu+m)+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda+4\mu^2)^{1/2}=-n_3,\quad n_3=0,1,2,\ldots,$$
$$\text{or}\quad \lambda=(n_3-\mu-m)(n_3-\mu-m+1)-\mu^2. \tag{37}$$

To see how we obtain condition (35), we rewrite the solution for Case (1) as follows:

$$m_1=m_2-2\mu=m\geq 0:$$

$$g_1=z^{m/2}(z-1)^{m/2+\mu}$$
$$\times{}_2F_1[(\mu+m)+\tfrac{1}{2}+\tfrac{1}{2}(1+4\lambda+4\mu^2)^{1/2},$$
$$(\mu+m)+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda+4\mu^2)^{1/2};\ m+1;z], \tag{38}$$

$$g_2=(z')^{m/2+\mu}(z'-1)^{m/2}$$
$$\times{}_2F_1[(\mu+m)+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda+4\mu^2)^{1/2},$$
$$(\mu+m)+\tfrac{1}{2}-\tfrac{1}{2}(1+4\lambda+4\mu^2)^{1/2};\ m+2\mu+1;z']. \tag{39}$$

As a function of z, g_1 and g_2 satisfy the same second order linear differential equation, in which g_1 is analytic at $z=0$, but g_2 is analytic at $z=1$, and from the boundary condition g_1 and g_2 take the same value and same derivative at the point $z=\frac{1}{2}$. Thus, the only possibility is that g_1 and g_2 as a function of z are polynomials. So we obtain the condition (35). The same argument can be applied for conditions (36) and (37).

We see from Eqs. (35), (36), and (37) that λ can take only the following values:

$$\lambda=(l+\mu)(l+\mu+1)-\mu^2, \tag{40}$$

where l is a positive integer or zero. It is easy to see that, for a given l, we can have $-l-2\mu\leq m\leq l$ corresponding to the same eigenvalue l. These solutions were discussed in Refs. 3 and 4, but using the singular potential $\mathbf{A}_1$.

4. CONCLUDING REMARKS

(1) We have successfully proposed a Schrödinger equation for a classical system which possesses no singularity-free Lagrangian over the whole configuration space, but if we divide the configuration space into overlapped regions, then, for each region, the classical system possesses a singularity-free Lagrangian.

(2) We have demonstrated that the proposed Schrödinger equation for the charge-monopole case gives Dirac's charge quantization conditions $2\mu = n$ and leads to quantization of the angular-momentum-like quantity $\lambda = (l+\mu)(l+\mu+1) - \mu^2$.

ACKNOWLEDGMENTS

Interesting discussions with professors Ni Wei-Tou, Lee Yee-Yen, and Shaw Jin-Chiang are deeply appreciated by the author.

*Works partially supported by the National Science Council of Taiwan, Rep. of China.

[1]P. A. M. Dirac, Phys. Rev. **74**, 817 (1948).
[2]Kuo-Shung Cheng, J. Math. Phys. **15**, 808 (1974).
[3]M. Fierz, Helv. Phys. Acta **17**, 27 (1944).
[4]C. A. Hurst, Ann. Phys. (N.Y.) **50**, 51 (1968).
[5]Tai-Tsun Wu and Chen-Ning Yang, Phys. Rev. D **12**, 3845 (1975).

Eigenvalue problem for Lagrangian systems. VIII. Eigenvalue bounds with application to fluid and plasma flows

E. M. Barston

Department of Mathematics, University of Illinois at Chicago Circle, Chicago, Illinois 60680
(Received 1 June 1976)

Circle theorems are obtained which furnish bounds for the complex eigenvalues arising from the linear stability analysis of states of steady motion of conservative dynamical systems. These results generalize the well-known circle theorems of Howard, Eckart, and others for fluid flows to arbitrary gyroscopic Lagrangian systems. Explicit bounds are obtained for the plane parallel flows of an inviscid fluid and a multicomponent plasma.

I. INTRODUCTION

Numerous theorems have appeared in the recent literature that locate, within certain circles in the complex plane, the complex eigenvalues arising in the linear stability analysis of steady flows of conservative fluids and magnetofluids.[1-11] The progenitor of these results is the well-known circle theorem of Howard[1] for inviscid, incompressible, plane parallel flow.

In this paper, we obtain circle theorems valid for a far larger class of systems, including rotating elastic bodies, macroscopic and microscopic models of N-component plasmas, as well as ideal fluids and magnetofluids. We achieve this generality by making use of the fact that the linearized equations governing small perturbations about a state of steady motion of a conservative dynamical system, when expressed in appropriate variables, assume the canonical form[12-20]

$$P\ddot{\xi}+A\dot{\xi}+H\xi(t)=0, \tag{1}$$

where P, iA, and H are time-independent, linear, formally self-adjoint operators with domain and range in an inner product space E. [A system satisfying Eq. (1) is said to be a gyroscopic Lagrangian system.] The search for solutions of the form $\xi(t)=\zeta\exp(i\omega t)$, where ζ is a fixed nonzero element of E and ω is a real or complex constant, leads to the quadratic eigenvalue problem[21]

$$(\omega^2P-\omega iA-H)\zeta=0 \tag{2}$$

for the eigenvalue ω and associated eigenvector ζ. Various localization theorems for the complex eigenvalues are readily obtained from Eq. (2), using only elementary algebra and a simple translation of the eigenvalues. This is done in Sec. II. In particular, we show that every complex eigenvalue ω ($\mathrm{Im}\omega\neq 0$) must lie in the intersection S of an infinite set of circles centered along the real axis, give formulas for the parameters of the circle of minimum radius, and discuss some of the characteristics of the set S. Two applications are discussed in Sec. III. Howard's circle theorem[1] is obtained as a direct consequence of the results of Sec. II, and the complex eigenvalues arising in a macroscopic model of an N-component plasma in plane parallel flow are shown to lie in the circle with center $(-k_z(u+U)/2, 0)$ and radius $|k_z|(U-u)/2$, where U denotes the maximum value of the streaming velocity (over all N streams), u is the minimum value of the streaming velocity, and k_z is the Fourier wavenumber in the direction of streaming. This reduces, in the case of a single uncharged fluid, to the circle theorem of Eckart[2] for compressible plane parallel flow.

Plasma dynamics would appear to be the most fertile area for applications. The formulation of magnetohydrodynamics, macroscopic N-component theory, and microscopic theory given in Refs. 14, 15, and 17 permits the use of our results at each of these levels of description. Hamieri[22] has recently obtained interesting results concerning the stability of the mhd cylindrical pinch with parallel or helical flows through the use of techniques similar to those presented here.

II. CIRCLE THEOREMS

Let E be an inner product space with inner product $(\ ,\ )$. Let P, iA, and H be formally self-adjoint linear operators with range in E and with domains of definition D_P, D_A, D_H, respectively (all linear subspaces of E), and suppose that $P>0$ on D_P. [An operator F is said to be formally self-adjoint provided $(\eta,F\zeta)=(F\eta,\zeta)$ for all $\eta,\zeta\in D_F$.] Let $D\equiv D_P\cap D_A\cap D_H$. For all nonzero $\zeta\in D$, we define

$$d(\zeta)\equiv\frac{(\zeta,iA\zeta)^2}{4(\zeta,P\zeta)^2}+\frac{(\zeta,H\zeta)}{(\zeta,P\zeta)}, \tag{3}$$

and

$$Q_\pm(\zeta)\equiv\frac{(\zeta,iA\zeta)}{2(\zeta,P\zeta)}\pm\sqrt{d(\zeta)}\,. \tag{4}$$

The inner products appearing in Eqs. (3)–(4) are all real-valued and thus $d(\zeta)$ is real-valued, due to the formal self-adjointness of P, iA, and H.

Let ζ be a fixed nonzero element of E, $\omega\in Z$ (the set of all complex numbers) and suppose that $\xi(t)=\exp(i\omega t)\zeta$ is a solution of the system (1). Then we must have $\zeta\in D$ ($\omega\neq 0$) and $H_\omega\zeta=0$, where $H_\omega\equiv H+\omega iA-\omega^2P$ is a linear operator from D into E, and H_ω is formally self-adjoint on D for real ω. It follows that $(\zeta,H_\omega\zeta)=0$, i.e.,

$$\omega^2(\zeta,P\zeta)-\omega(\zeta,iA\zeta)-(\zeta,H\zeta)=0 \tag{5}$$

and so $\omega=Q_+(\zeta)$ or $\omega=Q_-(\zeta)$. Hence if the extended real number $\Delta\equiv\inf_D d(\eta)\geq 0$ (it is to be understood that "$\inf_D$" means infimum over all nonzero elements of D),

Copyright © 1977 American Institute of Physics

then every root ω of Eq. (5) must be real. Clearly $\mathrm{Im}\omega \neq 0$ holds if and only if $d(\xi) < 0$; in that case we have

$$\mathrm{Re}\omega = \frac{(\xi, iA\xi)}{2(\xi, P\xi)}, \quad \mathrm{Im}\omega = \pm\sqrt{-d(\xi)} \quad (\mathrm{Im}\omega \neq 0) \tag{6}$$

the set $\tilde{D} \equiv \{\eta \mid \eta \in D,\ \eta \neq 0,\ d(\eta) < 0\}$ is nonempty, $\Delta \equiv \inf_D d(\eta) = \inf_{\tilde{D}} d(\eta) < 0$, and by Eq. (6),

$$|\omega|^2 = -\frac{(\xi, H\xi)}{(\xi, P\xi)} \leq \sup_{\tilde{D}} \frac{(\eta, -H\eta)}{(\eta, P\eta)}. \tag{7}$$

For future reference, we give a formal statement of this result.

Lemma 1: Let ω be an eigenvalue of Eq. (1) with corresponding eigenvector ξ, and let $\mathrm{Im}\omega \neq 0$. Then $\xi \in D$ and Eq. (7) holds.

We consider herein only eigenvalues ω with $\mathrm{Im}\omega \neq 0$. In view of the preceeding results, *we assume henceforth that* $\Delta < 0$. For all $\alpha \in R$ (the set of real numbers) and all nonzero $\eta \in D$,

$$\frac{(\eta, -H_\alpha\eta)}{(\eta, P\eta)} = \left(\alpha - \frac{(\eta, iA\eta)}{2(\eta, P\eta)}\right)^2 - d(\eta),$$

so that

$$d(\eta) = -\min_{\alpha \in R} \frac{(\eta, -H_\alpha\eta)}{(\eta, P\eta)} = \max_{\alpha \in R} \frac{(\eta, H_\alpha\eta)}{(\eta, P\eta)}. \tag{8}$$

Thus a necessary and sufficient condition that $\Delta < 0$ is that there exist a nonzero $\eta \in D$ such that $(\eta, H_\alpha\eta) < 0$ for all $\alpha \in R$ [or equivalently, $(\eta, -H_\alpha\eta) > 0$ for all $\alpha \in R$]. This statement remains valid with D replaced by $\tilde{D}$, so that $\Delta < 0$ implies that for all $\alpha \in R$, $\sup_{\tilde{D}}(\eta, -H_\alpha\eta)/(\eta, P\eta) > 0$. Indeed, for every $\eta \in \tilde{D}$, Eq. (8) gives

$$-d(\eta) \leq \frac{(\eta, -H_\alpha\eta)}{(\eta, P\eta)} \leq \sup_{\tilde{D}} \frac{(\phi, -H_\alpha\phi)}{(\phi, P\phi)}$$

so that

$$\sup_{\tilde{D}} \frac{(\phi, -H_\alpha\phi)}{(\phi, P\phi)} \geq \sup_{\tilde{D}}[-d(\eta)] = -\inf_{\tilde{D}} d(\eta)$$

$$= -\inf_D d(\xi) = -\Delta. \tag{9}$$

Assume that $\xi(t) = \exp(i\omega t)\xi$ is a solution of Eq. (1) as before, let α be a real number, and set

$$\eta(t) \equiv \exp(i(\omega - \alpha)t)\xi, \quad -\infty < t < \infty. \tag{10}$$

Then $(\partial^n\eta/\partial t^n)(t) \in D$ for all $n = 0, 1, 2, \cdots$ and each real t, $\xi(t) = \exp(i\omega t)\xi = \exp(i\alpha t)\eta(t)$ satisfies (1), and we have $\dot{\xi} = \exp(i\alpha t)(i\alpha\eta + \dot{\eta})$, $\ddot{\xi} = \exp(i\alpha t)(-\alpha^2\eta + 2i\alpha\dot{\eta} + \ddot{\eta})$. Equation (1) then implies that

$$P\ddot{\eta} + A_\alpha\dot{\eta} + H_\alpha\eta(t) = 0, \quad -\infty < t < \infty, \tag{11}$$

where $A_\alpha \equiv A + i2\alpha P$, and P, iA_α, H_α are all formally self-adjoint on D with $P > 0$ on D. Thus $\omega - \alpha$ is an eigenvalue of the system (11) with eigenvector ξ, and the application of Lemma 1 to the system (11) leads immediately to the following theorem (we have for convenience of reference, also included several previous results):

Theorem 1: (A) A necessary condition that the system (1) possess an eigenvalue ω with $\mathrm{Im}\omega \neq 0$ is that $\Delta \equiv \inf_D d(\phi) < 0$.

(B) A necessary and sufficient condition that $\Delta < 0$ is the existence of a $\phi \in D$ such that $(\phi, -H_\alpha\phi) > 0$ for all $\alpha \in R$.

Suppose that $\Delta < 0$. Define the extended real-valued function δ for all $\alpha \in R$ by

$$\delta(\alpha) \equiv \left(\sup_{\tilde{D}} \frac{(\phi, -H_\alpha\phi)}{(\phi, P\phi)}\right)^{1/2},$$

and let r be any extended real-valued function majorizing δ[i.e., $r(\alpha) \geq \delta(\alpha)$ for all $\alpha \in R$]. Then each of the following statements holds:

(C) Let $\alpha \in R$. Every eigenvalue ω of Eq. (1) with $\mathrm{Im}\omega \neq 0$ lies in the circle $C_\alpha(r) \equiv \{z \mid z \in Z,\ (\mathrm{Re}z - \alpha)^2 + (\mathrm{Im}z)^2 \leq r^2(\alpha)\}$ with center $(\alpha, 0)$ and radius $r(\alpha)$. Hence $\omega \in S(r) \equiv \cap_{\alpha \in R} C_\alpha(r)$.

(D) $r(\alpha) \geq \delta(\alpha) \geq (-\Delta)^{1/2}$, $-\infty < \alpha < \infty$.

(E) (semicircle theorem for exponentially growing modes) Let $\beta \in R$ such that $r(\beta) \leq r(\alpha)$ for all $\alpha \in R$, so that $C_\beta(r)$ is the circle of minimum radius (for the given majorizing function r). Then every eigenvalue ω of Eq. (1) with $\mathrm{Im}\omega < 0$ lies in the semicircle $C_\beta(r) \cap \{z \mid z \in Z,\ \mathrm{Im}z < 0\}$.

(F) Every eigenvalue of Eq. (1) with $\mathrm{Im}\omega \neq 0$ lies within the rectangle

$$M \equiv \left\{ z \mid z \in Z,\ |\mathrm{Im}z| \leq (-\Delta)^{1/2},\ \tfrac{1}{2}\inf_{\tilde{D}} \frac{(\phi, iA\phi)}{(\phi, P\phi)} \right.$$

$$\left. \leq \mathrm{Re}z \leq \tfrac{1}{2}\sup_{\tilde{D}} \frac{(\phi, iA\phi)}{(\phi, P\phi)} \right\} \text{ [see Eq. (6)]}.$$

The introduction of the auxiliary function r allows one the convenience of using an upper bound for $\delta(\alpha)$ rather than $\delta(\alpha)$ itself. In practice this is often a necessity, due to the difficulty of computing $\delta(\alpha)$ exactly.

Theorem 1 states that all eigenvalues with nonzero imaginary part must lie within each of the circles $C_\alpha(\delta)$, their intersection $S(\delta)$, and within the rectangle M as well. Among the circles $C_\alpha(\delta)$, the circle $C_\beta(\delta)$ of smallest radius is of primary interest. The inequality expressed in Theorem 1(D) suggests the possibility that the least radius is $(-\Delta)^{1/2}$. This is indeed correct. The next theorem establishes this fact and gives maximum and minimum principles for the center coordinate β.

Theorem 2: Let $-\infty < \Delta < 0$. The parameters β and $\delta(\beta)$ of the circle $C_\beta(\delta)$ of smallest radius are given by

$$\beta = \inf_D\left(\frac{(\xi, iA\xi)}{2(\xi, P\xi)} + \sqrt{d(\xi) - \Delta}\right)$$

$$= \sup_D\left(\frac{(\xi, iA\xi)}{2(\xi, P\xi)} - \sqrt{d(\xi) - \Delta}\right), \tag{12}$$

$$\delta^2(\beta) = \sup_{\tilde{D}} \frac{(\xi, -H_\beta\xi)}{(\xi, P\xi)} = \sup_D \frac{(\xi, -H_\beta\xi)}{(\xi, P\xi)}$$

$$= -\Delta = \inf_{\alpha \in R} \delta^2(\alpha). \tag{13}$$

Furthermore, if $\{\xi_n\}_{n=1}^\infty$ is any sequence from D such that $\lim_{n\to\infty} d(\xi_n) = \Delta$, then $\lim_{n\to\infty}[(\xi_n, iA\xi_n)/(\xi_n, P\xi_n)] = 2\beta$. The proof of this theorem is given in the Appendix.

The set $S(\delta)=\cap_{\alpha\in R}C_\alpha(\delta)\subset C_\beta(\delta)$ will usually provide a better result than $C_\beta(\delta)$. [For the case of inviscid, incompressible, plane parallel flow, $S(\delta)=C_\beta(\delta)$ and $S(\delta)$ provides no further information. See Sec. III.] However, $S(\delta)$ may be difficult to construct. We establish two elementary properties of $S(\delta)$: (1) $S(\delta)$ contains the vertical diameter $\{z\,|\,\mathrm{Re}z=\beta,\ |\mathrm{Im}z|\leq(-\Delta)^{1/2}\}$ of the circle $C_\beta(\delta)$; and (2), $S(\delta)\subset M\cap C_\beta(\delta)$. We have, for each $\alpha\in R$,

$$\delta^2(\alpha)=\sup_{\tilde D}\frac{(\zeta,-H_\alpha\zeta)}{(\zeta,P\zeta)}$$

$$=\sup_{\tilde D}\left(\frac{(\zeta,-H_\beta\zeta)}{(\zeta,P\zeta)}+\alpha^2-\beta^2+(\beta-\alpha)\frac{(\zeta,iA\zeta)}{(\zeta,P\zeta)}\right). \quad (14)$$

Let $\{\zeta_n\}_{n=1}^\infty$ be a sequence from $\tilde D$ such that $d(\zeta_n)\to\Delta$. By Theorem 2, $[(\zeta_n,iA\zeta_n)/(\zeta_n,P\zeta_n)]\to 2\beta$, and therefore

$$\frac{(\zeta_n,-H_\alpha\zeta_n)}{(\zeta_n,P\zeta_n)}=-d(\zeta_n)+\frac{(\zeta_n,iA\zeta_n)^2}{4(\zeta_n,P\zeta_n)^2}-\alpha\frac{(\zeta_n,iA\zeta_n)}{(\zeta_n,P\zeta_n)}+\alpha^2$$

$$\to-\Delta+(\alpha-\beta)^2=\delta^2(\beta)+(\alpha-\beta)^2.$$

Since

$$\delta^2(\alpha)=\sup_{\tilde D}\frac{(\zeta,-H_\alpha\zeta)}{(\zeta,P\zeta)}\geq\frac{(\zeta_n,-H_\alpha\zeta_n)}{(\zeta_n,P\zeta_n)},\quad n=1,2,3,\cdots,$$

we have $\delta^2(\alpha)\geq\delta^2(\beta)+(\alpha-\beta)^2$, and it follows from the Pythagorean theorem that $C_\alpha(\delta)$ contains the vertical diameter of $C_\beta(\delta)$. This proves (1). Our sequence $\{\zeta_n\}_{n=1}^\infty\subset\tilde D$ satisfies $d(\zeta_n)\to\Delta$ and $[(\zeta_n,iA\zeta_n)/(\zeta_n,P\zeta_n)]\to 2\beta$, which implies that Eq. (12) holds with D replaced by $\tilde D$. Hence

$$m_1\equiv\inf_{\tilde D}\frac{(\zeta,iA\zeta)}{2(\zeta,P\zeta)}\leq\beta\leq\sup_{\tilde D}\frac{(\zeta,iA\zeta)}{2(\zeta,P\zeta)}\equiv m_2.$$

Equation (14) gives

$$\delta^2(\alpha)\leq\sup_{\tilde D}\frac{(\zeta,-H_\beta\zeta)}{(\zeta,P\zeta)}+\alpha^2-\beta^2+\sup_{\tilde D}(\beta-\alpha)\frac{(\zeta,iA\zeta)}{(\zeta,P\zeta)},$$

so that

$$\delta^2(\alpha)\leq-\Delta+(\alpha-\beta)^2+2(\alpha-\beta)(\beta-m_1)\equiv r_1^2(\alpha),\quad\alpha\geq\beta, \quad (15)$$

$$\delta^2(\alpha)\leq-\Delta+(\alpha-\beta)^2+2(\beta-\alpha)(m_2-\beta)\equiv r_2^2(\alpha),\quad\alpha\leq\beta. \quad (16)$$

One readily deduces from the definitions of $r_1(\alpha)$ and $r_2(\alpha)$ that $\cap_{\alpha\geq\beta}C_\alpha(r_1)=C_\beta(\delta)\cap\{z\,|\,\mathrm{Re}z\geq m_1\}$ and $\cap_{\alpha\leq\beta}C_\alpha(r_2)=C_\beta(\delta)\cap\{z\,|\,\mathrm{Re}z\leq m_2\}$. By Eqs. (15)–(16),

$$S(\delta)\subset[\cap_{\alpha\geq\beta}C_\alpha(r_1)]\cap[\cap_{\alpha\leq\beta}C_\alpha(r_2)]=C_\beta(\delta)\cap M.$$

Thus $S(\delta)$ is the sharpest result given by Theorem 1, as it is a subset of every set constructed from intersections of M and the sets $C_\alpha(\delta)$.

III. APPLICATIONS

A. Plane parallel flow of an inviscid incompressible fluid

We consider an infinite horizontal layer of inviscid, incompressible, stratified fluid occupying the region $\{-\infty<x<\infty,\ y_1<y<y_2,\ -\infty<z<\infty\}$, which is bounded above and below by rigid walls at $y=y_2$ and $y=y_1$. The fluid is subjected to a constant vertical gravitational field $-g\mathbf{e}_y$, with unperturbed mass density $\rho_0(y)$ and unperturbed fluid velocity $U(y)\,\mathbf{e}_x$. We assume that $\rho_0(y)$ and $U(y)$ are twice continuously differentiable on $[y_1,y_2]$ and that $\rho_0(y)>0$ on $[y_1,y_2]$.

The consideration of small two-dimensional (no z dependence) disturbances of the fluid about the unperturbed state described above leads to the equation[23,24]

$$P\ddot\xi+A\dot\xi+H_t(y,t;k)=0, \quad (17)$$

where $\xi(y,t;k)$ is the Fourier transform of the y component of the displacement vector $\eta(x,y,t)$, k is the wavenumber (real), $P=-\partial\rho_0\partial+k^2\rho_0$, $A=2ik(-\partial\rho_0U\partial+k^2\rho_0U)$, $H=k^2(\partial\rho_0U^2\partial-k^2\rho_0U^2-g\rho_0')$, $\partial\equiv\partial/\partial y$, $'\equiv d/dy$, and $\cdot\equiv\partial/\partial t$. The quantity ξ satisfies the boundary conditions $\xi(y_1,t;k)=0=\xi(y_2,t;k)$. Accordingly, we take $D_P=D_A=D_H=D=\{f(y)\,|\,f\in C^2[y_1,y_2],\ f(y_1)=f(y_2)=0\}$. The operators P, A, and H map D into the inner product space $E=C[y_1,y_2]$ with inner product $(f,g)=\int_{y_1}^{y_2}\bar f(y)g(y)\,dy$. $D\subset E$, P, iA, and H are all formally self-adjoint on D, and P is positive definite on D.

Let k be fixed, $k\neq0$ ($k=0$ implies $A=H=0$ and $\Delta=0$), $c\in R$, $\phi\in D$, $\phi\neq0$. A straightforward integration by parts yields

$$\frac{(\phi,-H_{-kc}\phi)}{(\phi,P\phi)}=\frac{k^2\int_{y_1}^{y_2}[(U-c)^2Q(\phi)+g\rho_0'|\phi|^2]\,dy}{\int_{y_1}^{y_2}Q(\phi)\,dy}, \quad (18)$$

where $Q(\phi)\equiv\rho_0[|\phi'|^2+k^2|\phi|^2]$ and $\int_{y_1}^{y_2}Q(\phi)\,dy>0$. Define $F(c)\equiv\sup_D(\phi,-H_{-kc}\phi)/(\phi,P\phi)$, $-\infty<c<\infty$. It follows from Theorem 1 that for each $c\in R$, every eigenvalue ω of Eq. (17) with $\mathrm{Im}\omega\neq0$ lies within the circle C_{-kc} with center $(-kc,0)$ and radius $[F(c)]^{1/2}$ [$F(c)\leq0$ for some c implies that $\Delta\geq0$ and all eigenvalues must be real]. Let a and b denote the minimum and maximum values, respectively, of $U(y)$ on $[y_1,y_2]$. It is readily seen that

$$\sup_D\frac{\int_{y_1}^{y_2}(U-c)^2Q(\phi)\,dy}{\int_{y_1}^{y_2}Q(\phi)\,dy}=G(c)$$

$$\equiv\begin{cases}(b-c)^2, & c\leq\dfrac{a+b}{2},\\[2ex](c-a)^2, & c\geq\dfrac{a+b}{2}.\end{cases} \quad (19)$$

The minimum value of the function $G(c)$ is $[(b-a)/2]^2=G((a+b)/2)$.

If $g\rho_0'(y)\leq0$ for all $y\in[y_1,y_2]$, Eq. (18) implies that $F(c)\leq k^2G(c)$ for all $c\in R$, and therefore every complex eigenvalue ω lies within the circle with center $(-k(a+b)/2,0)$ and radius $|k|(b-a)/2$. This is the celebrated circle theorem of Howard.

Suppose that $g\rho_0'\equiv0$ on $[y_1,y_2]$ and that $U(y)$ is nonconstant on every open subinterval of $[y_1,y_2]$. Then $F(c)\equiv k^2G(c)$, H_α is negative definite on D for all $\alpha\in R$, and every nonzero $\phi\in D$ belongs to $\tilde D$. The function δ defined in Theorem 1 thus satisfies $\delta^2(-kc)=k^2G(c)$ for every $c\in R$, so that the circle of minimum radius $C_\beta(\delta)$ defined in Theorem 1 is precisely the circle constructed above. Furthermore, it follows easily from the definition of $G(c)$ that $S(\delta)\equiv\cap_{\alpha\in R}C_\alpha(\delta)=C_\beta(\delta)$.

In the circumstance that $g\rho_0'(y) > 0$ for some $y \in [y_1, y_2]$, Eq. (18) implies that $F(c) \leq k^2[G(c) + \lambda^{-1}M]$ for all $c \in R$, where $M \equiv \max_{[y_1,y_2]}[g\rho_0'(y)/\rho_0(y)] > 0$ and

$$\lambda \equiv \min_{\phi \in D} \frac{\int_{y_1}^{y_2} Q(\phi)\, dy}{\int_{y_1}^{y_2} \rho_0 |\phi|^2\, dy} > k^2.$$

Thus every complex eigenvalue lies in the circle with center $(-k(a+b)/2, 0)$ and radius $|k|[((b-a)/2)^2 + \lambda^{-1}M]^{1/2}$.

B. Plane parallel flow of an N-component plasma

We consider a plasma consisting of N species of charged particles confined to the region $\Omega = \{0 < x < a, -\infty < y < \infty, -\infty < z < \infty\}$ by perfectly reflecting walls at $x = 0$ and $x = a$, and satisfying therein the following set of equations (macroscopic plasma description, electrostatic approximation):

$$\frac{\partial \rho_j}{\partial t} + \nabla \cdot (\rho_j \mathbf{V}_j) = 0, \quad 1 \leq j \leq N, \tag{20}$$

$$\rho_j \left\{ \frac{\partial \mathbf{V}_j}{\partial t} + (\mathbf{V}_j \cdot \nabla)\mathbf{V}_j \right\} = -\nabla \Pi_j - q_j \rho_j \nabla \Phi, \quad 1 \leq j \leq N, \tag{21}$$

$$\left(\frac{\partial}{\partial t} + \mathbf{V}_j \cdot \nabla \right) \Pi_j \rho_j^{-\gamma_j} = 0, \quad 1 \leq j \leq N, \tag{22}$$

$$\nabla^2 \Phi = -\epsilon_0^{-1}\left[\sigma_0(x) + \sum_{j=1}^{N} q_j \rho_j \right]. \tag{23}$$

The quantity $\rho_j(\mathbf{x}, t)$ denotes the mass density of the jth species, $\mathbf{V}_j(\mathbf{x}, t)$ the mean velocity, $\Pi_j(\mathbf{x}, t)$ the pressure, q_j the charge-to-mass ratio (a constant), γ_j the ratio of specific heats (a constant ≥ 1), $\Phi(\mathbf{x}, t)$ the electric potential, and $\sigma_0(x)$ the fixed background charge density. Poisson's equation (23) is written in mks units; ϵ_0 is the permittivity of free space. The jth plasma component is assumed to have the temperature T_j in the unperturbed state and to be in steady flow along the z axis. The unperturbed quantities (distinguished by a 0 subscript) are independent of t and the coordinates y and z, and are allowed to be twice continuously differentiable functions of x on $[0, a]$, consistent with Eqs. (20)–(23), with $\rho_{j0}(x) > 0$ on $[0, a]$, $1 \leq j \leq N$. Thus

$$\mathbf{V}_{j0} = v_{j0}(x)\mathbf{e}_z, \qquad 1 \leq j \leq N, \tag{24}$$

$$\Pi_{j0}(x) = c_j^2 \rho_{j0}(x), \quad 1 \leq j \leq N, \tag{25}$$

where c_j is the thermal velocity of the jth species of a particle,

$$\rho_{j0}(x) = \hat{\rho}_j \exp[-q_j \Phi_0(x)/c_j^2], \quad 1 \leq j \leq N, \tag{26}$$

where $\hat{\rho}_j$ is a constant, and

$$\frac{d^2\Phi_0}{dx^2} = -\epsilon_0^{-1}[\sigma_0 + \sum_j q_j \rho_{j0}]. \tag{27}$$

We assume small motions about the unperturbed state described above, and linearize Eqs. (20)–(23) about that state. The resulting equations, when written in terms of the Lagrangian displacement vector $\boldsymbol{\xi}(\mathbf{x}, t)$, take the form

$$\rho_{j0}\left(\frac{\partial}{\partial t} + v_{j0}\frac{\partial}{\partial z} \right)^2 \boldsymbol{\xi}_j - (\gamma_j - 1)\, c_j^2 \nabla(\rho_{j0} \nabla \cdot \boldsymbol{\xi}_j)$$

$$- c_j^2 \rho_{j0} \nabla[\rho_{j0}^{-1} \nabla \cdot (\rho_{j0}\boldsymbol{\xi}_j)] + q_j \rho_{j0} \nabla \Phi_1 = 0, \quad 1 \leq j \leq N \tag{28}$$

$$\nabla^2 \Phi_1 = \epsilon_0^{-1} \sum_j q_j \nabla \cdot (\rho_{j0}\boldsymbol{\xi}_j), \tag{29}$$

where the subscript 1 denotes a perturbed quantity. The boundary conditions imposed on $\boldsymbol{\xi}$ and Φ_1 are that

$$\xi_{jx}(0, y, z, t) = 0 = \xi_{jx}(a, y, z, t), \quad 1 \leq j \leq N \tag{30}$$

and that Φ_1 and $\partial\Phi_1/\partial x$ be continuous across the planes $x = 0$ and $x = a$. Outside the region Ω, $\nabla^2\Phi_1 = 0$, and we require that $\Phi_1 \to 0$ as $|x| \to \infty$. We look for solutions of the form $\boldsymbol{\xi}_j(\mathbf{x}, t) = \boldsymbol{\eta}_j(x, t; \mathbf{k}) \exp(i\mathbf{k} \cdot \mathbf{x})$, $\Phi_1 = \phi(x, t; \mathbf{k}) \times \exp(i\mathbf{k} \cdot \mathbf{x})$, where $\mathbf{k} = k_y\mathbf{e}_y + k_z\mathbf{e}_z$, and obtain the following set of equations (*Note*: we use the abbreviated notation $\rho_j \equiv \rho_{j0}$, $v_j \equiv v_{j0}$ henceforth):

$$\rho_j\left(\frac{\partial}{\partial t} + ik_z v_j \right)^2 \eta_{j1} - c_j^2(\gamma_j - 1)[\rho_j(\eta_{j1}' + ik\eta_{j2})]'$$

$$- c_j^2 \rho_j \{\rho_j^{-1}[(\rho_j \eta_{j1})' + ik\rho_j \eta_{j2}]\}'$$

$$+ q_j \rho_j \phi' = 0, \quad 1 \leq j \leq N, \tag{31}$$

$$\rho_j\left(\frac{\partial}{\partial t} + ik_z v_j \right)^2 \eta_{j2} - ikc_j^2\,(\gamma_j - 1)\,\rho_j(\eta_{j1}' + ik\eta_{j2})$$

$$- ikc_j^2[(\rho_j \eta_{j1})' + ik\rho_j \eta_{j2}] + ikq_j\rho_j\phi = 0, \quad i \leq j \leq N, \tag{32}$$

$$\phi'' - k^2\phi = \epsilon_0^{-1} \sum_j q_j[(\rho_j \eta_{j1})' + ik\rho_j \eta_{j2}], \tag{33}$$

where $\eta_{j1} \equiv \eta_{jx}$, $\eta_{j2} \equiv k^{-1}(k_y\eta_{jy} + k_z\eta_{jz})$, $k = (k_y^2 + k_z^2)^{1/2} > 0$, and $' \equiv \partial/\partial x$. The boundary conditions on ϕ imply that $\phi'(0, t; \mathbf{k}) - k\phi(0, t; \mathbf{k}) = 0$ and $\phi'(a, t; \mathbf{k}) + k\phi(a, t; \mathbf{k}) = 0$. The differential operator $L \equiv -d^2/dx^2 + k^2$ with domain $D_L = \{f(x) \mid f \in C^2[0, a],\ f'(0) - kf(0) = 0 = f'(a) + kf(a)\}$ admits the positive compact Hermitian inverse Γ (an integral operator whose kernel is the Green's function for L) which maps $C[0, a]$ into D_L. With the aid of Γ, Eqs. (31)–(33) take the following matrix form in the inner product space

$E = \{\underset{\sim}{F}(x) \mid \underset{\sim}{F}(x) = (\underset{\sim}{f}_1(x), \underset{\sim}{f}_2(x), \ldots, \underset{\sim}{f}_N(x))^T,\ \underset{\sim}{f}_j(x) = (f_{j1}(x), f_{j2}(x))^T \in C[0, a] \times C[0, a],\ 1 \leq j \leq N\}$:

$$P\ddot{\underset{\sim}{\xi}} + A\dot{\underset{\sim}{\xi}} + H\underset{\sim}{\xi}(x, t; \mathbf{k}) = 0, \tag{34}$$

$$P = \begin{pmatrix} P_1 & & 0 \\ & \ddots & \\ 0 & & P_N \end{pmatrix}, \quad A = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_N \end{pmatrix}, \quad H = (H_{jm}), \quad P_j = \rho_j \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_j = 2ik_z v_j P_j,$$

$$H_{jj}=\begin{pmatrix} \gamma_j c_j^2[-\partial\rho_j\partial+\rho_j\Psi_j'']-k_z^2\rho_j v_j^2-\frac{q_j^2}{\epsilon_0}\rho_j\partial\Gamma\partial\rho_j & ikc_j^2[\rho_j'-\gamma_j\partial\rho_j]-\frac{q_j^2}{\epsilon_0}ik\rho_j\partial\Gamma\rho_j \\ -ikc_j^2[\gamma_j\partial\rho_j-(\gamma_j-1)\rho_j']-\frac{q_j^2}{\epsilon_0}ik\rho_j\Gamma\partial\rho_j & (\gamma_j c_j^2k^2-k_z^2v_j^2)\rho_j+\frac{q_j^2}{\epsilon_0}k^2\rho_j\Gamma\rho_j \end{pmatrix},$$

$$H_{jm}=-\epsilon_0^{-1}q_jq_m\rho_j\begin{pmatrix}\partial\Gamma\partial\rho_m & ik\partial\Gamma\rho_m \\ ik\Gamma\partial\rho_m & -k^2\Gamma\rho_m\end{pmatrix},\quad j\neq m,\quad \partial\equiv\frac{\partial}{\partial x},\quad \Psi_j\equiv-\gamma_j^{-1}\ln(\rho_j/\hat{\rho}_j),\quad D_P=D_A=E,$$

$$D=D_H=\{\underset{\sim}{F}\mid \underset{\sim}{F}\in E,\ f_{j1}\in C^2[0,a],\ f_{j1}(0)=0=f_{j1}(a),\ 1\leq j\leq N\}.$$

The inner product on E is given by

$$(\underset{\sim}{F},\underset{\sim}{G})=\sum_{j=1}^{N}\int_0^a \underset{\sim}{f}_j^{*}\cdot\underset{\sim}{g}_j\,dx=\sum_{j=1}^{N}\int_0^a[\bar{f}_{j1}g_{j1}+\bar{f}_{j2}g_{j2}]\,dx.$$

The operators P, iA, and H are formally self-adjoint on their domains of definition and hence on D, and $P>0$. Let $\mu\in R$ and $\underset{\sim}{F}$ be a nonzero element of D. A straightforward computation yields

$$\frac{(\underset{\sim}{F},-H_{-k_z\mu}\underset{\sim}{F})}{(\underset{\sim}{F},P\underset{\sim}{F})}=\frac{k_z^2\sum_j\int_0^a(v_j-\mu)^2\rho_j|f_j|^2\,dx-\epsilon_0^{-1}\int_0^a(\sum_j q_j\bar{r}_j)\Gamma(\sum_j q_jr_j)\,dx}{\sum_j\int_0^a\rho_j|\underset{\sim}{f}_j|^2\,dx}$$

$$-\frac{\sum_j c_j^2\int_0^a[(\gamma_j-1)\rho_j|f_{j1}'+\gamma ikf_{j2}|^2+\rho_j^{-1}|r_j|^2]\,dx}{\sum_j\int_0^a\rho_j|\underset{\sim}{f}_j|^2\,dx}\leq k_z^2\frac{\sum_j\int_0^a(v_j-\mu)^2\rho_j|\underset{\sim}{f}_j|^2\,dx}{\sum_j\int_0^a\rho_j|\underset{\sim}{f}_j|^2\,dx},$$

where $r_j\equiv(\rho_jf_{j1})'+ik\rho_jf_{j2}$. Let $u_j\equiv\min_{[0,a]}v_j(x)$, $U_j\equiv\max_{[0,a]}v_j(x)$, $u\equiv\min_j u_j$, $U\equiv\max_j U_j$. Then for $\mu\in R$,

$$\sup_{\underset{\sim}{F}\in D}\frac{\sum_j\int_0^a(v_j-\mu)^2\rho_j|f_j|^2\,dx}{\sum_j\int_0^a\rho_j|\underset{\sim}{f}_j|^2\,dx}=h(\mu)\equiv\begin{cases}(U-\mu)^2, & \mu\leq(u+U)/2,\\ (\mu-u)^2, & \mu\geq(u+U)/2.\end{cases}$$

The minimum value of $h(\mu)$ is $[(U-u)/2]^2=h((u+U)/2)$. It follows that $\sup_D[(\underset{\sim}{F},-H_{-k_z\mu}\underset{\sim}{F})/(\underset{\sim}{F},P\underset{\sim}{F})]\leq k_z^2h(\mu)$, $-\infty<\mu<\infty$, and Theorem 1 implies that every eigenvalue ω of the system (34) with $\mathrm{Im}\omega\neq0$ lies in the circle with center $(-k_z(u+U)/2,0)$ and radius $|k_z|(U-u)/2$.

ACKNOWLEDGMENT

The work presented here was supported by the National Science Foundation under Research Grant MPS75-07391.

APPENDIX

Theorem: Let P, A, H, H_ω, D, d, and Δ be defined as in the first two paragraphs of Sec. II. Suppose that $-\infty<\Delta<\infty$. Then

$$-\Delta=\inf_{\alpha\in R}\sup_D\frac{(\zeta,-H_\alpha\zeta)}{(\zeta,P\zeta)}=\sup_D\frac{(\zeta,-H_\beta\zeta)}{(\zeta,P\zeta)},\tag{A1}$$

where

$$\beta\equiv\inf_D\left\{\frac{(\zeta,iA\zeta)}{2(\zeta,P\zeta)}+[d(\zeta)-\Delta]^{1/2}\right\}$$

$$=\sup_D\left\{\frac{(\zeta,iA\zeta)}{2(\zeta,P\zeta)}-[d(\zeta)-\Delta]^{1/2}\right\}.\tag{A2}$$

Furthermore, if $\{\zeta_n\}_{n=1}^{\infty}$ is any sequence from D such that $d(\zeta_n)\to\Delta$, then $(\zeta_n,iA\zeta_n)/(\zeta_n,P\zeta_n)\to2\beta$. If $-\infty<\Delta<0$, then

$$-\Delta=\inf_{\alpha\in R}\sup_{\hat{D}}\frac{(\zeta,-H_\alpha\zeta)}{(\zeta,P\zeta)}=\sup_{\hat{D}}\frac{(\zeta,-H_\beta\zeta)}{(\zeta,P\zeta)}.\tag{A3}$$

Proof: Let $\Delta\in R$. For $\alpha\in R$ and nonzero $\zeta\in D$, we define the real-valued function

$$f_\zeta(\alpha)\equiv\frac{(\zeta,-H_\alpha\zeta)}{(\zeta,P\zeta)}+\Delta=\frac{(\zeta,[-H_\alpha+\Delta P]\zeta)}{(\zeta,P\zeta)}$$

$$=\alpha^2-\alpha\frac{(\zeta,iA\zeta)}{(\zeta,P\zeta)}-\frac{(\zeta,\hat{H}\zeta)}{(\zeta,P\zeta)},$$

where $\hat{H}\equiv H-\Delta P$ on D. For each fixed nonzero $\zeta\in D$, the graph of $y=f_\zeta(\alpha)$ in the αy plane is a parabola that is concave up, with axis of symmetry parallel to the y axis and vertex at $((\zeta,iA\zeta)/2(\zeta,P\zeta),-\hat{d}(\zeta))$, where $\hat{d}(\zeta)\equiv d(\zeta)-\Delta$. Since $\Delta\equiv\inf_D d(\zeta)$, $0=\inf_D\hat{d}(\zeta)=-\sup_D[-\hat{d}(\zeta)]$; hence the vertex of each such parabola lies on or below the α axis, and the zeros of $f_\zeta(\alpha)$ are real and are given by $\hat{Q}_\pm(\zeta)\equiv(\zeta,iA\zeta)/2(\zeta,P\zeta)\pm[\hat{d}(\zeta)]^{1/2}$. Let η and ζ be two nonzero elements of D. We show that $\hat{Q}_+(\zeta)\geq\hat{Q}_-(\eta)$. If η and ζ are linearly dependent, then $\eta=z\zeta$ for some nonzero $z\in Z$, so that $\hat{Q}_+(\zeta)-\hat{Q}_-(\eta)=\hat{Q}_+(\zeta)-\hat{Q}_-(\zeta)=2\hat{d}(\zeta)\geq0$. Suppose that η and ζ are linearly independent. Then $S\equiv\mathrm{span}\{\eta,\zeta\}$ is a two-dimensional subspace of D. Let ϕ_1 and ϕ_2 be an orthonormal basis for S, and let Π be

the projection operator from E into S defined by $\Pi\xi = (\phi_1, \xi)\phi_1 + (\phi_2, \xi)\phi_2$ for all $\xi \in E$. The operators $a \equiv \Pi P \Pi$, $b \equiv \Pi i A \Pi$, and $c \equiv \Pi \hat{H} \Pi$ are Hermitian on D and map S into itself, with $a > 0$ on S. For any nonzero $\xi \in S$, we have $\hat{Q}_\pm(\xi) = (\xi, b\xi)/2(\xi, a\xi) \pm [\hat{d}(\xi)]^{1/2}$ and $0 \leq \hat{d}(\xi) = (\xi, b\xi)^2/4(\xi, a\xi)^2 + (\xi, c\xi)/(\xi, a\xi)$, so that by Lemma 3 of Ref. 25 and Theorem 8 of Ref. 26, $\sup_S \hat{Q}_-(\xi) \leq \inf_S \hat{Q}_+(\xi)$. Therefore, $\hat{Q}_+(\zeta) \geq \hat{Q}_-(\eta)$, as we wished to prove. Since η and ζ were arbitrary nonzero elements of D, we conclude that $-\infty < \sup_D \hat{Q}_-(\xi) \leq \inf_D \hat{Q}_+(\xi) < \infty$. Let $\epsilon > 0$. Since $\inf_D \hat{d}(\xi) = 0$, there exists a nonzero $\zeta \in D$ such that $0 \leq \hat{d}(\zeta) < \epsilon$, and therefore $\hat{Q}_+(\zeta) = \hat{Q}_-(\zeta) + 2[\hat{d}(\zeta)]^{1/2} < \hat{Q}_-(\zeta) + 2\epsilon^{1/2}$. Hence $\inf_D \hat{Q}_+(\xi) < \sup_D \hat{Q}_-(\xi) + 2\epsilon^{1/2}$. As this holds for arbitrary $\epsilon > 0$, $\inf_D \hat{Q}_+(\xi) \leq \sup_D \hat{Q}_-(\xi)$, which, together with our previous result, implies that $\inf_D \hat{Q}_+(\xi) = \sup_D \hat{Q}_-(\xi)$. Let $\beta \equiv \inf_D \hat{Q}_+(\xi) = \sup_D \hat{Q}_-(\xi)$ [note that this is Eq. (A2)], and suppose that $\{\zeta_n\}_{n=1}^{\infty}$ is a sequence of nonzero elements of D such that $d(\zeta_n) \to \Delta$. Then $\hat{d}(\zeta_n) \to 0$, and for all $n = 1, 2, 3, \cdots$,

$$\frac{(\zeta_n, iA\zeta_n)}{2(\zeta_n, P\zeta_n)} - [\hat{d}(\zeta_n)]^{1/2} \leq \beta \leq \frac{(\zeta_n, iA\zeta_n)}{2(\zeta_n, P\zeta_n)} + [\hat{d}(\zeta_n)]^{1/2},$$

which implies that $(\zeta_n, iA\zeta_n)/(\zeta_n, P\zeta_n) \to 2\beta$. We proceed to establish (A1). Let ζ be any nonzero element of D. Then $\hat{Q}_-(\zeta) \leq \beta \leq \hat{Q}_+(\zeta)$, and $f_\zeta(\alpha) \leq 0$ for all $\alpha \in [\hat{Q}_-(\zeta), \hat{Q}_+(\zeta)]$. Thus $f_\zeta(\beta) \leq 0$ for all nonzero $\zeta \in D$, and therefore

$$\sup_D \frac{(\zeta, -H_\beta \zeta)}{(\zeta, P\zeta)} = \sup_D f_\zeta(\beta) - \Delta \leq -\Delta. \tag{A4}$$

For each nonzero $\zeta \in D$, $\min_{\alpha \in R} f_\zeta(\alpha) = -\hat{d}(\zeta)$, so that $f_\zeta(\alpha) \geq -\hat{d}(\zeta)$ for all $\alpha \in R$. Hence $\sup_D f_\zeta(\alpha) \geq -\inf_D \hat{d}(\zeta) = 0$, i.e., $\sup_D (\zeta, -H_\alpha\zeta)/(\zeta, P\zeta) \geq -\Delta$ for all $\alpha \in R$, and therefore $\inf_{\alpha \in R} \sup_D (\zeta, -H_\alpha\zeta)/(\zeta, P\zeta) \geq -\Delta$. Equation (A1) now follows from Eq. (A4). There remains only to prove Eq. (A3). Suppose $\Delta < 0$. Then $\tilde{D}$ is nonempty, and for each nonzero $\zeta \in \tilde{D} \subset D$, $f_\zeta(\alpha) \geq -\hat{d}(\zeta)$ for all $\alpha \in R$. Hence $\sup_{\tilde{D}} f_\zeta(\alpha) \geq -\inf_{\tilde{D}} \hat{d}(\xi) = -\inf_{\tilde{D}} d(\xi) + \Delta = 0$, i.e., $\sup_{\tilde{D}} (\zeta, -H_\alpha\zeta)/(\zeta, P\zeta) \geq -\Delta$ for all $\alpha \in R$. Thus $\inf_{\alpha \in R} \sup_{\tilde{D}} (\zeta, -H_\alpha\zeta)/(\zeta, P\zeta) \geq -\Delta$. Since $\tilde{D} \subset D$, $\sup_{\tilde{D}} (\zeta, -H_\beta\zeta)/(\zeta, P\zeta) \leq \sup_D (\zeta, -H_\beta\zeta)/(\zeta, P\zeta) = -\Delta$, which implies Eq. (A3).

[1] L.N. Howard, J. Fluid. Mech. 10, 509 (1961).
[2] C. Eckart, Phys. Fluids 6, 1042 (1963).
[3] M.E. Stern, Phys. Fluids 6, 636 (1963).
[4] J. Pedlosky, J. Atmos. Sci. 21, 201 (1964).
[5] T.J. Eisler, Phys. Fluids 8, 1635 (1965).
[6] P.G. Gilman, J. Atmos. Sci. 24, 101 (1967).
[7] G.S. Agrawal, J. Phys. Soc. Jpn. 26, 1519 (1969).
[8] L.N. Howard, Stud. Appl. Math. 52, 39 (1973).
[9] D.J. Acheson, J. Fluid Mech. 61, 609 (1973).
[10] W. Blumen, J. Fluid Mech. 68, 177 (1975).
[11] D.P. Lalas, J. Fluid Mech. 69, 65 (1975).
[12] E.T. Routh, *The Advanced Part of a Treatise on the Dynamics of a System of Rigid Bodies* (Macmillan, New York, 1905), 6th ed., Chaps. 3, 6, 7.
[13] E.T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies* (Cambridge U.P., New York, 1937), 4th ed., Chap. 7.
[14] P.A. Sturrock, Ann. Phys. (N.Y.) 4, 306 (1958).
[15] F.E. Low, Proc. R. Soc. **A248**, 283 (1958).
[16] C. Eckart, Phys. Fluids 3, 421 (1960).
[17] E.A. Frieman and M. Rotenberg, Rev. Mod. Phys. 32, 898 (1960).
[18] F.E. Low, Phys. Fluids 4, 842 (1961).
[19] W.A. Newcomb, Nucl. Fusion Suppl., Part 2, 451 (1962).
[20] E.M. Barston, J. Math. Phys. 8, 523 (1967).
[21] E.M. Barston, J. Math. Phys. 15, 675 (1974).
[22] E. Hamieri, private communication.
[23] J.W. Miles, J. Fluid Mech. 10, 496 (1961).
[24] C.-S. Yih, *Dynamics of Nonhomogeneous Fluids* (Macmillan, New York, 1965), Chap. 4.
[25] E.M. Barston, Int. J. Eng. Sci. 10, 947 (1972).
[26] E.M. Barston, Int. J. Eng. Sci. 12, 413 (1974).

The Zeno's paradox in quantum theory

B. Misra and E. C. G. Sudarshan*

Center for Particle Theory, University of Texas at Austin, Austin, Texas 78712
(Received 24 February 1976)

We seek a quantum-theoretic expression for the probability that an unstable particle prepared initially in a well defined state ρ will be found to decay *sometime during a given interval.* It is argued that probabilities like this which pertain to continuous monitoring possess operational meaning. A simple natural approach to this problem leads to the conclusion that an unstable particle which is continuously observed to see whether it decays will never be found to decay! Since recording the track of an unstable particle (which can be distinguished from its decay products) *approximately* realizes such continuous observations, the above conclusion seems to pose a paradox which we call Zeno's paradox in quantum theory. The relation of this result to that of some previous works and its implications and possible resolutions are briefly discussed. The mathematical transcription of the above-mentioned conclusion is a structure theorem concerning semigroups. Although special cases of this theorem are known, the general formulation and the proof given here are believed to be new. We also note that the known "no-go" theorem concerning the semigroup law for the reduced evolution of any physical system (including decaying systems) is subsumed under our theorem as a direct corollary.

1. INTRODUCTION

The object of this paper is to discuss a seemingly paradoxical result in quantum theory concerning temporal evolution of a dynamical system under continuous observation during a period of time. For reasons that will become clear shortly we call this complex of deductions *Zeno's paradox in quantum theory.*

Let us consider schematically the theory of an unstable quantum system. Naturally the states corresponding to the decay products also should be included in the space of all states which we take to be a Hilbert space $\mathcal{H}$. Let us denote the (orthogonal) projection onto the subspace spanned by the *undecayed* (unstable) states of the system by E. This projection E thus represents the observable that corresponds to the "yes-no experiment" for determining whether the system is in an *undecayed* state or in a *decayed* state. The evolution in time of the states of the total system will be described by a unitary group $U(t) \equiv \exp(-iHt)$ labeled by the real time parameter t. In this setting the quantity

$$q(t) = \mathrm{Tr}[\rho U^*(t) E U(t)] \tag{1}$$

is interpreted as the probability, *at the instant* t, for finding the system undecayed when at time 0 it was prepared in the state ρ. Correspondingly, the probability that, *at the instant* t, the system will be found to have decayed is the complementary quantity

$$p(t) = 1 - q(t) = \mathrm{Tr}[\rho U^*(t) E^{\perp} U(t)], \quad E^{\perp} = I - E. \tag{2}$$

All these are, of course, standard.

Quantum theory, in fact, provides an unambiguous algorithm for computing the probability distributions of *time*, given the knowledge of the initial state of the of) *time,* given the knowledge of the initial state of the same system and its law of time evolution. Expressions (1) and (2) are only particular instances of this well-known algorithm.

In contrast to the above-mentioned probabilities which refer to a specified instant of time we may consider the following probabilities for which quantum theory has no ready expressions:

(1) The probability that the system prepared in the undecayed state ρ at time 0 is found to decay *sometime during the interval* $\Delta = [0, t]$. We denote this by $P(0, t; \rho)$.

(2) The probability $Q(0, t; \rho)$ that no decay is found *throughout the interval* Δ when the initial state of the system was known to be ρ.

(3) The probability that the system prepared initially in the state ρ will be found to be undecayed *throughout* $[0, t_1] \equiv \Delta$, but found to decay *sometime* during the subsequent period $[t_1, t] \equiv \Delta_2$, $0 < t_1 < t$. We denote this by $R(0, t_1, t; \rho)$.

It is important to distinguish the probabilities $Q(0, t; \rho)$ from $q(t)$ since there is the temptation to identify them [and hence also $P(0, t; \rho)$ with $p(t)$].[1] The probability $q(t)$, however, refers to outcomes of measurement of E at the time t, *the system being left unobserved* after the initial state preparation until t.

The operational meaning (if any!) of the probabilities P, Q, R on the other hand is to be found in terms of the outcomes of continuously ongoing measurement of E during the entire interval of time Δ. The notion of such continuously ongoing observations (or, equivalently, measurements) is obviously an idealization.

We may consider the process of continuing observation as the limiting case of successions of (practically) instantaneous measurements (of E) as the intervals between successive measurements approach zero. Since there does not seem to be any principle, internal to quantum theory, that forbids the duration of a single measurement or the dead time between successive measurements from being arbitrarily small, the process of continuous observation seems to be an admissible process in quantum theory.

It may be argued, however, that what are measurable or not are governed not only by the fundamental principles of quantum theory but also by the actual constituents of the real world and their interactions.

 Copyright © 1977 American Institute of Physics

The concept of continuous observation would indeed be bereft of any physical meaning *if* it could be established that the fundamental constituents of the real world and the interaction between them are such as to exclude the possibility of arbitrarily frequent observations. But, on the one hand, we cannot claim as final our present knowledge of the constituents and interactions of the real world. On the other hand, to agree that there is a limitation on the frequency of observation amounts to claiming the existence of an elementary and indivisible unit of time. Though the existence of an elementary interval of time is an exciting possibility, it is not part of the currently accepted and tested physical theories.

We, thus, feel that the notion of continuous observation should be accepted, at least for the present, as physically meaningful and quantum theory should be pressed to yield an answer to questions relating the probabilities pertaining to such observations.

Continuous observation processes seem to be realized in practice also, at least approximately, by the tracks of unstable charged particles in bubble chambers and other detecting media. The observation of the track amounts practically to a more or less continuous monitoring of the existence of the unstable particle and thus a measurement of E during the period of the particle's flight through the detection chamber. We are therefore led to accept as operationally meaningful the $P(0, t;\rho)$, $Q(0, t;\rho)$, and $R(0, t_1, t;\rho)$. To be a *complete* theory, quantum theory must provide an algorithm for computing these probabilities.

In the next section we describe what appears to be the natural approach to determining quantum-theoretic expressions for these probabilities. Our investigation leads to the paradoxical result mentioned at the beginning of this section: An unstable particle observed continuously whether it has decayed or not will *never* be found to decay! Since this evokes the famous paradox of Zeno denying the possibility of motion to a flying arrow, we call this result the Zeno's paradox in quantum theory

In fact, if E is taken to be the projection to the set of localized states of a particle (or, a quantum arrow) in a given region D of space, then one concludes that the particle will never be found to arrive in a disjoint region D' provided it is *continuously observed* whether it has entered D' or not: The "arrow" cannot move to where it is not!

This result acquires an even more picturesque and paradoxical formulation when it is applied to the "hellish contraption" considered in the Schrödinger's cat paradox.[2] It may be recalled that the contraption consists of an unstable (quantum) particle placed in a box equipped with an efficient counter and a cat inside a steel chamber. If the particle decays, the counter triggers and, in its turn, activates a tiny hammer which breaks a container of cyanide in the steel chamber. Monitoring the vital functions of the cat amounts to observing if the particle has decayed or not. In view of the Zeno's paradox formulated above, should we conclude that the particle will never decay? Will the cat escape the cruel death awaiting it, against which it has no defense, provided its vital signs are constantly watched with loving care?

The mathematical transcription of Zeno's paradox is a structure theorem concerning a class of strongly continuous semigroups. This theorem is formulated and proved in Sec. 3 of this paper and may possess some intrinsic interest apart from its application in the present context of a theory of continuous observation. As a by-product of this investigation we find that the known result[3] concerning the incompatibility of the semiboundedness of the Hamiltonian H with the requirement that $E\exp(iHt)E$ form a strictly contractive semigroup can be subsumed under the above-mentioned theorem as one of its direct corollaries.

Some of the implications and possible resolutions of the quantum Zeno's paradox will be briefly discussed in the concluding section of the paper.

Finally it may be mentioned that the conclusion called here the Zeno's paradox in quantum theory has been noted in some previous works,[4-6] but the present analysis of the problem is carried out in a more general and mathematically rigorous setting than the previous works. This, we feel, is not merely a dispensable luxury, but is necessary to locate the precise assumptions on which the "Zeno's paradox" rests.

2. QUANTUM THEORETICAL EXPRESSIONS FOR $P(0, t;\rho)$ AND RELATED PROBABILITIES

The three probability functions P, Q, R introduced in the previous section relate to the results of continuous observation throughout an interval of time. By their very definitions they must obey the relations

$$P(0, t;\rho) + Q(0, t;\rho) = 1$$

and

$$R(0, t_1, t;\rho) = Q(0, t_1;\rho)\, P(0, t - t_1;\rho_1),$$

where ρ_1 is the state in which the system (prepared initially in the state ρ) finds itself at t_1 after being continuously observed *and* found to be undecayed throughout $[0, t_1]$. We may therefore concentrate our attention on calculating Q and ρ_1.

We start with the system in the state ρ and make a series of $n+1$ measurements, which are idealized to be instantaneous, at times 0, t/n, $2t/n, \ldots, (n-1)t/n$, and t. We seek the probability $Q(\Delta, n;\rho)$ that it be found undecayed in *each* of these measurements. It is natural to assume that $Q(\Delta;\rho) \equiv Q(0, t;\rho)$ can be evaluated as the limit of $Q(\Delta, n;\rho)$ when $n \to \infty$, provided the limit exists.

Let us denote by $\rho(n, t)$ the state in which the system finds itself after the $(n+1)$ measurements at 0, t/n, $2t/n, \ldots, t$ *and* being found to be undecayed in each of these measurements. Now, according to the orthodox theory of measurement, if a measurement of E on the system is carried out yielding the result "yes" (that is, "undecayed"), then the state of the system collapses to a new (unnormalized!) state ρ' of the form

$$\rho' = \sum_j A_j^* \rho A_j \tag{3}$$

with

$$\sum_j A_j^* A_j = E. \tag{4}$$

The collapsed state ρ' given by (3) is, in general, not uniquely determined by the measured observable E and the observed outcome but depends also on the details of the measuring apparatus. This circumstance is reflected in the nonuniqueness of the operators A_j satisfying (4).

The mapping (3) of the density matrices is very closely related to the "completely positive maps" defined by

$$\rho \to \sum_\alpha V_\alpha \rho V_\alpha^* = \mathcal{A}(V_\alpha)\rho,$$

$$\sum_\alpha V_\alpha^* V_\alpha = 1.$$

The "state collapse" caused by "nonselective" measurements of E is described by such maps. They will be considered in a future publication in the context of repeated and continuous nonselective measurements.

Quantum theory envisages also the possibility of *ideal* measurements under which the collapse of the state proceeds according to the simple law

$$\rho \to \rho' = E\rho E \tag{5}$$

when the measurement of E on the state ρ yields the result "undecayed."

The considerations of this paper will be restricted to such ideal measurements only, since in such cases we can exploit the positive definiteness of the Hamiltonian in a direct manner. If we were to consider the more general collapses (3) we would have to proceed more indirectly using the von Neumann–Liouville generator which is however not positive definite. The study of the probabilities $Q(\Delta, \rho)$, etc. would then involve new technical problems obscuring the essentials of Zeno's paradox. We plan to present the study of the more general situation in a subsequent paper.[7]

Accordingly, to determine $\rho(n, t)$ we allow the system to collapse at each measurement according to (5) but at the intervening time intervals it undergoes the usual Schrödinger time development. The (unnormalized) state $\rho(n, t)$ is then easily seen to be

$$\rho(n, t) = T_n(t)\rho T_n^*(t), \tag{6}$$

where

$$T_n(t) = [EU(t/n)E]^n \equiv [E\exp(-iHt/n)E]^n. \tag{7}$$

Moreover, it is also easy to show that the standard interpretation of the quantum theoretical formalism entails the formula

$$Q(n, \Delta; \rho) = \mathrm{Tr}[T_n(t)\rho T_n^*(t)]. \tag{8}$$

In fact, (8) is a special case of a more general formula for the probability connections between several successive observations.[8] It is important, however, to bear in mind that the general formula discussed in[8] [and, *a fortiori*, formula (8)] holds only under the assumption that the successive measurements under consideration are *ideal* in the sense described above. For nonideal successive measurements these formulas do not yield correct probability connections.

Returning to ideal measurements we have to proceed to the limit for $n \to \infty$. We define

$$\rho(t) = \text{s-}\lim_{n\to\infty} \rho(n, t), \tag{9}$$

$$Q(\Delta; \rho) = \lim_{n\to\infty} Q(\Delta, n; \rho), \tag{10}$$

provided the limits on the right-hand side exist. Hence, if the limit

$$\text{s-}\lim_{n\to\infty} T_n(t) \equiv \text{s-}\lim_{n\to\infty} [EU(t/n)E]^n = T(t)$$

exists for $t \geq 0$, then we may make the identification

$$\rho(t) = \{\mathrm{Tr}[T(t)\rho T^*(t)]\}^{-1} \cdot T(t)\rho T^*(t) \tag{11}$$

for the resultant (normalized) state obtained as a result of *continuous observation* and verification that the system remained *undecayed throughout the interval*. The probability $Q(\Delta; \rho)$ for this outcome is given by

$$Q(\Delta; \rho) = \lim_{n\to\infty} \mathrm{Tr}[T_n(t)\rho T_n^*(t)]$$

$$= \mathrm{Tr}[\rho T^*(t) T(t)]. \tag{12}$$

Once $Q(\Delta; \rho)$ is obtained in this manner we may calculate $P(\Delta; \rho)$ to be

$$P(\Delta; \rho) = \mathrm{Tr}[\rho(I - T^*(t))T(t)]. \tag{13}$$

For a given group $U(t)$ of time-evolution the existence of the operator $T(t)$ for $t \geq 0$ imposes a nontrivial restriction on the projection E. This restriction may be viewed as a *necessary* condition in order that the observable represented by E admits a *continuous ideal* measurement.

It is known[9] that the operators $T(t)$ (if they exist) form a strongly continuous semigroup for $t > 0$. The continuity of $T(t)$ at $t = 0$ does not generally follow from the existence of $T(t)$, but on physical grounds, we shall *assume it*;

$$\text{s-}\lim_{t\to 0_+} T(t) = E. \tag{14}$$

This condition expresses the essentially desirable requirement that the probability $Q(\Delta; \rho)$ given by (12) approaches the probability $\mathrm{Tr}(\rho E)$ as $t \to 0_+$ that the system is undecayed initially.

To prove the existence of $T(t)$ and its continuity at the origin, (14), in specific examples of physical interest poses nontrivial mathematical problems. We hope to consider these in a subsequent paper.

3. ZENO'S PARADOX IN QUANTUM THEORY

In the preceding section we arrived at formula (13) for the probability $P(\Delta; \rho)$ that the system prepared initially in the undecayed state ρ will be observed to decay *sometime* during the interval $\Delta = [0, t]$. Despite the natural derivation of (13) we now show that the probability $P(\Delta; \rho)$ vanishes for all finite intervals Δ provided that the initial state was undecayed,

$$\mathrm{Tr}(\rho E) = 1. \tag{15}$$

We are thus led to the paradoxical conclusion that an unstable particle will not decay as long as it is kept under continuous observation as to whether it decays

or not. The mathematical transcription of this statement is the following theorem.

Theorem 1: Let $U(t) \equiv \exp(-iHt)$, t real, designate a strongly continuous one-parameter group of unitary operators in the (separable) Hilbert space $\mathcal{H}$. Let E denote an orthogonal projection in $\mathcal{H}$. Assume that:

(i) The self-adjoint generator H of the group $U(t)$ is *semibounded.*

(ii) There exists an (antiunitary) operator θ such that

$$\theta E \theta^{-1} = E,$$

$$\theta U(t) \theta^{-1} = U(-t) \quad \text{for all } t.$$

(iii) $\text{s-lim}_{n\to\infty} [EU(t/n)E]^n \equiv T(t)$ exists for all $t \geq 0$.

(iv) $\text{s-lim}_{n\to 0_+} T(t) = E$.

Then $\text{s-lim}_{n\to\infty}[EU(t/n)E]^n \equiv T(t)$ exists for *all* real t and possesses the following properties:

(a) The function $t \to T(t)$ is strongly continuous and for *all* real t and s satisfies the semigroup law:

$$T(t)\,T(s) = T(t+s),$$

(b) and

$$T^*(t) = T(-t).$$

Remarks: (1) The conclusions of the theorem imply the relation:

$$T^*(t)\,T(t) = E \quad \text{for } \textit{all real } t \tag{16}$$

so that $P(\Delta;\rho) = \text{Tr}[\rho(I-E)] = 0$ for all ρ satisfying (15).

(2) With θ interpreted as the time-reversal (or CPT) operation, the assumption (ii) of the theorem turns out to be only a weak version of T or CPT invariance of the theory. Moreover it should be noted that assumption (iii) is used only once in the proof for concluding the existence of the strong limit (iii) for $t < 0$ as well.

(3) It is easy to give a relatively elementary proof of the theorem under the additional assumption that E is a one-dimensional projection onto a vector in the domain of H. The theorem is also known to hold in the special case that $H = -\nabla^2$ and E is given by

$$(E\psi)(x) = \chi(x)\,\psi(x), \quad \psi \in L^2(R^3),$$

where χ is the characteristic function of a (suitably smooth) region of R^3.[10] Theorem 1 generalizes this result to arbitrary semibounded H and arbitrary projection E.

(4) The semiboundedness of H is necessary: Consider the following counterexample. Let $V(t)$ be the operator family

$$(V(t)\,\psi) = \psi(x-t), \quad \psi \in L^2(R).$$

Let E be defined by

$$(E\psi)(x) = \begin{cases} \psi(x) & x \leq 0, \\ 0 & x > 0. \end{cases}$$

It is then easy to verify that

$$EV(t)\,EV(s)\,E = EV(t+s)\,E$$

for all $t, s \geq 0$, so that

$$\text{s-lim}_{n\to\infty}[EV(t/n)E]^n = EV(t)E = T(t)$$

for all $t \geq 0$, but

$$T^*(t)\,T(t) = EV(t^*)\,EV(t)\,E$$
$$= V^*(t)\,EV(t) \neq E \quad \text{for all } t > 0.$$

Thus the conclusion of the theorem is violated, though the assumptions in its formulation except the semiboundedness of the self-adjoint generator $V(t)$ are met. [Strictly speaking, assumption (ii) about the existence of θ is also not satisfied, but it was necessary only to prove the existence of $T(-t)$ and $T(-t)$ is trivially verified to exist in the present example.]

We now turn to the

Proof of Theorem 1: The existence of

$$T(t) \equiv \text{s-lim}_{n\to\infty}[EU(t/n)E]^n \tag{17}$$

for *all* real t follows immediately from the assumed existence of $T(t)$ for *positive* t and assumption (ii). In fact for $t \geq 0$

$$T(-t) \equiv \text{s-lim}_{n\to\infty}[EU(-t/n)E]^n$$
$$= \text{s-lim}_{n\to\infty}\theta[EU(t/n)E]^n\theta^{-1}$$
$$= \theta T(t)\,\theta^{-1}. \tag{18}$$

To prove assertion (b) we observe that

$$[EU(-t/n)E]^n = \{[EU(t/n)E]^n\}^*$$
$$\to T^*(t) \quad \text{weakly as } n \to \infty.$$

On the other hand,

$$[EU(-t/n)E]^n \to T(-t) \quad \text{strongly as } n \to \infty.$$

The assertion (b) follows immediately.

It remains to prove (a), especially assertion (16). Let us make a slight notational change and write

$$W(t) \equiv T(-t) = \text{s-lim}_{n\to\infty}[E\exp(iHt/n)E]^n. \tag{19}$$

The statement (a) can be transcribed into a corresponding statement on $W(t)$. We shall also assume, without loss of generality, that H is self-adjoint and positive.

For convenience in exposition we shall break up the proof of statement (a) into the following three lemmas:

Lemma 1: Let H be a positive self-adjoint operator of the Hilbert space $\mathcal{H}$ and let $F_n(z)$ be the operator-valued function defined by

$$F_n(z) \equiv [E\exp(iHz/n)E]^n. \tag{20}$$

Then

(1) $F_n(z)$ is defined and continuous (in the strong operator topology) for all complex z with $\text{Im}z \geq 0$ and it is *holomorphic* in the open upper half-plane $\text{Im}z > 0$.

(2) The function $F_n(z)$ has the integral representation

$$F_n(z) = \frac{(z+i)^2}{2\pi i}\int_{-\infty}^{\infty} \frac{F_n(t)}{(t+i)^2(t-z)}\,dt, \quad \text{Im}z > 0. \tag{21}$$

(3)

$$\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{F_n(t)\,dt}{(t+i)^2(t-z)}=0,\quad \mathrm{Im}z<0. \tag{22}$$

Proof of Lemma 1: The assertion (1) follows from the positive self-adjointness of H and its proof is standard. To prove assertion (2) we start with Cauchy's integral formula for the function $F_n(z)/(z+i)^2$ which is holomorphic in the open upper half-plane,

$$\frac{F_n(z)}{(z+i)^2}=\frac{1}{2\pi i}\oint_C\frac{F_n(z')}{(z'+i)^2(z'-z)}\,dz',\quad \mathrm{Im}z>0,$$

where C is *any* simple *closed* rectifiable contour enclosing the point z and contained entirely in the open upper half-plane. A similar integral representation holds of course for the holomorphic function $F_n(z)$ itself. But with the choice we have made the integrand vanishes faster than $|z'|^{-1}$ as $|z'|\to\infty$. Hence if we choose the closed contour C to be the axis running from $-\infty+i\epsilon$ to $+\infty+i\epsilon$ and an infinite semicircle we could rewrite the contour integral as an open line integral

$$F_n(z)=\frac{(z+i)^2}{2\pi i}\int_{-\infty}^{\infty}\frac{F_n(t+i\epsilon)}{(t+i+i\epsilon)^2(t-z+i\epsilon)}, \tag{21'}$$

$\mathrm{Im}z>\epsilon>0.$

The (operator) norm of this integrand is dominated by the integrable function

$$(1+t^2)^{-1}\circ(\mathrm{Im}z-\epsilon_0)^{-1}$$

for all ϵ with $0\leq\epsilon<\epsilon_0<\mathrm{Im}z$. Moreover,

$$\text{s-}\lim_{\epsilon\to 0_+}F_n(t+i\epsilon)=F_n(t).$$

Hence the conditions for the application of Lebesgue's dominated convergence theorem for operator-valued integrals[11] are met and (21′) goes over to the desired representation (21) in the limit $\epsilon\to 0_+$. The relation (22) is similarly obtained from the vanishing of the contour integral

$$\frac{1}{2\pi i}\oint_C\frac{F_n(z')}{(z'+i)^2(z'-z)}\,dz'\quad\text{for}\quad \mathrm{Im}z<0.$$

Lemma 2: With the same notation as in Lemma 1 let us assume that

$$W(t)\equiv\text{s-}\lim_{n\to\infty}F_n(t)\equiv\text{s-}\lim_{n\to\infty}[E\exp(iHt/n)\,E]^n$$

exists for all real t. Then:

(1) $W(z)\equiv\text{s-}\lim_{n\to\infty}F_n(z)$ exists for all z with $\mathrm{Im}z>0$.

(2) The function $W(z)$ is holomorphic in the *open* upper half-plane and satisfies the semigroup composition law

$$W(z_1)\,W(z_2)=W(z_1+z_2). \tag{23}$$

(3) There exists a nonnegative and self-adjoint operator B and a projection G such that

$$GB=BG=B,$$

and

$$W(z)=G\exp(iBz)\,G,\quad \mathrm{Im}z>0. \tag{24}$$

Proof of Lemma 2: To prove (1) we start with the representation (21) for $F_n(z)$. By assumption, $W(t)\equiv\text{s-}\lim_{n\to\infty}F_n(t)$ for all real t and

$$\left\|\frac{F_n(t)}{(t+i)^2(t-z)}\right\|\leq\frac{(1+t^2)^{-1}}{|\mathrm{Im}z|}\quad\text{for all } n.$$

We can therefore apply again the Lebesgue theorem on dominated convergence and conclude that

$$W(z)=\text{s-}\lim_{n\to\infty}F_n(z)$$

exists and has the representation

$$W(z)=\frac{(z+i)^2}{2\pi i}\int_{-\infty}^{\infty}\frac{W(t)}{(t+i)^2(t-z)}\,dt,\quad \mathrm{Im}z>0. \tag{25}$$

From the well known Vitali's theorem[11] we can conclude that $W(z)$ is holomorphic in the open upper half-plane.

To prove the semigroup property of $W(z)$ we show first that this law holds for pure imaginary values,

$$W(is)\,W(it)=W(i(t+s))$$

for all positive t and s.

To this end, first consider the case where t and s are rationally related so that there exist positive integers p,q for which

$$\frac{s+t}{r(p+q)}=\frac{s}{rp}=\frac{t}{rq}$$

for all integers r. For such s,t we can deduce

$$\left[E\exp\left(-H\frac{t+s}{r(p+q)}\right)E\right]^{r(p+q)}=\left[E\exp\left(-H\frac{s}{rp}\right)E\right]^{rp}\left[E\exp\left(-H\frac{t}{rq}\right)E\right]^{rq}$$

which, in the limit $r\to\infty$ yields

$$W(i(t+s))=W(is)\,W(it).$$

Once this is established for rationally related positive s and t by continuity it can be extended to all positive s and t. Since $W(is)$ is holomorphic it is, *a fortiori*, continuous for $s>0$.

To prove assertion (3) we observe that the operators

$$W(is)\equiv\text{s-}\lim_{n\to\infty}[E\exp(-Hs/n)\,E]^n$$

form a semigroup of self-adjoint operators for $s>0$. According to a well known structure theorem for such semigroups[12] there exists a *self-adjoint nonnegative* operator B and a projection G such that

$$BG=GB=B$$

and

$$W(is)=G\exp(-Bs)\,G,\quad s>0.$$

The function $z\to G\exp(iBz)\,G$ is holomorphic in the open upper half-plane and assumes the same values $W(is)$ as the holomorphic function $W(z)$ for $z=is$ $(s>0)$. The uniqueness of holomorphic functions, then, immediately establishes the representation (24). The semigroup property (23) for z_1,z_2 in the open upper half-plane follows from (24) and the commutativity of G and B.

Lemma 3: Under the assumptions of Lemma 2, we have the weak limit for operators along the real axis

$$\text{w-}\lim_{\eta\to 0_+} W(s+i\eta) \equiv W(s) \quad \text{for almost all real } s. \tag{26}$$

Proof of Lemma 3: To obtain this weak limit let us start from the integral representation (25) rewritten in the form

$$W(s+i\eta) = \frac{(s+i+i\eta)^2}{2\pi i}\int_{-\infty}^{\infty} \frac{W(t)}{(t+i)^2(t-s-i\eta)}\,dt, \quad \eta>0.$$

On the other hand, from (22) and the Lesbegue dominated convergence theorem

$$0 = \frac{(s+i+i\eta)^2}{2\pi i}\int_{-\infty}^{\infty} \frac{W(t)}{(t+i)^2(t-s+i\eta)}\,dt, \quad \eta>0.$$

Therefore,

$$W(s+i\eta) = \frac{(s+i+i\eta)^2}{\pi}\int_{-\infty}^{\infty} \frac{W(t)}{(t+i)^2}\,\frac{\eta}{(t-s)^2+\eta^2}\,dt.$$

For any two vectors ψ, ϕ in $\mathcal{H}$ we may write

$$(\psi, W(s+i\eta)\,\phi) = \frac{(s+i+i\eta)^2}{\pi}\int_{-\infty}^{\infty} \frac{(\psi, W(t)\,\phi)}{(t+i)^2}\,\frac{\eta}{(t-s)^2+\eta^2}\,dt.$$

Since the quantity $(\psi, W(t)\,\phi)/(t+i)^2$ considered as a function of t is integrable, it follows that

$$\lim_{\eta\to 0_+}(\psi, W(s+i\eta)\,\phi) = (\psi, W(s)\,\phi) \tag{27}$$

for almost all s.

To complete the assertion of Lemma 3 a technical difficulty is to be resolved. For a given pair ψ, ϕ of vectors, the assertion (29) has been shown to hold for almost all s. The exceptional set (of measure zero) where this result may not hold may appear to depend on the pair ψ, ϕ chosen. To show that there is at most a single *null* set outside which (27) holds for *all* pairs ψ, ϕ we proceed as follows: Let $\mathcal{D}$ be a *countable* dense subset of the separable Hilbert space $\mathcal{H}$ and let $\mathcal{N}$ be the union of the *countable* family of exceptional null sets corresponding to all pairs ψ, ϕ with $\psi\in\mathcal{D}$, $\phi\in\mathcal{D}$. This set $\mathcal{N}$ is a set of measure zero and the weak limit (26) holds everywhere outside this set for ψ, ϕ in $\mathcal{D}$, but then (27) will hold in the complement of $\mathcal{N}$ for all pairs ψ, ϕ not necessarily in $\mathcal{D}$. In fact, writing

$$A(s,\eta) = W(s+i\eta) - W(s)$$

we may obtain

$$(\psi, A(s,\eta)\,\phi) = (\psi-\psi_n, A(s,\eta)\,\phi) + (\psi_n, A(s,n)(\phi-\phi_n)) + (\psi_n, A(s,\eta)\,\phi_n).$$

We see that for s outside the exceptional set $\mathcal{N}$ the first two terms on the right-hand side tend to zero as $n\to\infty$, since we may choose

$$\text{s-}\lim_{n\to\infty}\psi_n = \psi, \quad \text{s-}\lim_{n\to\infty}\phi_n = \phi.$$

The third term, by hypothesis, goes to zero as $\eta\to 0_+$ since ψ_n, ϕ_n are chosen to lie in $\mathcal{D}$. [The proof of this lemma incorporates a suggestion due to K. Sinha.]

The proof of assertion (a) of the theorem may now be easily completed by combining the conclusions of the preceding lemmas,

$$W(s) = \text{w-}\lim_{\eta\to 0_+} W(s+i\eta) = \text{w-}\lim_{\eta\to 0_+} G\exp[iB(s+i\eta)]\,G$$
$$= G\exp(iBs)\,G \quad \text{for almost all real } s. \tag{24'}$$

Thus $W(s)\,W^*(s) = G$ for almost all real s. According to assumption (iv) in the statement of Theorem 1,

$$\text{w-}\lim_{s\to 0_+} W(s)\,W^*(s) = \text{w-}\lim_{s\to 0_+} T(-s)\,T^*(-s)$$
$$= \text{w-}\lim_{s\to 0_+} T^*(s)\,T(s) = E. \tag{28}$$

Thus $G=E$ and we may rewrite (24) in the form

$$W(s) = E\exp(iBs)\,E \text{ for almost all } s,$$
$$EB = BE = B. \tag{29}$$

But we can now strengthen this relation for $W(s)$ to read

$$W(s) = E\exp(iBs)\,E \quad \text{for } \textit{all } s \tag{30}$$

in view of the strong continuity of $W(s)$. Combining (29) and (30) we immediately deduce the validity of assertion (a).

Although not of primary interest for the discussion in this paper, we recall the known result[3] that if H is semibounded, the operators $E\mathrm{U}(t)\,E$ cannot form a semigroup for $t\geq 0$ except in the event of E commuting with $\mathrm{U}(t)$ for all real t. We may subsume this result as a corollary to Theorem 1.

Corollary: Let the self-adjoint operator H be semibounded, let E be an orthogonal projection and let $\mathrm{U}(t)$ stand for $\exp(-iHt)$. If $\{E\mathrm{U}(t)\,E \mid t\geq 0\}$ form a semigroup, then

$$E\mathrm{U}(t) = \mathrm{U}(t)\,E \quad \text{for all real } t. \tag{31}$$

Proof: The semigroup property for $E\mathrm{U}(t)\,E$, i.e., the relation

$$E\mathrm{U}(t)\,E\mathrm{U}(s)\,E = E\mathrm{U}(t+s)\,E \quad \text{for } \textit{all real } s,t \tag{32}$$

will imply

$$E\mathrm{U}(t)\,E\mathrm{U}^*(t)\,E = E \quad \text{for } \textit{all real } t \tag{32'}$$

and hence

$$E \leq \mathrm{U}(t)\,E\mathrm{U}^*(t) \quad \text{for } \textit{all real } t.$$

Multiplying this equation from the left by $\mathrm{U}(-t)$ and the right by $\mathrm{U}^*(-t)$ will yield

$$\mathrm{U}(-t)\,E\mathrm{U}^*(-t) \leq E \quad \text{for } \textit{all real } t.$$

Together these two inequalities imply

$$E = \mathrm{U}(t)\,E\mathrm{U}^*(t) \quad \text{for } \textit{all real } t$$

or, equivalently,

$$E\mathrm{U}(t) = \mathrm{U}(t)\,E \quad \text{for } \textit{all real } t.$$

The proof of the corollary is thus reduced to the proof of (32) or (32').

Since the operators $E\mathrm{U}(t)\,E$ are assumed to form a semigroup for $t\geq 0$ and $[E\mathrm{U}(t)\,E]^* = E\mathrm{U}(-t)\,E$, for all positive integers n and *all real* t we have

$$[E\mathrm{U}(t/n)\,E]^n = E\mathrm{U}(t)\,E.$$

Hence

$$\text{s-}\lim_{n\to\infty}[EU(t/n)E]^n = EU(t)E \equiv T(t)$$

exists for *all* real t and all the assumptions of Theorem 1 are verified, except for (ii). But as we have pointed out, this assumption itself was needed for the sole purpose of guaranteeing the existence of $T(t)$ for *all real* t. Thus we can safely conclude that Theorem 1 applies in this case also and hence (32) holds for *all* real t and s.

4. CONCLUDING DISCUSSION

What conclusions must we draw from Zeno's paradox in quantum theory? Is it a curious but innocent mathematical result or does it have something to say about the foundation of quantum theory? Does it, for example, urge us to have a principle in the formulation of quantum theory that forbids the continuous observation of an observable that is not a constant of motion?

The answer to the first two questions appears to depend on whether it is operationally meaningful to seek the probability that the particle makes a transition from a preassigned subspace of states $E\mathcal{H}$ to the orthogonal subspace $E^\perp\mathcal{H}$ *sometime* during a given period of time. We have endeavored to present arguments that such probabilities possess operational meaning in terms of the outcome of successive (in the limit, continuous) measurements of an appropriate quantum mechanical observable. If this is accepted, it follows that to be a *complete* theory quantum mechanics must provide an algorithm for computing these probabilities. The quantum Zeno's paradox shows that the seemingly natural approach to this problem discussed in the preceding sections leads to bizarre and physically unacceptable answers. We thus lack a trustworthy quantum-theoretic algorithm for computing such probabilities. Until such a trustworthy algorithm is developed the *completeness* of quantum theory must remain in doubt.

The lack of a trustworthy quantum-theoretic algorithm for probabilities like $P(0,t;\rho)$ is intimately connected with the difficulties involved in defining an operator of "arrival time" (or, more generally, "time of transition") in quantum theory.[4] Let us briefly discuss this problem in the context of the "time of decay" of an unstable particle.

From the definition of $P(0,t;\rho)$ it must have the following properties:

$$\begin{aligned}
&\text{(i) } P(0,t;\rho) \geq 0 \text{ for all } t \geq 0\\
&\text{(ii) } P(0,t;\rho) \geq P(0,s;\rho) \text{ for } t \geq s\\
&\text{(iii) } P(0,t;\rho) \to 1 \text{ for } t \to \infty\\
&\text{(iv) } P(0,t;\rho) \to \mathrm{Tr}[\rho E^\perp] \text{ for } t \to 0_+.
\end{aligned} \tag{33}$$

In addition, $P(0,t;\rho)$ may be assumed to be continuous as a function of t. If we were to succeed in finding a formula

$$P(0,t;\rho) = \mathrm{Tr}[\rho B(0,t)], \tag{34}$$

then the operator $B(0,t)$ would have the following properties:

(i) $B(0,t) \geq 0$,

(ii) $B(0,t) \geq B(0,s)$, $t \geq s$,

(iii) $B(0,t) \to I$ (strongly), $t \to \infty$,

(iv) $B(0,t) \to E^\perp$ (strongly), $t \to 0_+$,

(v) $B(0,t)$ is a strongly continuous function of t.

The family $B'(0,t) \equiv EB(0,t)E$ restricted to the subspace $E\mathcal{H}$ of the (unstable) undecayed states will then form a "generalized resolution of the identity" (GRI). Unlike the more familiar (projection-valued) resolution of the identity, a GRI does not necessarily determine a densely defined operator, but under some mild additional assumptions (which we need not specify explicitly here) the GRI $B'(0,t)$ will determine a Hermitian (though not necessarily self-adjoint) operator τ so that

$$(\psi, \tau\psi) = \int t\, d(\psi, B'(0,t)\psi) \text{ for all } \psi \in \mathcal{D}(\tau). \tag{35}$$

The operator τ thus defined may then be interpreted as the operator of "time of decay."

Conversely, if there exists a positive Hermitian "time of decay" operator τ associated with the subspace $E\mathcal{H}$ of undecayed states and $B'(0,t)$ denotes a GRI associated with it, then through (33) we may define the probability $P(0,t;\rho)$ which may be interpreted as the probability that the system prepared initially in the (undecayed) state ρ will be found to decay *sometime* during the interval $[0,t]$.

Looked at from this point of view, the Zeno's paradox thus strengthens and sharpens the pessimistic conclusion of Allcock[4] and others concerning the possibility of introducing an observable of "arrival time" in quantum theory. We must emphasize that in our study here, the conclusion is *not* based on certain *a priori*, but questionable, assumptions about τ; such, for instance, as the assumption that τ be "canonically conjugate" to the Hamiltonian, or that τ be a *self-adjoint* operator in the Hilbert space. In the literature such requirements were implicitly or explicitly placed on τ.

We have so far supposed that it is operationally meaningful to ask about probabilities such as $P(0,t;\rho)$ and $Q(0,t;\rho)$. We have also taken the stance that the observed tracks of unstable particles in a bubble chamber or photographic emulsion is in contradiction with the conclusion we have called Zeno's paradox in quantum theory. It is, however, possible to adopt one of the following attitudes:

(1) Probabilities such as $P(0,t;\rho)$ have no operational meaning: There is a fundamental principle in quantum theory that denies the possibility of continuous observation.

Since so far no such principle has been derived from or incorporated into quantum theory, this is not a satisfactory way of resolving the paradox at the present time.

(2) Zeno's paradox is based on the assumption that the continuous measurements are ideal measurements. But measurements (or, observations) involved in the recording of the track of an unstable particle in a detecting medium are nonideal in the sense of (3).

This is a tenable view and it would deny the validity of Theorem 1 as stated and proved in this paper. It has the somewhat unsettling side effect that $P(0, t; \rho)$ and hence the "observed lifetime" of an unstable particle is not a property of the particle (and its Hamiltonian) only, but depends on the details of the observation process. At the present time we have no indication that this is so.

(3) The record of the track of a particle is not a *continuous* observation that the particle has not decayed, but only a discrete sequence of such observations; while Zeno's paradox obtains only in the limit of continuous observations.

While this is tenable, the sufficiently repeated monitoring of the particle should again lengthen the lifetime. There is, however, no indication that the lifetime of a (charged) unstable particle (say, a muon) is appreciably increased in the process of its track formation through bubble chamber. To shed additional light on this question a quantitative investigation of the effect of repeated monitoring on the lifetime of particles (in specific models) is in progress.[13]

(4) Natural though it seems, it is wrong to assume that the temporal evolution of a quantum system under continuing observation can be described by a linear operator of time-evolution such as $T(t)$. It can be described only in terms of a persistent interaction between the quantum system and the *classical measuring apparatus*. When this is done the quantum Zeno's paradox will either disappear or if it survives, at least, it will be understandable as the drastic change in the behavior of the quantum system caused by its continuous interaction with a classical measuring apparatus.

This point of view is at present only a program since there is no standard and detailed theory for the actual coupling between quantum systems with *classical measuring apparatus*. A beginning in this direction is made in a forthcoming paper.[14]

Having been forced into such unusual points of view by the quantum Zeno's paradox one is prompted to draw also some parallels between it and certain empirical findings in the study of human awareness. We shall present such close parallels between the quantum Zeno's paradox and the findings of sensory deprivation and other experiments pertaining to the study of consciousness in a separate publication.

In conclusion, it seems to us that the problems posed by Zeno's paradox have no clean cut resolution at the present time and deserve further discussions. It may also be reemphasized that the probabilities such as $Q(\Delta; \eta, \rho)$ [or $Q(\Delta; \rho)$] that pertain to the outcomes of successive measurements (or continuous measurements) depend on the law according to which "state collapses" occur at the time of measurements. Thus one may say that the "collapse of state vector" caused by measurement, which has haunted the foundation of quantum mechanics like an invisible ghost becomes visible through probabilities such as $Q(\Delta; \eta, \rho)$, etc. The probabilities pertaining to the outcomes of several successive (as well as continuous) measurements therefore deserve further theoretical as well as experimental study than they have received so far.

ACKNOWLEDGMENTS

It is a pleasure to thank Professors C. Chiu, M. Guenin, K. Gustafson, Ph. Martin, C Piron, and K. Sinha for their critical comments as well as constructive suggestions. We are also grateful to the referee for some useful comments, in particular, for bringing to our notice Refs. 5 and 6. One of us (B. M.) was partially supported by Swiss NSF during summer 1975.

*Work supported in part by Energy Research and Development Administration Contract E(40-1)3992.

[1]Another possible suggestion is to interpret

$$\frac{1}{t}\int_0^t \mathrm{Tr}[\rho U^*(s) E U(s)]\, ds$$

as the desired probability $Q(0, t; \rho)$. Apart from the fact that there is really no convincing reason for this interpretation this expression is *not* generally a monotone (decreasing) function of time t, a property which $Q(0, t; \rho)$ should possess.

[2]E. Schrödinger, Naturwissenschaften 23, 807 (1935).
[3]D. Williams, Commun. Math. Phys. 21, 314 (1971); L. Horwitz, J. P. Marchand, and J. Lavita, Rocky Mount. J. Math. 1, 225 (1975).
[4]G. R. Allcock, Ann. Phys. (N.Y.) 53, 251—348 (1969).
[5]W. Yorgrau in *Problems in Philosophy in Science*, edited by I. Lakatos and A. Musgrave (North-Holland, Amsterdam, 1968), pp. 191—92.
[6]H. Ekstein and A. Seigert, Ann. Phys. (N.Y.) 68, 509 (1971).
[7]B. Misra and E. C. G. Sudarshan, in preparation.
[8]E. P. Wigner, in *Foundations of Quantum Mechanics*, edited by B. d'Espgnat (Academic, New York, 1971), especially formulas 14 and 14(a), p. 16.
[9]P. R. Chernoff, Mem. Am. Math. Soc. 140 (1974).
[10]C. N. Friedman, Indiana Univ. Math. J. 21, 1001—11 (1972).
[11]E. Hille and R. Phillips, *Functional Analysis and Semigroups* (Am. Math. Soc. Colloq. Publ., Providence, R.I., 1957).
[12]Ref. 11, Sec. 22.3.
[13]C. Chiu, B. Misra, and E. C. G. Sudarshan, in preparation.
[14]E. C. G. Sudarshan and B. Misra, in preparation.

Convolution and H-equations for operator-valued functions with applications to neutron transport theory

C. T. Kelley

Purdue University, West Lafayette, Indiana
(Received 11 June 1976)

The Wiener–Hopf factorization of certain-valued functions is related to operator-valued generalizations of Chandrasekhar's H-functions. These functions satisfy a nonlinear system and may be computed by an iterative scheme. A general transport equation is solved in terms of these functions. The equation for steady-state transport in one space dimension with isotropic scattering and continuous energy dependence is discussed as an application.

I. INTRODCUTION

Study of multigroup transport problems and the corresponding convolution equation have led to matrix-valued generalizations of the Chandrasekhar H-function.[1] These matrices satisfy a nonlinear system, which we write in matrix form as

$$\begin{pmatrix} H_l(\mu,\zeta) & 0 \\ 0 & \tilde{H}_r(\mu,\zeta) \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \mu\zeta \begin{pmatrix} H_l(\mu,\zeta) & 0 \\ 0 & \tilde{H}_r(\mu,\zeta) \end{pmatrix} \times \int_0^1 \frac{1}{\mu+\nu} \begin{pmatrix} D(\nu) & 0 \\ 0 & \tilde{D}(\nu) \end{pmatrix} \begin{pmatrix} H_r(\nu,\zeta) & 0 \\ 0 & \tilde{H}_l(\nu,\zeta) \end{pmatrix} d\nu, \tag{1.1}$$

where I is the $n\times n$ identity matrix, $\sim$ denotes transpose, and D is an integrable matrix-valued function defined on $[0,1]$.

Mullikin,[2] Victory,[3] and Mullikin and Victory[4] have studied Eq. (1.1) in detail for matrix-valued H-functions. In this paper we extend part of that analysis. In particular we observe the operator-valued analog of Eq. (1.1) and show that the H-operators provide a Wiener-Hopf factorization of an operator-valued function. Notation has been chosen to emphasize this similarity of the operator-matrix equations to the scalar equations.[5,6]

We then extend a result of Zweifel and Bowden,[7] showing that the nonlinear equation for the H-operators converge under iteration to the physical solution, i.e., the solution arising from the Wiener-Hopf factorization of an operator-valued function.

As an application we solve a transport equation of the form

$$\mu\frac{\partial\psi}{\partial x} = \Sigma\psi - C\int_{-1}^{1}\psi(x,\mu')\,d\mu'. \tag{1.2}$$

Here ψ is Banach-space valued, and Σ and C are operators. Such problems occur in steady-state transport problems in one space dimension with isotropics scattering and continuous energy dependence.[2,8] In the scalar and multigroup cases, our solution becomes the Case singular eigenfunction expansion[9] after a contour integration.[10] In these cases this seems to be a different way of obtaining the results of Larsen and Habetler[11] and Bowden, Sancaktar, and Zweifel.[12,13]

In this paper, integrals of Banach-valued and operator-valued functions are to be understood as Bochner integrals. A good general reference for this is the book by Hille and Phillips.[14]

II. WIENER-HOPF THEORY AND H-OPERATORS

Let $\mathcal{B}$ be a Banach space, $\mathcal{B}'$ its dual. We denote for $1\le p<\infty$ by $\mathbf{B}_p(a,b),\mathcal{B})$ the space of Bochner p-integrable functions from (a,b) to $\mathcal{B}$. Let $\mathcal{L}(\mathcal{B})$ and $\mathrm{Com}(\mathcal{B})$ be the algebras of bounded (resp. compact) operators on $\mathcal{B}$. If $A\in\mathcal{L}(\mathcal{B})$, A' will denote its adjoint.

Let X denote the space of 2×2 diagonal matrices of the form

$$A = \begin{pmatrix} A_l & 0 \\ 0 & A_r' \end{pmatrix}, \tag{2.1}$$

where A_l and $A_r\in\mathcal{L}(\mathcal{B})$. X is a Banach algebra under matrix multiplication with norm given by

$$\|A\|_X = \max(\|A_l\|_{\mathcal{L}(\mathcal{B})},\ \|A_r\|_{\mathcal{L}(\mathcal{B})}). \tag{2.2}$$

We denote by X^0 the subalgebra of X given by

$$X^0 = \{A\in X \mid A_l, A_r\in \mathrm{Com}(\mathcal{B})\}.$$

For

$$A = \begin{pmatrix} A_l & 0 \\ 0 & A_r' \end{pmatrix}\in X$$

we define

$$A^* = \begin{pmatrix} A_r & 0 \\ 0 & A_l' \end{pmatrix}. \tag{2.3}$$

We note that, for $A, B\in X$, $(AB)^* = B^*A^*$.

We let $\mathcal{X}_p$ [resp. $\mathcal{X}_p^0$] $= \mathbf{B}_p(\mathbf{R},X)$ [resp. $\mathbf{B}_p(\mathbf{R},X^0)$].

$\mathcal{X}_p$ and $\mathcal{X}_p^0$ are Banach spaces with norm

$$\|T\|_{\mathcal{X}_p(X,0)} = \left(\int_{-\infty}^{\infty}\|T(x)\|_{X(X^0)}^p\,dx\right)^{1/p}. \tag{2.4}$$

We note that $\mathcal{X}_1$ and $\mathcal{X}_1^0$ are Banach algebras under convolution and $\mathcal{X}_1^0$ is a subalgebra of $\mathcal{X}_1$. For $T\in\mathcal{X}_1$ we may define

$$\hat{T}(\lambda) = \int_{-\infty}^{\infty}\exp(i\lambda x)T(x)\,dx. \tag{2.5}$$

We then have that if $F(x) = \int_{-\infty}^{\infty}T(x-y)S(y)\,dy$ with $T, S\in\mathcal{X}_1$, then $F\in\mathcal{X}_1$ and

$$\hat{F} = \hat{T}\hat{S}. \tag{2.6}$$

Copyright © 1977 American Institute of Physics

Also if $T\in\mathcal{X}_1$, $\hat{T}^*=(\hat{T})^*$.

We define $\mathcal{X}_C(\mathcal{X}_C^0)$ by $\mathcal{X}_C(\mathcal{X}_C^0)=\{T:\mathbf{R}\to X(X^0)\mid T$ is continuous in the $\|\ \|_X$ topology and $\sup_{x\in\mathbf{R}}\|T(x)\|_{X(X^0)}<\infty\}$. $\mathcal{X}C$ is a Banach algebra under pointwise multiplication with norm given by

$$\|T\|_{\mathcal{X}_C}=\sup_{x\in\mathbf{R}}\|T(x)\|_X. \tag{2.7}$$

$\mathcal{X}_C^0$ is a subalgebra of $\mathcal{X}_C$.

If T is in any of the spaces $\mathcal{X}$, define

$$\rho T(x)=\begin{cases}T(x), & x>0,\\ 0, & x\le 0.\end{cases} \tag{2.8}$$

Let $\rho\mathcal{X}=\{\rho T\mid T\in\mathcal{X}\}$. $\rho\mathcal{X}_1$ ($\rho\mathcal{X}_1^0$) are subalgebras of $\mathcal{X}_1$ ($\mathcal{X}_1^0$).

Now let $\mathcal{X}$ denote any of the spaces $\mathcal{X}_\rho$, $\mathcal{X}_\rho^0$, $\mathcal{X}_C$, $\mathcal{X}_C^0$. For $T\in\mathcal{X}_1$, $K\in\mathcal{X}$ define

$$\mathcal{T}K(x)=\int_{-\infty}^{\infty}T(x-y)K(y)\,dy. \tag{2.9}$$

For $L\in\rho\mathcal{X}$ define

$$\mathbb{P}\mathcal{T}L(x)=\rho(\mathcal{T}L)(x). \tag{2.10}$$

It is easy to prove the following result.

Theorem 2.1: Let $\mathcal{T}$ and $\mathbb{P}\mathcal{T}$ be as above. Then $\mathcal{T}\in\mathcal{L}(\mathcal{X})$, $\mathbb{P}\mathcal{T}\in\mathcal{L}(\rho\mathcal{X})$, and

$$\|\mathbb{P}\mathcal{T}\|_{\mathcal{L}(\rho\mathcal{X})}\le\|\mathcal{T}\|_{\mathcal{L}(\mathcal{X})}\le\|T\|_{\mathcal{X}_1}. \tag{2.11}$$

We now prove a factorization theorem due to Gohberg,[15] using an argument essentially due to Mullikin.[16] We let $I_\mathcal{B}$ be the identity operator in $\mathcal{L}(\mathcal{B})$ and let $I\in X$ be given by

$$I=\begin{pmatrix}I_\mathcal{B} & 0\\ 0 & (I_\mathcal{B})'\end{pmatrix}. \tag{2.12}$$

Theorem 2.2: Let $T\in\mathcal{X}_1^0$. Assume that

(i) $T^*(-x)=T(x)$,

(ii) $\|T\|_{\mathcal{X}_1}<1$.

Then there is a $\Gamma\in\rho\mathcal{X}_1^0$ so that $\mathcal{H}(\lambda)=I+\hat{\Gamma}^*(\lambda)$ satisfies

$$(I-\hat{T}(\lambda))\mathcal{H}^*(\lambda)\mathcal{H}(-\lambda)=I. \tag{2.13}$$

Proof: As $\|T\|_{\mathcal{X}_1}<1$, the equation

$$(I-\mathbb{P}\mathcal{T})\Gamma=\rho T \tag{2.14}$$

has a unique solution in $\rho\mathcal{X}_1^0$.

For $\mathrm{Im}z\ge 0$, let $J(x,z)$ be the unique solution in $\rho\mathcal{X}_C$ to

$$(I-\mathbb{P}\mathcal{T})J(\cdot,z)=\exp[i(\cdot)z]I\quad(x>0). \tag{2.15}$$

Note that if $\mathrm{Im}z>0$, $J\in\mathcal{X}_1$. We have

$$\int_0^\infty\Gamma^*(x)J(x,z)\,dx=\int_0^\infty\exp(izx)\Gamma^*(x)\,dx+\int_0^\infty\Gamma^*(x)[\mathcal{T}J(\cdot)](x)\,dx. \tag{2.16}$$

The last integral in Eq. (2.16) is

$$\int_0^\infty\Big[\int_0^\infty\Gamma^*(x)T(x-y)\,dx\Big]J(y,z)\,dy$$

$$=\int_0^\infty(\mathbb{P}\mathcal{T}\Gamma)^*(y)J(y,z)\,dy$$

$$=\int_0^\infty[\Gamma^*(y)-T^*(y)]J(y,z)\,dy.$$

The relation $T^*(-x)=T(x)$ was used in the second equality.

Hence, (2.16) becomes

$$\int_0^\infty\Gamma^*(y)J(y,z)\,dy$$

$$=\int_0^\infty\exp(iyz)\Gamma^*(y)+\int_0^\infty[\Gamma^*(y)-T^*(y)]J(y,z)\,dy. \tag{2.17}$$

Cancellation implies

$$\Gamma^*(z)=\int_0^\infty T(-y)J(y,z)\,dy$$

$$=\lim_{x\to 0^+}(\mathbb{P}\mathcal{T}J)=J(0^+,z)-I. \tag{2.18}$$

Hence,

$$J(0^+,z)=\mathcal{H}(z)\quad(\mathrm{Im}z\ge 0). \tag{2.19}$$

An argument due to Mullikin[16] will show that

$$\frac{\partial}{\partial x}J(x,z)=izJ(x,z)+\Gamma(x)\mathcal{H}(z). \tag{2.20}$$

We integrate both sides of Eq. (2.20) and obtain, for $\mathrm{Im}z\ge 0$,

$$J(x,z)\exp(-ixz)=\Big[\int_0^x\exp(-izy)\Gamma(y)\,dy+I\Big]\mathcal{H}(z). \tag{2.21}$$

Now as

$$\Gamma=\begin{pmatrix}\Gamma_l & 0\\ 0 & \Gamma_r'\end{pmatrix}\in\mathcal{X}_1^0,$$

the operator valued functions, $\hat{\Gamma}_r(z)$ and $\hat{\Gamma}_l(z)$, are compact for all values of z, $\mathrm{Im}z\ge 0$. Hence, if $\mathcal{H}^{-1}(z_0)$ fails to exist, there are φ and φ' in $\mathcal{B}$ and $\mathcal{B}'$ respectively such that

$$\mathcal{H}(z_0)\begin{pmatrix}\varphi\\ \varphi'\end{pmatrix}=J(0^+,z_0)\begin{pmatrix}\varphi\\ \varphi'\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}$$

as an element of $\mathcal{B}\times\mathcal{B}'$. By Eq. (2.21), however, this implies

$$J(x,z_0)\begin{pmatrix}\varphi\\ \varphi'\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}\text{ for every }x>0.$$

Hence

$$\begin{pmatrix}\varphi\\ \varphi'\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}$$

and $\mathcal{H}^{-1}(z)$ exists for $\mathrm{Im}z\ge 0$ and is analytic for $\mathrm{Im}z>0$.

Also from Eq. (2.21) we have that, for z real,

$$\lim_{x\to\infty}J(x,z)\exp(-ixz)=\mathcal{H}^*(-z)\mathcal{H}(z). \tag{2.22}$$

We apply (2.21) to the equation for J to get, for $x>0$, z real,

$$\begin{aligned} &J(x,z)\exp(-ixz)\\ &\quad = I+\exp(-ixz)\int_0^\infty T(x-y)J(y,z)\,dy\\ &\quad = I+\int_0^\infty T(x-y)\exp[iz(x-y)]\\ &\qquad \times[H(z)+\int_0^y \Gamma(u)\exp(izu)\,du H(z)]\,dy\\ &\quad = I+\int_{-\infty}^x T(y)\exp(-izy)\\ &\qquad \times[H(z)+\int_0^{x-y}\Gamma(u)\exp(-izu)\,du H(z)]\,dy. \end{aligned} \tag{2.23}$$

As $T\in X_1^0$ we may apply the dominated convergence theorem to Eq. (2.23) and let x tend to ∞, to get

$$H^*(-z)H(z)=I+\hat{T}(-z)H^*(-z)H(z).$$

This completes the proof.

Following Krien,[6] we have

Corollary 2.1: If $T\in X_1^0$ is as in Theorem 2.2, then $(I-\mathbb{P}\mathcal{T})^{-1}=I+\mathcal{G}$, where, for $S\in PX$

$$\mathcal{G}S(x)=\int_0^\infty G(x,y)S(y)\,dy \tag{2.24}$$

and for $x,y>0$

$$\begin{aligned} G(x,y)=&\begin{Bmatrix}\Gamma(x-y), & x>y\\ \Gamma^*(y-x), & x<y\end{Bmatrix}\\ &+\int_0^{\min\{x,y\}}\Gamma(x-r)\Gamma^*(y-r)\,dr. \end{aligned} \tag{2.25}$$

Note that $G^*(x,y)=G(y,x)$.

If $K\in \mathbf{B}_1(\mathbf{R},\ \mathrm{Com}(\mathcal{B}))$, we set

$$T(x)=\begin{pmatrix}T_l(x) & 0\\ 0 & T_r'(x)\end{pmatrix}=\begin{pmatrix}K(x) & 0\\ 0 & K'(-x)\end{pmatrix},$$

note that $\int_{-\infty}^{\infty}\|K(x)\|_{\mathcal{L}(\mathcal{B})}\,dx=\|T\|_{X_1^0}$, and apply Theorem 2.2 to get

Corollary 2.2: If K is as above and $\int_{-\infty}^{\infty}\|K(x)\|_{\mathcal{L}(\mathcal{B})}\,dx<1$, then there are operator valued functions $H_r(\lambda)$ and $H_l(\lambda)$, continuous for $\mathrm{Im}\lambda\geq 0$, analytic for $\mathrm{Im}\lambda>0$, such that

(i) $H_l^{-1}(\lambda)$ and $H_r^{-1}(\lambda)$ are continuous for $\mathrm{Im}\lambda\geq 0$, analytic for $\mathrm{Im}\lambda>0$.

(ii) $H_r(\lambda)-I$ and $H_l(\lambda)-I$ are compact for every λ, $\mathrm{Im}\lambda\geq 0$.

(iii) $[I-\hat{K}(\lambda)]H_r(\lambda)H_l(-\lambda)=I$ for $\lambda\in\mathbf{R}$.

From the proof of Theorem 2.2 we obtain

Corollary 2.3:

$$\int_0^\infty \exp(i\lambda x)J(x,z)\,dx=\hat{J}(\lambda,t)=\frac{i}{z+\lambda}H^*(\lambda)H(z).$$

In order to get a nonlinear equation similar to the classical scalar equation,[1,5] we require that $T(x)$ have a special form. Let $D\in\mathbf{B}_1([0,1],\ \mathrm{Com}(\mathcal{B}))$ and define $\mathbf{D}\in\mathbf{B}_1([0,1],X^0)$ by

$$\mathbf{D}(\nu)=\begin{pmatrix}D(\nu) & 0\\ 0 & D'(\nu)\end{pmatrix}. \tag{2.26}$$

Note that $\mathbf{D}^*=\mathbf{D}$.

If

$$\begin{aligned} T(x)&=\int_0^1\exp(-|x|/\nu)\mathbf{D}(\nu)\frac{d\nu}{\nu},\\ H(z)&=I+\int_0^\infty T(x)J(x,z)\,dx\\ &=I+\int_0^\infty T(x)(I-\mathbb{P}\mathcal{T})^{-1}[\exp(ixz)\theta(x)I]\,dx, \end{aligned} \tag{2.27}$$

where $\theta(x)$ is the characteristic function of $(0,\infty)$. Corollary (2.1) implies that

$$\begin{aligned} H(z)&=I+\int_0^\infty T(x)\exp(ixz)\,dx\\ &\quad+\int_0^\infty T(x)\int_0^\infty G(x,y)\exp(iyz)\,dy\,dx\\ &=I+\int_0^\infty T(x)\exp(ixz)\,dx\\ &\quad+\int_0^\infty\int_0^\infty[G^*(x,y)T(x)]^*\,dx\exp(iyz)\,dy. \end{aligned}$$

Now,

$$\begin{aligned} &\int_0^\infty G^*(x,y)T(x)\,dx\\ &\quad=\int_0^\infty\int_0^1 G(y,x)\exp(-x/\nu)\mathbf{D}(\nu)\frac{d\nu}{\nu}dx\\ &\quad=\int_0^1\int_0^\infty G(y,x)\exp(-x/\nu)\,dx\,\mathbf{D}(\nu)\frac{d\nu}{\nu}\\ &\quad=\int_0^1 J\left(x,\frac{i}{\nu}\right)\mathbf{D}(\nu)\frac{d\nu}{\nu}-T(x). \end{aligned}$$

The change in order of integration is justified as

$$\left\|G(y,x)\exp(-x/\nu)\mathbf{D}(\nu)\frac{1}{\nu}\right\|_x\in L_1((0,\infty)\times[0,1])$$

for every $x>0$.

Hence,

$$\begin{aligned} H(z)=I+\int_0^\infty&\left[\int_0^1\exp(-x/\nu)+J\left(x,\frac{i}{\nu}\right)\mathbf{D}(\nu)\frac{d\nu}{\nu}\right]\\ &\times\exp(ixz)\,dx. \end{aligned} \tag{2.28}$$

Now $J(x,i/\nu)(1/\nu)\in\mathbb{B}_1((0,\infty)\times[0,1],X)$ by the equation for J and the fact that $(I-\mathbb{P}\mathcal{T})^{-1}$ exists in $\mathcal{L}(PX_1)$. Hence we may interchange the order of integration in Eq. (2.28) to get

$$\begin{aligned} H(z)&=I+\int_0^1\mathbf{D}(\nu)\hat{J}^*\left(z,\frac{i}{\nu}\right)\frac{d\nu}{\nu}\\ &=I+\int_0^1\mathbf{D}(\nu)H^*\left(\frac{i}{\nu}\right)H(z)\left(\frac{i}{\nu}\right)\frac{d\nu}{z+i/\nu}. \end{aligned} \tag{2.29}$$

If, for $\mathrm{Re}\mu$, we set $H(\mu)=H(i/\mu)$, Eq. (2.29) becomes

$$H(\mu)=I+\int_0^1\frac{\mu}{\mu+\nu}\mathbf{D}(\nu)H^*(\nu)\,d\nu H(\mu). \tag{2.30}$$

As H commutes with itself and I, we may also write Eq.

(2.30) as

$$H(\mu)=I+\mu H(\mu)\int_0^1 \mathbf{D}(\nu)H^*(\nu)\frac{d\nu}{\mu+\nu}. \tag{2.31}$$

In order to discuss uniqueness questions related to Eq. (2.30), we assume that $\|T\|=1$, and introduce a parameter, ζ ($|\zeta|<1$) as a multiplicative factor of T.

We define, for $\|\zeta T\|=|\zeta|<1$,

$$H(z,\zeta)=\begin{pmatrix} H_l(z,\zeta) & 0 \\ 0 & H_r'(z,\zeta)\end{pmatrix} \tag{2.32}$$

by $(I-\zeta T(\lambda))H^*(\lambda,\zeta)H(-\lambda,\zeta)=I$, for real λ, with H_r and H_l, as in Corollary (2.2). It is a result of Gohberg[15] that this uniquely defines $H(z,\zeta)$.

It is easy to show that $H(\mu,\zeta)=H(i/\mu,\zeta)$ satisfies, for $|\zeta|<1$, $\mathrm{Re}\,\mu>0$,

$$H(\mu,\zeta)=I+\mu\zeta\int_0^1 \mathbf{D}(\nu)H^*(\nu,\zeta)\frac{d\nu}{\mu+\nu}H(\mu,\zeta). \tag{2.33}$$

In the next section we show that H may be found by an iterative technique.

III. ITERATIVE SOLUTION OF EQ. (2.33)

In this section we extend results of Bowden and Zweifel[7] to the case of Eq. (2.33). In their work Bowden and Zweifel divide by the scalar analog of $\mathbf{D}$. As $\mathbf{D}$ is compact operator-valued, we cannot do this here. Instead, an analytic continuation argument in the parameter ζ is used.

As in Bowden and Zweifel's work, we have

Theorem 3.1: For $|\zeta|<1$, the equation

$$K(\nu,\zeta)=\zeta\mathbf{D}(\nu)+\mu\int_0^1\frac{1}{\mu+\nu}K^*(\nu,\zeta)\,d\nu K(\mu,\zeta) \tag{3.1}$$

has a unique solution in the ball

$$\{K\in\mathbf{B}_1([0,1],X)\mid\|K-\zeta\mathbf{D}\|_{\mathbf{B}_1([0,1],X)}<\tfrac{1}{2}\}. \tag{3.2}$$

Moreover, $K=\lim_{n\to\infty}K_n$, where

$$\begin{aligned}&K_0(\mu,\zeta)=0,\\ &K_n(\mu,\zeta)=\zeta\mathbf{D}(\mu)+\mu\int_0^1\frac{1}{\mu+\nu}K_{n-1}^*(\nu,\zeta)\\ &\qquad\times d\nu K_{n-1}(\mu,\zeta).\end{aligned} \tag{3.3}$$

The convergence is in $\mathbf{B}_1([0,1],X)$ uniformly for ζ in compact subsets of $\{|\zeta|<1\}$.

Now $K=\zeta H\mathbf{D}$ satisfies Eq. (3.1). Also

$$\begin{aligned}&\|\zeta H\mathbf{D}-\zeta\mathbf{D}\|_{\mathbf{B}_1([0,1],X)}\\ &\le\tfrac{1}{2}|\zeta|\sup_{\mu\in([0,1],X)}\|H(\mu,\zeta)-I\|_X)\\ &\le\tfrac{1}{2}|\zeta|\sup_{\mu\in[0,1]}\int_0^\infty\exp(-x/\mu)\|\Gamma^*(x,\zeta)\|_X\,dx\\ &\le\tfrac{1}{2}|\zeta|\,\|\Gamma(\cdot,\zeta)\|_{X_1}.\end{aligned}$$

Now $(I-\zeta\mathbf{P}\mathcal{T})\Gamma=\zeta\rho\Gamma$; hence

$$\|\Gamma(\cdot,\zeta)\|_{X_1}\le\left|\frac{\zeta}{1-\zeta}\right|.$$

So

$$\|\zeta H\mathbf{D}-\zeta\mathbf{D}\|_{\mathbf{B}_1([0,1],X)}<\tfrac{1}{2}$$

for $|\zeta|<\frac{2}{3}$.

So, for $|\zeta|<\frac{2}{3}$, we have

$$H^{-1}(\mu,\zeta)=\left[I-\int_0^1\frac{\mu}{\mu+\nu}K^*(\nu,\zeta)\,d\zeta\right]^{-1} \tag{3.4}$$

where $K=\lim_{n\to\infty}K_n$ as in Eq. (3.3).

Now the right side of Eq. (3.4) is analytic for $|\zeta|<1$; hence, Eq. (3.4) holds for all $|\zeta|<1$, by analytic continuation. The dominated convergence theorem implies

Theorem 3.2: Let K_n be as in 3.3. Then for each $\alpha\in(0,1)$,

$$H(\mu,\zeta)^{-1}=\lim_{n\to\infty}\left[I-\int_0^1\frac{\mu}{\mu+\nu}K_n^*(\nu,\zeta)\,d\nu\right] \tag{3.5}$$

uniformly for $|\zeta|\le\alpha$, $0\le\mu\le1$.

IV. SOLUTION OF A TRANSPORT PROBLEM

The following equation is a model for steady-state neutron transport in our space dimension with isotropic scattering and continuous energy dependence[2,8]:

$$\begin{aligned}&\mu\frac{\partial}{\partial x}\psi(x,\mu,E)+\sigma(E)\psi(x,\mu,E)\\ &\quad=\zeta\int_{E_0}^{E_1}\int_{-1}^{1}k(E,E')\psi(x,\mu',E')\,d\mu'\,dE'.\end{aligned} \tag{4.1}$$

The function ψ is subject to the boundary conditions

$$\begin{aligned}&\psi(0,\mu,E)=\psi_0(\mu,E), && \mu>0,\\ &\lim_{x\to\infty}\psi(x,\mu,E)=0, && \mu<0.\end{aligned} \tag{4.2}$$

σ is bounded on $[E_0,E_1]$ and $\sigma(E)\ge1$; k is continuous on $[E_0,E_1]\times[E_0,E_1]$.

We seek a solution ψ so that both ψ and $(\partial/\partial x)\psi$ are continuous in E and μ and p-integrable ($1\le p<\infty$) in x, or continuous and bounded in x.

If we write the right side of Eq. (4.1) as $F(x,E)$ we find, as in the scalar case,[5]

$$\begin{aligned}F(x,E)-\zeta\int_0^\infty\int_{E_0}^{E_1}k(E,E')\int_0^1\exp[-\sigma(E')|x-y|/\mu]\\ \times\frac{d\mu}{\mu}F(y,E')\,dE'\,dy=\zeta\int_{E_0}^{E_1}k(E,E')\\ \times\int_0^1\exp[-\sigma(E')x/\mu]\psi_0(\mu,E')\,d\mu\,dE'.\end{aligned} \tag{4.3}$$

We rewrite Eq. (4.3) in operator notation by suppressing the E variables. Let β_0 denote the Banach space of continuous functions on $[E_0,E_1]$. We define $C\in\mathrm{Com}(\beta_0)$, $\Sigma\in\mathcal{L}(\beta_0)$, $D\in\mathbf{B}_1([0,1],\mathrm{Com}(\beta_0))$ and $K\in X_1$ as follows.

Let $\varphi \in \mathcal{B}$, then

$$C\varphi(E) = \int_{E_0}^{E_1} k(E, E')\varphi(E')\,dE',$$
$$\Sigma\varphi(E) = \sigma(E)\varphi(E),$$
$$D(\mu)\varphi(E) = \int_{E_0}^{E_1} k(E, E')\theta(\sigma(E')^{-1} - \mu)\varphi(E')\,dE', \tag{4.4}$$
$$K(x)\varphi(E) = \int_0^1 \exp(-|x|/\mu)\,D(\mu)\frac{d\mu}{\mu}\varphi(E).$$

As before, θ is the Heaviside function.

For $\psi_0 \in \mathbf{B}_1([0, 1], \mathcal{B}_0)$ define

$$\psi_\Sigma(\mu, E) = \int_{E_0}^{E_1} k(E, E')\sigma(E')^{-1}\psi_0\left(\frac{\mu}{\sigma(E')}, E\right)dE'. \tag{4.5}$$

With this notation we may write (4.3) as

$$F(x) - \zeta \int_0^\infty K(x-y)F(y)\,dy = \int_0^1 \exp(-x/\mu)\psi_\Sigma(\mu)\,d\mu \quad (x > 0). \tag{4.6}$$

We solve a more general problem. Let $\mathcal{B}$ be any Banach space. For $\Lambda \in \mathcal{L}(\mathcal{B})$ we let $\sigma(\Lambda)$ denote the spectrum of Λ.

Now let $\Sigma \in \mathcal{L}(\mathcal{B})$, $C \in \mathrm{Com}(\mathcal{B})$ be such that

$$\Sigma \text{ has bounded inverse with } \|\Sigma^{-1}\| = 1 \text{ and } \sigma(\Sigma) \subseteq [1, M] \text{ for some finite } M > 1. \tag{4.7a}$$

For every $\psi_0 \in \mathbf{B}_1([0, 1], \mathcal{B})$ there is

$$\psi_\Sigma \in \mathbf{B}_1([0, 1], \mathcal{B}) \text{ so that } C\int_0^1 \exp(-\Sigma x/\nu)\psi_0(\nu)\,d\nu = \int_0^1 \exp(-x/\nu)\psi_\Sigma(\nu)\,d\nu. \tag{4.7b}$$

If $K(x) = C\int_0^1 \exp(-|x|\Sigma/\nu)\,d\nu/\nu$, then there is $D \in \mathbf{B}_1([0, 1], \mathrm{Com}(\mathcal{B}))$ so that

$$K(x) = \int_0^1 \exp(-|x|/\nu)D(\nu)\frac{d\nu}{\nu}, \tag{4.7c}$$

$$\int_{-\infty}^{\infty} \|K(x)\|_{\mathcal{L}(\mathcal{B})}\,dx \le 1. \tag{4.7d}$$

On these assumptions we solve the following problem. Find $\psi(x, \mu)$, $\mathcal{B}$-valued, differentiable once in x, so that both ψ and $\partial\psi/\partial x$ are Bochner integrable in μ either Bochner p-integrable ($1 \le p < \infty$) or continuous and bounded in x, and so that

$$\mu\frac{\partial\psi}{\partial x} + \Sigma\psi = C\int_{-1}^{1} \psi(x, \mu')\,d\mu' \equiv F(x),$$
$$\psi(0, \mu) = \psi_0(\mu), \quad \mu > 0, \tag{4.8}$$
$$\lim_{x\to\infty} \psi(x, \mu) = 0, \quad \mu < 0.$$

The function $\psi_0 \in \mathbf{B}_1([0, 1], \mathcal{B})$ is given.

An argument similar to that given by Burniston, Mullikin, and Siewert[17] will show that, in the special problem given by Eq. (4.1), the assumptions (4.7) will hold if $k(E, E')$ is strictly positive and

$$\lim_{n\to\infty} ((\Sigma^{-1}C)^n\|_{\mathcal{L}(\mathcal{B}_0)})^{1/n} \le \tfrac{1}{2}, \tag{4.9}$$

where Σ and C are given by (4.4) and $\mathcal{B}_0$ is the space of continuous functions on the interval $[E_0, E_1]$.

Under the assumptions (4.7), $F(x)$ satisfies

$$F(x) - \int_0^\infty K(x-y)F(y)\,dy = \int_0^1 \exp(-x/\nu)\psi_\Sigma(\nu)\,d\nu \quad \text{for } x > 0. \tag{4.10}$$

We define $T \in \mathcal{X}_1^0 \cap \mathcal{X}_2^0$ by

$$T = \begin{pmatrix} T_l & 0 \\ 0 & T_r' \end{pmatrix} = \begin{pmatrix} K & 0 \\ 0 & K' \end{pmatrix} \tag{4.11}$$

and apply Sec. II and Fubini's theorem to obtain

$$F(x) = \int_0^1 J_l(x, i/\nu)\psi_\Sigma(\nu)\,d\nu. \tag{4.12}$$

In order to give an iterative procedure for finding F, and hence ψ, it will suffice, by the result of Sec. III, to express J in terms of H. We require the following lemma, which is an easy consequence of the classical Parseval relation.

Lemma (4.1): Let

$$f = \begin{pmatrix} f_l I_\mathcal{B} & 0 \\ 0 & f_r I_\mathcal{B} \end{pmatrix},$$

where f_l and $f_r \in L_1(\mathbf{R}) \cap L_2(\mathbf{R})$. Let $T \in \mathcal{X}_1 \cap \mathcal{X}_2$. Then

$$\int_{-\infty}^{\infty} \begin{pmatrix} \bar{f}_l(x)I_\mathcal{B} & 0 \\ 0 & \bar{f}_r(x)I \end{pmatrix} T(x)\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} \begin{pmatrix} \hat{\bar{f}}_l(\lambda)I_\mathcal{B} & 0 \\ 0 & \hat{\bar{f}}_r(\lambda)I \end{pmatrix} \hat{T}(\lambda)\,d\lambda. \tag{4.13}$$

We let $\Gamma_0(x, \zeta)$ be the extension of Γ to all of $\mathbf{R}$ by the defining equation

$$(I - \zeta\mathbb{P}\mathcal{T})\Gamma = P T. \tag{4.14}$$

We have $\hat{\Gamma}_0(\lambda, \zeta) = \zeta T(\lambda)H^*(\lambda, \zeta)$. We apply the lemma to (2.21) and obtain, for $\mathrm{Im} z \ge 0$,

$$J(x, z, \zeta) = (I - \zeta\mathbb{P}\mathcal{T})^{-1}(\exp(ixz)\theta(x)) = \left[\exp(ixz) - \frac{\zeta}{2\pi i}\int_{-\infty}^{\infty} \hat{T}(\lambda)H^*(-\lambda, \zeta) \times \frac{1}{z-\lambda}[\exp(ix\lambda) - \exp(ixz)]\,d\lambda\right]H(z, \zeta). \tag{4.15}$$

Hence,

$$J_l(x, z, \zeta) = \exp(ixz) - \frac{\zeta}{2\pi i}\int_{-\infty}^{\infty} \hat{K}(\lambda)H_r(-\lambda, \zeta) \times \frac{1}{z-\lambda}[\exp(ix\lambda) - \exp(ixz)]\,d\lambda H_l(z). \tag{4.16}$$

We have

Theorem 4.1: The unique solution, ψ, to Eq. (4.8) under the assumptions (4.7) is given by

$$\psi(x, \mu) = \begin{cases} \displaystyle\int_0^x \exp[-\Sigma(x-y)/\mu]F(y)\frac{dy}{\mu} + \exp(-\Sigma x/\mu)\psi_0(\mu), & \mu > 0 \\ \displaystyle -\int_x^\infty \exp[\Sigma(y-x)/\mu]F(y)\frac{dy}{\mu}, & \mu < 0. \end{cases} \tag{4.17}$$

F is given in terms of H_r and H_l by Eq. (4.12) and Eq. (4.16).

Mullikin has pointed out[10,18] that, in the scalar and multigroup cases, Eq. (4.10) may be used to derive the Case eigenfunction expansion[9] in a manner somewhat different from that used by Larsen and Habetler[11] and Bowden, Sancaktar, and Zweifel.[12,13] Mullikin's approach is to use the relation

$$\hat{K}(\lambda)H_r(-\lambda,\zeta)=\hat{K}(\lambda)[I-\zeta\hat{K}(\lambda)]^{-1}H_l^{-1}(\lambda,\zeta) \tag{4.18}$$

in Eq. (4.16) and move the contour to the branch cut $[i,i\infty)$ of $H_r(-\lambda,\zeta)$. In the operator case, however, incomplete knowledge of the singularities of $[I-\zeta\hat{K}(\lambda)]^{-1}$ in the complex plane makes it unclear how either the Mullikin approach or the Larsen resolvent integration technique can be extended.

In a forthcoming paper we intend to exploit the special properties of the problem given by Eq. (4.1) for positive k to show that $\zeta=1$ is a branch point of order 2 for Eq. (2.33). This result will be applied to extend criticality results of Victory[3] and Mullikin and Victory.[4]

ACKNOWLEDGMENT

The author wishes to thank Professor T. W. Mullikin for suggesting this problem to him and for many helpful suggestions.

[1]S. Chandrasekhar, *Radiative Transfer* (Oxford U. P., Oxford, 1950).
[2]T. W. Mullikin, Transport Theory Stat. Phys. 3(4), 215 (1973).
[3]H. D. Victory, "Multigroup Critical Problems for Slabs in Neutron Transport Theory," thesis (Purdue University, 1974).
[4]T. W. Mullikin and H. D. Victory, "*N*-Group Neutron Transport Theory: A Criticality Problem in Slab Geometry" (to appear in J. Math. Anal. Appl.).
[5]I. W. Busbridge, *The Mathematics of Radiative Transfer* (Cambridge U. P., Cambridge, 1960), p. 4.
[6]M. G. Krien, Am. Math. Soc. Trans. Ser. **22**, 163 (1962).
[7]R. L. Bowden and P. F. Zweifel, "A Banach Space Analysis of the Chandrasekhar *H*-equation (to be published).
[8]A. M. Wienberg and F. P. Wigner, *The Physical Theory of Neutron Chain Reactors* (University of Chicago Press, Chicago, 1958), p. 228.
[9]K. M. Case and P. F. Zweifel, *Linear Transport Theory* (Addison-Wesley, Reading, Mass., 1967), p. 43.
[10]T. W. Mullikin, "Wiener-Hopf Technique," "Standard Procedures to Compute Radiative Transfer in a Scattering Atmosphere" (report of ad hoc committee of IAMAP, J. Lenoble, Chairman, Univ. of Lille, France, 1974).
[11]E. W. Larsen and G. J. Habetler, Comm. Pure Appl. Math. **26**, 525 (1973).
[12]R. L. Bowden, S. Sancaktar, and P. F. Zweifel, J. Math. Phys. **17**, 76 (1976).
[13]R. L. Bowden, S. Sancaktar, and P. F. Zweifel, J. Math. Phys. **17**, 82 (1976).
[14]E. Hille and R. S. Phillips, *Functional Analysis and Semigroups* (Am. Math. Soc., Providence, R.I., 1957).
[15]I. C. Gohberg, Am. Math. Soc. Transl. Ser. 2 **49**, 130 (1966).
[16]T. W. Mullikin, "Mathematics of Radiative Transfer" (to appear in Proceedings of Symposium on Theoretical Astrophysics and Relativity, University of Chicago, 1975).
[17]E. G. Burniston, T. W. Mullikin, and C. F. Siewert, J. Math. Phys. **13**, 1461 (1972)
[18]T. W. Mullikin, private communication.

Relativistic rotation and the anholonomic object

James F. Corum

Electrical Engineering Department, West Virginia University, Morgantown, West Virginia 26506
(Received 29 April 1975; revised manuscript received 14 June 1976)

The purpose of this communication is to call attention to the conceptual economy provided by the *object of anholonomity* for the theory of relativity. This geometric object expresses certain consequences of relativity theory and provides a single, simple framework for discussing a variety of phenomena. It particularly clarifies the description of relativistic rotation. The relativistic rotational transformation of the four coordinate differentials of flat space–time generates a set of anholonomic, or inexact differentials, whose duals are an orthogonal set of basis vectors. How should a rotating observer interpret physical events referred to such orthogonal, but anholonomic frames? The answer to this question rests upon the origin and physical significance of the object of anholonomity. It is demonstrated that not only is the rotational Lorentz transformation an anholonomic transformation, but that the intrinsic anholonomic effects are essential to interpreting rotational phenomena. In particular, the Sagnac effect may be interpreted as the physical manifestation of temporal anholonomity under rotation. The Thomas precession of a reference axis may be interpreted as a consequence of the spatial anholonomity of the rotating frame. Further, the full four-dimensional covariance of Maxwellian electrodynamics, under a relativistic Lorentz rotation, is possible only with the inclusion of anholonomic effects. Schiff attempted to explain the vanishing of the electromagnetic field tensor in the frame of an observer rotating about a charged spherical capacitor, in terms of an interaction with the distant stars. However, if one employs a Lorentzian rotation and the ensuing orthogonal anholonomic reference frames, an explanation is possible without gravitationally induced currents. The anholonomic approach clarifies the distinction between the physically different operations of source rotation and observer rotation in a flat space–time. It is finally concluded that a consistant theory of relativistic rotation, satisfying the principle of general covariance, inherently requires the presence of the *object of anholonomity*.

1. INTRODUCTION

The continuing interest in relativistic rotation which has been evident in the literature of theoretical physics for almost seven decades, has recently gained a wider audience.[1–7] This particular aspect of relativity theory has several engineering applications.[8–10] Consequently, it appears proper to examine the presuppositions inherent in the analysis of rotational phenomena. The following investigation is intended to not only clarify the classical concepts, but also to simplify the computations of rotational electrodynamics and resolve an apparent paradox.[11]

At the turn of the century, discussions of *inertial* motion led to the group of space–time transformations called the Lorentz transformations. Carried out in Minkowski space, they demonstrate an intrinsic marriage of space and time and their success has demonstrated the impotency of the Galilean presuppositions of absolute space and universal absolute time. Upon thoughtful reflection, it appears that the possibility of global coordinatization and temporal synchronization in moving frames of reference arises because of the holonomicity, or integrability, of the transformed inertial coordinates.

Almost from the beginning, however, Ehrenfest,[12] Planck,[13] and subsequently a surfeit of other authors recognized a conceptual difficulty concerning uniform rotation, or noninertial motion. These early authors were concerned with the non-Euclidean nature of rotating frames of reference.

Why is a new examination appropriate? It is the author's position that an important feature of relativistically rotating frames is their anholonomicity. In particular, it is maintained that a consistent theory of relativistic rotation, satisfying the principle of general covariance, inherently requires the presence of the *object of anholonomity*. This object serves, not to modify the theory of special relativity in any way, but rather to express certain of its consequences in a more transparent and useful form.[14]

The orthodox approach to relativistic rotation in the past has followed either of two distinct paths. One might follow a direct Galilean rotational *coordinate* transformation,

$$r' = r, \tag{1a}$$

$$\theta' = \theta \pm \omega t, \tag{1b}$$

$$z' = z, \tag{1c}$$

$$t' = t. \tag{1d}$$

(The primed coordinates indicate the reference coordinates of the rotating observer.) Such an approach, however, suppresses phenomena whose explanation is dependent upon the relativity of simultaneity (e.g., Thomas precession, Sagnac effect). We call this approach the "method of Galilean frames."

In flat space–time, an inertial observer may span the entire space–time manifold with the usual set of basis vectors (see Fig. 1) and their duals, the differential forms $\sigma^{\mu} = (dr,\ d\theta,\ dz,\ cdt)$. These inertial basis vectors form an orthogonal set and, under a Galilean rotation, these reference frames (which are the tangents to natural coordinate curves, and infinitesimal patches parallel to natural coordinate surfaces) are correlated to the frames employed by a rotating observer, and one obtains the usual rotating frame metric from

$$g_{ab} = \mathbf{e}_a \cdot \mathbf{e}_b. \tag{2}$$

 Copyright © 1977 American Institute of Physics

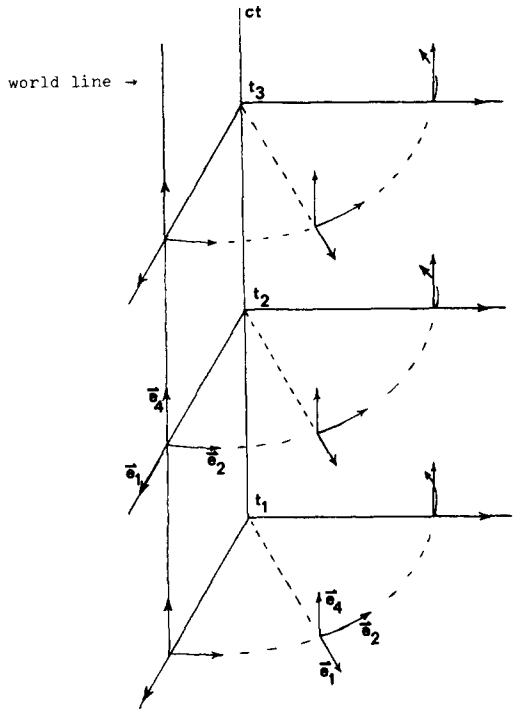


FIG. 1. The world view of the inertial observer with several of the quadruply infinite tetrad frames and his world line. Note that $\mathbf{e}_4$ is the tangent vector to his world line.

Instead of this method, which corresponds to a simple rotation of the spatial axes, some authors employ an approach often called the "method of instantaneous frames." (Although called by this name, it should not be inferred that these are the only frames "instantaneously" at rest with respect to the rotating system.) Where the time coordinate is absolute in the "method of Galilean frames," the t' variable is not a physical or clock time in the "method of instantaneous frames," but provides a measurement of "coordinate interval." Under a Galilean rotation, dt' equals dt and is integrable so that $t' = t$. With the "method of instantaneous frames" the relation between inertial time dt, measured by a clock which is located on the axis of rotation, and local "coordinate interval time," is $dt' = [1 - (r^2\omega^2)/c^2]^{1/2}\, dt_{\text{axis}}$. This is not an exact differential, and consequently dt' is a nonintegrable, or *anholonomic* quantity. Notice that with this approach a Lorentz-like transformation is made of time, but only a Galilean transformation is made of the θ coordinate.

The presuppositions made concerning the structure of the reference frames employed in these two conceptual descriptions of rotation ought to be commented upon. We will assume that the inertial observer is at rest and employ holonomic coordinates (with Greek superscripts) x^μ. In cylindrical polar coordinates the inertial observer has the coordinates $(r,\ \theta,\ z,\ ct)$, with coordinate differentials $dr,\ d\theta,\ dz,\ cdt$. These are the differential forms $\sigma^\mu = (dr,\ d\theta,\ dz,\ cdt)$. The metrical line element of the inertial observer is given by

$$-ds^2 = g_{\mu\nu}\sigma^\mu \otimes \sigma^\nu = g_{\mu\nu}dx^\mu dx^\nu. \tag{3}$$

By employing the duality relation one may obtain the duals to the differential forms, which are the natural basis vectors, $\mathbf{e}_\mu = \partial/\partial x^\mu$. Notice that the subscript μ numbers the basis vector and not the components. This set of vectors spans the space–time manifold and forms a set of orthogonal basis vectors for the rest observer. The Galilean method then correlates these frames to those of a rotating observer by means of a Galilean transformation of basis vectors

$$\mathbf{e}_a = G_a^\mu \mathbf{e}_\mu. \tag{4}$$

(We have chosen the notation μ and a to refer to the reference frame of the inertial and rotating observer, respectively and G_a^μ is the usual Galilean rotation.) The rotating observer employs this $\mathbf{e}_a$ field of frames at every point to span the same space–time events as the rest observer spans with his field of frames $\mathbf{e}_\mu$. Furthermore, the metrical line element of the rotating observer is given by

$$-ds^2 = g_{ab}\sigma^a \otimes \sigma^b, \tag{5}$$

where g_{ab} is obtained from Eq. (2). Hence, the line element of the observer in the Galilean rotating frame is

$$-ds^2 = dr'^2 + r'^2 d\theta'^2 + dz'^2 - \left(1 - \frac{r'^2\omega^2}{c^2}\right) c^2 dt'^2 + 2r'^2\omega\, d\theta'\, dt'. \tag{6}$$

This metric, which is associated with the basis vectors of Eq. (4) should not be confused with the similar metrical components of Eq. (10) below. Equation (10), as is now demonstrated, is associated with the basis vectors of Eq. (9).

A similar analysis now gives the "instantaneous frame" results where

$$\mathbf{e}_a = I_a^\mu \mathbf{e}_\mu \tag{7}$$

and the transformation between local tetrad frames is

$$I_\mu^a = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & \gamma r^2\omega/c & 0 & \gamma \end{pmatrix}, \tag{8}$$

where $\gamma = (1 - r^2\omega^2/c^2)^{-1/2}$. After carrying out this transformation, one finds the tetrad frames:

$$\mathbf{e}_1 = \frac{\partial}{\partial r}, \tag{9a}$$

$$\mathbf{e}_2 = \frac{\partial}{\partial \theta} - \frac{\gamma r^2\omega}{c^2}\frac{\partial}{\partial t}, \tag{9b}$$

$$\mathbf{e}_3 = \frac{\partial}{\partial z}, \tag{9c}$$

$$\mathbf{e}_4 = \frac{1}{\gamma c}\frac{\partial}{\partial t}. \tag{9d}$$

One may obtain the metrical components for the rotating observer from Eq. (2). As a result, the metric employed by this rotating observer has the form

$$g_{ab} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & r^2 & 0 & r^2\omega \\ 0 & 0 & 1 & 0 \\ 0 & r^2\omega & 0 & -1/\gamma^2 \end{pmatrix}. \tag{10}$$

Note that the basis frames and differential forms will be considerably different from those in the "method of Galilean frames," even though the metrical components g_{ab} are the same. It may be instructive to sketch these frames to convince oneself of their distinction. In particular, as stated above, $dt' \neq dt$. This method is of historical interest in that these frames $\mathbf{e}_a$ are *in reality* the ones employed in many relativistic discussions of rotation.[15–19]

There are, however, several characteristic weaknesses in this approach. First, it is anholonomic (as is the approach proposed below) and therefore its use requires the presence of an anholonomic object (a fact which seems to have been previously overlooked). Secondly, the basis vectors which it inherently presupposes are oblique rather than orthogonal ($g_{24} \neq 0$). Consequently tensor equations on these basis vectors are quite different when written in either covariant or contravariant form. The approach now proposed, called the "method of anholonomic frames," provides a computational technique which, it is maintained, is not only simpler to apply, but also conceptually clearer.

2. METHOD OF ANHOLONOMIC FRAMES

This approach seems to have first been employed by Pirani[20] and later by Irvine.[21] However, it appears that both authors did not explicitly employ the anholonomicity of their results. Irvine, in fact, computes an electromagnetic field tensor without including the anholonomic object.

One may obtain a field of orthogonal frames by letting the helical world line of the rotating observer provide the timelike direction and then construct orthogonal spatial vectors from the formulas of Frenet and Serret.[21] The result is the field of frames

$$\mathbf{e}_{r'} = \frac{\partial}{\partial r}, \tag{11a}$$

$$\mathbf{e}_{\theta'} = \gamma \frac{\partial}{\partial \theta} + \frac{\gamma r^2 \omega}{c} \frac{\partial}{\partial t}, \tag{11b}$$

$$\mathbf{e}_{z'} = \frac{\partial}{\partial z}, \tag{11c}$$

$$\mathbf{e}_{4'} = \frac{\gamma}{c} \frac{\partial}{\partial t} + \frac{\gamma \omega}{c^2} \frac{\partial}{\partial \theta}. \tag{11d}$$

These are orthogonal and at every space–time event they are related to the natural basis vectors of the inertial observer by a Lorentz-like transformation (an "instantaneous" Lorentz transformation). That is,

$$\mathbf{e}_a = h_a^{\mu} \mathbf{e}_{\mu}, \tag{12}$$

where

$$h_a^{\mu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \gamma & 0 & \gamma\omega/c \\ 0 & 0 & 1 & 0 \\ 0 & \gamma r^2 \omega/c & 0 & \gamma \end{pmatrix} \tag{13}$$

is the rotational Lorentz transformation. Compare Figs. 1 and 2. Furthermore, the orthogonal basis vectors provide the cylindrical coordinate diagonal

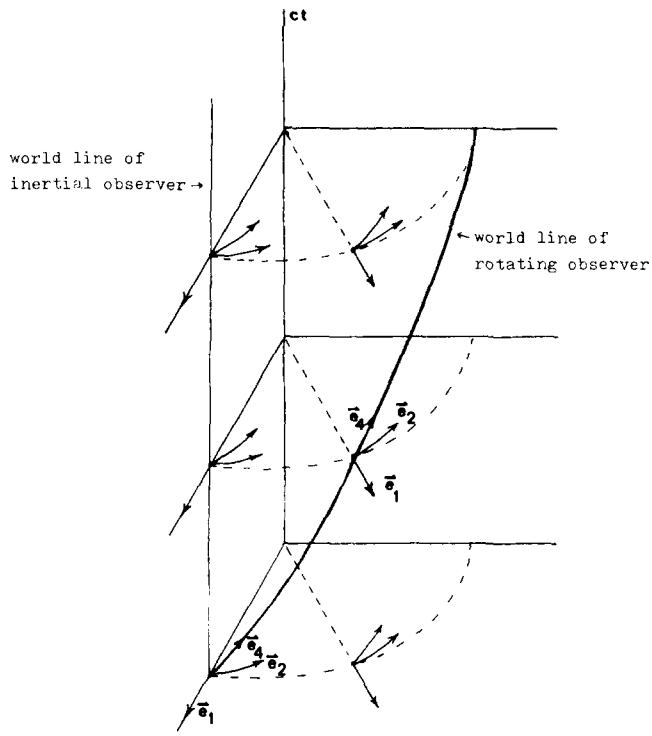


FIG. 2. World view of the inertial observer with the rotating observer's world line and several of his orthogonal tetrad frames.

metric at each point on the space–time manifold (which, of course, is still flat).

In general, it is not possible to find a natural system of coordinates which incorporates the above tetrad vectors employed by the rotating observer. These orthogonal basis vectors are not the tangents to the coordinate axes of any global coordinate system. However, they span every local tangent space and are a much more useful device for measuring physical quantities than the skewed basis vectors employed in many discussions of rotation.

In order to clarify the discussion of anholonomity which follows, the reader is referred to Refs. 22–24, where he should note the distinction between *natural coordinate* systems and what Schouten often calls anholonomic "coordinates." In this regard the discussions presented concerning the *Lie bracket* are of special significance. We will not derive the geometrical properties of the anholonomic object here.

The coordinate independent covariant derivative is a concept familiar from classical tensor analysis. In relativistic field theory one often employs the covariant curl of the electric vector potential A_a to form the electromagnetic field tensor. With respect to an arbitrary tetrad frame, this operation is given by

$$F_{ab} = \nabla_a A_b - \nabla_b A_a = \frac{\partial A_b}{\partial x^a} - \frac{\partial A_a}{\partial x^b} + 2\,\Omega^c_{ab} A_c, \tag{14}$$

where we write the quantity

$$\Omega^c_{ba} = \tfrac{1}{2}(\Gamma^c_{ab} - \Gamma^c_{ba}). \tag{15}$$

This quantity is a nontensorial object called the *object of anholonomity*.[23,24] The affine connection is symmetri-

cal in a natural (coordinate) basis and so the anholonomic object vanishes when a natural (or holonomic) frame of reference is employed. Note that Ω^c_{ba} is not a tensor. (In particular, it is *not* the torsion tensor, whose existence depends upon properties of the manifold and not the choice of local tetrads.) There are several ways that one may actually compute the Ω^c_{ba}. Given any tetrad frame one may compute the Lie bracket

$$[\mathbf{e}_a, \mathbf{e}_b] = 2\,\Omega^c_{ab}\,\mathbf{e}_c. \tag{16}$$

Given any set of differential forms, one may compute the exterior derivative[23]

$$\mathbf{d}\omega^a = 2\,\Omega^a_{bc}\,\omega^b \wedge \omega^c. \tag{17}$$

Or, given the transformation h^μ_a between a natural frame $\mathbf{e}_\mu$ and an arbitrary tetrad frame $\mathbf{e}_a$, one may perform the transformation

$$\Gamma^a_{bc} = h^a_\alpha h^\mu_b h^\nu_c \Gamma^\alpha_{\mu\nu} - h^\mu_b h^\nu_c \frac{\partial h^a_\nu}{\partial x^\mu} \tag{18}$$

and similarly find Γ^a_{cb}. The difference is then formed and halved to give Ω^a_{bc} as defined above.

The choice of basis vectors will have an important effect upon the interpretation of field components. Since the basis vectors $\mathbf{e}_\mu$ can be arbitrarily chosen at each space–time event, the corresponding metrical components $g_{\mu\nu}$ are fairly arbitrary. How then shall one choose his local reference frames? In such a way as to make his computations as simple as possible to execute, and to interpret physically.

3. THE ANHOLONOMIC OBJECT OF RELATIVISTIC ROTATION

It is now possible to interpret the procedure called the "method of anholonomic frames." The natural 1-forms employed by the rotating observer are found from

$$\omega^a = h^a_\mu \sigma^\mu \tag{19}$$

and are given by

$$\omega^1 = dr, \tag{20a}$$

$$\omega^2 = \gamma(d\theta - \omega\,dt), \tag{20b}$$

$$\omega^3 = dz, \tag{20c}$$

$$\omega^4 = \gamma c\left(dt - \frac{r^2\omega}{c^2}\,d\theta\right). \tag{20d}$$

Noting that

$$\mathbf{d}\omega^1 = \mathbf{d}\omega^3 = 0, \tag{21a}$$

$$\mathbf{d}\omega^2 = \frac{r\omega^2}{c^2}\,\gamma^2\,\omega^1 \wedge \omega^2, \tag{21b}$$

$$\mathbf{d}\omega^4 = \left(-\frac{r\omega^2}{c^2}\gamma^2\right)\omega^1 \wedge \omega^4 - \left(2\frac{r\omega}{c}\gamma^2\right)\omega^1 \wedge \omega^2, \tag{21c}$$

by Eq. (17), one may now read off the nonzero components of the anholonomic object of the relativistically rotating observer as

$$\Omega^2_{12} = -\Omega^2_{21} = \tfrac{1}{2}\gamma^2 r\omega^2/c^2, \tag{22a}$$

$$\Omega^4_{12} = -\Omega^4_{21} = -\gamma^2 r\omega/c, \tag{22b}$$

$$\Omega^4_{14} = -\Omega^4_{41} = -\tfrac{1}{2}\gamma^2 r\omega^2/c^2. \tag{22c}$$

One might also verify that the anholonomic object is zero for the (nonrelativistic) "method of Galilean frames," and is *nonzero* for the (oblique) "method of instantaneous frames."

4. KINEMATIC EFFECTS OF ANHOLONOMITY

A. Thomas precession—A spatial effect

If one carries a defined axis in a given frame of reference along a spatially closed curve with an acceleration not parallel to his velocity, then, as seen by an inertial observer, the reference axis will have rotated through a small angle, $\Delta\alpha$. This effect is called Thomas precession.[25,26] Uhlenbeck and Goudsmit introduced the concept of quantized electron spin angular momentum in 1926 and they were successful in explaining the anomalous Zemann effect, however the spin-orbit interaction that they obtained was twice too large.[27] Thomas' discovery of a purely kinematic precession, independent of the cause of acceleration, reduced the spin-orbit splitting. The Thomas precession arises because under successive Lorentz transformations a compass orientation changes by an amount equivalent to that obtained by a Lorentz transformation plus a spatial rotation.

To see how the Thomas precession may be interpreted as a direct result of anholonomity, one may carry a reference axis along a circular path. After proceeding over the closed path, he will find that his reference axis is not at the starting orientation. The net gap or angular defect is given by

$$\Delta\alpha = \mathbf{d}\omega^2 = 2\int\int \Omega^2_{ab}\,\omega^a \wedge \omega^b. \tag{23}$$

Thus, the rotation produces an angular defect of order

$$\Delta\alpha \cong \frac{\pi r^2\omega^2}{c^2}, \tag{24}$$

in accordance with Thomas' results. This result is, of course, a relativistic effect and is not contained within the "method of Galilean frames."

B. Sagnac effect—As temporal anholonomity

If one follows along a closed path on a rotating disk, one finds upon returning to the starting point that his clock will differ from an inertial clock by an amount Δt. (Furthermore, on the rotating disk, proper time passes differently at r than at $r + dr$. Consequently, global synchronization of clocks in the rotating frame is impossible.)

Confining ourselves to a fixed distance from the origin on the rotating disk, and proceeding to synchronize clocks along a closed curve, upon returning to the starting point, one finds a time differing from the initial clock by the amount

$$\Delta t = \frac{1}{c}\iint \mathbf{d}\omega^4 = \frac{2}{c}\iint \Omega^4_{ab}\,\omega^a \wedge \omega^b. \tag{25}$$

That is,

$$\Delta t \cong \pm 2\omega A/c^2 \tag{26}$$

for $r\omega/c \ll 1$. The $\pm$ refers to following the circuit with or opposed to the direction of rotation.

Sagnac[28] caused two beams of monochromatic light to interfere after passing a closed circular path in opposite directions on a rotating disk. From the above, the time difference for the propagation is

$$\Delta t \cong 4\omega A/c^2, \tag{27}$$

where A is the area enclosed by the circuit.[29,30] This time difference appears as a difference in wavelength (and frequency) between the two beams.[5,31,32] Now,

$$\frac{\Delta f}{f_0} = \frac{\Delta t}{t} \cong \frac{4\omega A}{cL}, \tag{28}$$

where L is the circuit length on a nonrotating disk. The frequency difference between the two beams is then

$$\Delta f \cong 4\omega A/\lambda_0 L. \tag{29}$$

This expression gives the beat frequency between two electromagnetic waves which propagate in opposite directions along a closed path on a rotating disk. Recently, Newburgh *et al.* have employed acoustic and magnetic surface waves for constructing rate gyros.[10] Presumably, the "creeping wave" associated with the geometrical theory of diffraction would also suffer a spectral translation depending, in part, upon the radial dimension of the rotating surface upon which the wave propagates. Apparently this phenomenon has not yet been observed.

The above results are often computed by other methods, where one inherently considers the skewed or off-diagonal terms of the metric.[33] (Remember that $g_{42} \neq 0$ if one has chosen oblique tetrad reference frames.) Employing the metric of Eq. (10), one may compute

$$\Delta t = \frac{1}{c}\int \frac{g_{4a}}{g_{44}}\, dx^a \cong \pm \frac{2\omega A}{c} \tag{30}$$

for $r\omega/c \ll 1$. This approach, however, is not without certain ambiguities. When interpreting the electromagnetic field tensor and the electromagnetic force density, one must bear in mind the oblique character of the local reference directions. For example, the choice of an oblique tetrad frame inherent in the "method of instantaneous frames" will lead to electromagnetic forces proportional to particle velocities, even when a charged particle moves in a purely electric field. Further, the components of the field tensor in the rotating frame, F^{ab}, will have to be interpreted differently from the way one interprets the components of the field tensor, $F^{\mu\nu}$, in an inertial frame.

This problem does not arise in the anholonomic formulation. In that approach, every local tetrad frame is orthogonal and related to the inertial field of frames at every space–time event by a Lorentz transformation, and every local metric (corresponding to the local tangent space spanned by these tetrads) is Minkowskian. As desirable as these properties may be, they are obtained at a high price. These relativistic orthogonal basis vectors are not the tangents to the coordinate axes of any natural coordinate system. (Those of the "method of instantaneous frames" aren't either.) That is, it is not possible to create a global coordinate system which incorporates the local curves to which the tetrad vectors are tangent, but, these tetrads do have the desirable property of orthogonality, and greatly simplify the interpretation of electromagnetic measurements made by rotating observers.

5. ELECTROMAGNETIC EFFECTS OF ANHOLONOMITY

One may obtain Maxwell's source equations, from an invariant action principle, as

$$\nabla_b F^{ab} = (4\pi/c) J^a, \tag{31}$$

where the field tensor is obtained from Eq. (14). Now, the affine connection is unsymmetric in its lower indices in an anholonomic frame, and one must be careful to explicitly display the anholonomic object. This object vanishes in any holonomic frame and we have the definition so familiar from special relativity,

$$F_{ab} = \frac{\partial A_b}{\partial x^a} - \frac{\partial A_a}{\partial x^b}. \tag{32}$$

However, Eq. (32) is not acceptable as a definition of F_{ab} in an anholonomic frame. As Schouten points out, "If any expression with respect to holonomic coordinates is transformed with respect to an anholonomic system, correction terms appear, all containing the object of anholonomity."[34] Since the last term in Eq. (14) is not in general zero, Eq. (32) is not generally valid. Further, since the object of anholonomity is nonvanishing for the "method of instantaneous frames," it is clear that Eq. (32) will not satisfy the principle of general covariance and does not constitute a viable definition of the electromagnetic field tensor.

The "method of anholonomic frames," as we have called it, greatly simplifies the covariance of electrodynamical equations. *All definitions of field variables, and the techniques for handling them, are now exactly the same as the covariant definitions in inertial frames.* The interpretation of physical results is clearly much less complicated when all results are referred to orthogonal reference frames.

At this point it might be instructive to compare the actual step by step computations and physical interpretations of the various methods previously indicated, for some particular example. Without explicitly performing these operations, however, we point out that only for the "method of anholonomic frames" will any path in Fig. 3 lead to an expression for the field tensor whose terms may be uniquely interpreted (just as they are with inertial motion in special relativity), and which satisfies the covariant form of Maxwell's equations. Employing the orthogonal basis vectors of this technique and following any path in Fig. 3 gives an internally consistent field theory.

This is not the case with the "method of instantaneous frames." Furthermore, with that approach, and with the "method of Galilean frames," one employs skewed basis vectors and electromagnetic forces appear which are proportional to charge velocities, even when the charges move in a purely electrostatic field. This curious result, and all the seemingly arbitrary ambiguities must, of course, be interpreted in light of the fact that the reference tetrads are not orthogonal, i.e., $g_{24} \neq 0$. When using these approaches, one has committed himself to describing the universe from an

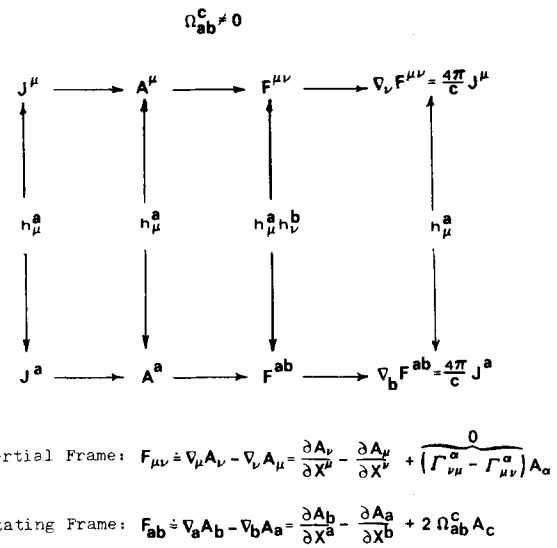


FIG. 3. The anholonomic electromagnetic program. Under the anholonomic Lorentz transformation, h^μ_a, any of the indicated paths lead to a consistent expression for F_{ab}. Furthermore, the rotational paradoxes no longer arise, and the final results satisfy the principle of general covariance.

oblique frame of reference. In this respect, the "method of anholonomic frames" provides one with a simpler set of tetrads for reference.

Which method should one employ in any given situation? This question might be debated purely on pragmatic grounds. However, it is maintained that one should use that technique which is consistent with the application, and the simplest to interpret physically. Conceptual simplicity, orthogonality of reference, and similarity to special relativity all weigh heavily in favor of the "method of anholonomic frames." Only with this method do the *components of the field tensor retain a unique meaning in all frames of reference*. Clearly, it is the logical extension of relativity theory to noninertial systems. Let us now apply the "method of anholonomic frames" to Schiff's paradox.

6. SCHIFF'S PARADOX

In 1939, Schiff[11] attempted to explain the vanishing of the electromagnetic field tensor in the frame of an observer externally rotating about a charged spherical capacitor in terms of Mach's principle. Consider a spherical shell of radius a and negative charge $-Q$ surrounded by a spherical charge distribution of radius b and net charge $+Q$. Three situations present themselves: (1) shell and external observer at rest; (2) external observer at rest, shell in relative uniform motion; (3) shell at rest, external observer in relative rotation with respect to the rest frame of the shell. Schiff's paradox concerns the following arguments. In situation (1), by Gauss' law, there is no external field. Hence, the field tensor $F^{\mu\nu}$ vanishes in both the frame of the observer and in that of the shell ($F^{ab} = F^{\mu\nu} = 0$). In the second situation, the observer at rest sees rotating charges and observes an external magnetic field given by

$$B(r,\theta) = \left[\frac{2\omega Q(b^2-a^2)}{3r^3c}\cos\theta\right]\hat{r} + \left[\frac{\omega Q(b^2-a^2)}{3r^3c}\right]\sin\theta\,\hat{\theta}. \tag{33}$$

Thus, the field tensor is nonzero for the rest observer.

In the third situation, however, the observer circling about sees no magnetic field since, if the tensor $F^{\mu\nu}$ vanishes in the inertial system, then so must the tensor F^{ab} in the rotating system. Schiff's paradox may be stated as a question. "Since the relative motion between charges and observer is the same in both case (2) and (3), should not the observer have the same field tensor in both cases?"

Schiff answered no as follows. In the rotating observer's frame, the distant stars are rotating and this warps his metric tensor (introducting off-diagonal or non-Euclidean terms). If one defines the covariant components of the field tensor to satisfy Maxwell's source equation and then, by means of this warped metric, obtains the contravariant components of the field tensor, the second Maxwell equation will have extra terms. Schiff interpreted these as the physical manifestations of an extra current, induced by the action of the rotating distant masses on the electromagnetic field. The vanishing of the external field tensor is then due to the cancellation of the actual physical current and gravitationally induced, or Machian current.

The "method of anholonomic frames" also provides a negative answer of the above question, but for less philosophical reasons. The anholonomic formulation provides an electromagnetic field tensor whose components give the negative shell's fields as

$$E_r = -\gamma Q/r, \tag{34a}$$

$$B_\theta = -\gamma Q\omega\sin\theta/rc \quad (r > a), \tag{34b}$$

and the positive shell's fields as

$$E_r = \gamma Q/r, \tag{35a}$$

$$B_\theta = \gamma Q\omega\sin\theta/rc \quad (r > b), \tag{35b}$$

in the frame of the rotating observer. Hence, by the principle of superposition, the fields vanish external to the capacitor.

Furthermore, if one examines the reference frames that Schiff employed, he finds that they are oblique and arise as a result of a local coordinate rotation. Consequently his metric should not be diagonal ($\mathbf{e}_2 \cdot \mathbf{e}_4 = g_{24} \neq 0$). However, if one employs a relativistic Lorentz rotation, he obtains an orthogonal field of anholonomic reference tetrads. The electromagnetic field tensor, produced by the charged capacitor, vanishes with respect to these frames simply by the principle of linear superposition, as stated above. The paradox is then resolved locally without the use of absolute space or distant masses.[35]

7. CONCLUSION

The essence of this investigation has been that the Lorentz transformation, for rotational motion, is an anholonomic transformation. Consequently, there

exists no global t' coordinate corresponding to an inertial t coordinate. It has been argued that a holonomic approach is not satisfactory for relativistic rotation, the fundamental distinction being that the anholonomic formulation provides an orthogonal reference frame while the holonomic approach employs an oblique frame of reference. The price one pays is that now the object of anholonomity *must* be employed. That is, properties of the affine connection may not be neglected in anholonomic frames. This problem does not arise under linear, uniform motion since the rectilinear Lorentz transformations are holonomic and the transformed affinity remains symmetric. The anholonomic formulation of electrodynamics (see Fig. 3) is posed as a tenable explanation of rotational relativity in flat space–time. It should also be noted that unless one employs those anholonomic objects which are associated with the commonly used "method of instantaneous frames," the electrodynamical equations will not be covariant. If one neglects the Ω^{a}_{bc}, instead of Fig. 3, he will be faced with an inconsistent, noncovariant electrodynamics. The situation for skewed basis vectors can be made covariant simply by remembering to include any anholonomic effects. But, then one should ask himself, "If the situation is already anholonomic, why not use orthogonal reference vectors?"

ACKNOWLEDGMENT

The author wishes to gratefully acknowledge beneficial discussions with the following people: Professors H. C. Ko, C. V. Heer, and U. H. Gerlach, all of the Ohio State University, and Dr. C. Chuang of the OSU Electroscience Laboratory.

[1]J. van Bladel, Proc. IEEE **61**, 260–68 (1973).
[2]T. Shiozawa, Proc. IEEE **61**, 1694–702 (1973).
[3]J. C. Hafele, Am. J. Phys. **40**, 81–85 (1972).
[4]T. C. Mo, J. Math. Phys. **11**, 2589–610 (1970).
[5]W. H. Weihofen, Am. J. Phys. **43**, 39–46 (1975).
[6]H. A. Atwater, Proc. IEEE **63**, 316–17 (1975).
[7]T. Shiozawa, remarks in reply to Ref. 6, Proc. IEEE **63**, 317–18 (1975).
[8]J. van Bladel, Proc. IEEE **64**, 301–18 (1976).
[9]J. Killpatrick, IEEE Spectrum **4**, 44–55 (1967).
[10]R. Newburgh, P. Blacksmith, A. J. Budreau, and J. C. Sethares, Proc. IEEE **62**, 1621–28 (1974).
[11]L. I. Schiff, Proc. Natl. Acad. Sci. **25**, 391–95 (1939).
[12]P. Ehrenfest, Phys. Z. **10**, 918 (1909).
[13]M. Planck, Phys. Z. **11**, 294 (1910).
[14]The author is grateful to an anonymous reviewer for his comments on this point.
[15]R. Adler, M. Bazin, and M. Schiffer, *Introduction to General Relativity* (McGraw-Hill, New York, 1965), pp. 110–16, 118–20.
[16]W. Pauli, *Theory of Relativity* (Pergamon, New York, 1958), p. 151.
[17]C. Møller, *The Theory of Relativity* (Oxford U. P., New York, 1952), p. 247.
[18]Y. B. Zeldovich and I. D. Novikov, *Relativistic Astrophysics, Vol. 1-Stars and Relativity* (University of Chicago Press, Chicago, 1971), pp. 9–16, 20.
[19]L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* (Addison-Wesley, Reading, Massachusetts, 1962), pp. 259–62, 296–98.
[20]F. A. E. Pirani, Bull. Acad. Pol. Sci. **5**, 143–46 (1957).
[21]W. M. Irvine, Physica **30**, 1160–70 (1964).
[22]C. Misner, K. Thorne, and J. Wheeler, *Gravitation* (Preliminary version, duplicated Fall 1970), p. 124.
[23]J. A. Schouten, *Ricci Calculus* (Springer-Verlag, Berlin, 1954), pp. 99–110, 117–21, 169–73.
[24]J. A. Schouten, *Einfuhrung in die Neueren Methoden der Differential geometrie* (Noordhoff, Groningen, 1935), Vol. 1, pp. 138–40.
[25]L. H. Thomas, Nature **117**, 514 (1926).
[26]L. H. Thomas, Phil. Mag. **7**, 1 (1927).
[27]J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1963).
[28]G. Sagnac, C. R. Acad. Sci. (Paris) **157**, 708, 1410 (1913); J. Phys. Radium **4**, 177–95 (1914).
[29]E. J. Post, Rev. Mod. Phys. **39**, 475–93 (1967).
[30]E. J. Post, J. Opt. Soc. Am. **62**, 234–39 (1972).
[31]C. V. Heer, Phys. Rev. **134A**, 799–804 (1964).
[32]C. V. Heer, J. Little, and N. Bupp, Ann. Inst. H. Poincaré, Ser. A, **8**, 311–26 (1968).
[33]Ref. 19, p. 296.
[34]Ref. 23, p. 100.
[35]In this respect, compare the comments of DeSitter in: W. DeSitter, Proc. R. Acad. Amsterdam **19**, 527–32 (1916); **19**, 1217–25 (1917).

Discrete spectrum for a system of electrically charged particles

Metin Demiralp

Applied Mathematics Division, Marmara Scientific and Industrial Research Institute, P.K. 141 Kadiköy-Istanbul, Turkey

Erdoğan Şuhubi

Applied Mathematics Division, Marmara Scientific and Industrial Research Institute and Department of Theoretical Mechanics, Faculty of Sciences, Technical University of Istanbul, Turkey
(Received 12 April 1976; revised manuscript received 28 June 1976)

For a system of electrically charged particles, the time-independent Schrödinger's equation is transformed into hyperspherical coordinates. By using the hyperharmonic expansion, the hyperradial equation is obtained, and its discrete spectrum is investigated. To this end, only negative energy values are taken into consideration. The functional equation which yields the discrete spectrum of the hyperradial equation is obtained by means of a matching technique in the Laplace transform domain of hyperradial Schrödinger's equation.

1. INTRODUCTION

In recent years there has been an increased interest in the use of hyperspherical coordinates for evaluation of the energy spectra of several atoms or molecules.[1–6] In Knirk's paper[3] a general approach to the problem for atomic systems is proposed. In addition, some authors made use of variational test functions expressed in these coordinates to obtain the energy spectra of rather simple systems such as the helium atom.[7] Indeed, the use of hyperspherical coordinates facilitates the general treatment of such problems in that a system composed of electrically charged particles, as well as atoms or molecules, can be investigated in a general way and several integrals appearing in the analysis can be easily evaluated.

In this paper, we somewhat generalize this problem to electrically charged particles by transforming the coordinate frame to a hyperspherical one and, using the hyperharmonic expansion, we obtain an infinite system of coupled ordinary differential equations. By truncating this system of equations and by reducing it to a first order system, we arrive at a differential equation in the same form as the one studied by Okubo.[8] By relating the regular solutions to asymptotic ones of this equation and apply the boundary condition at infinity we obtain a functional relation for possible discrete energy values of the system.

2. DEFINITION OF SYSTEM AND SCHRÖDINGER'S EQUATION

We consider a system composed of $N+1$ electrically charged particles under no external perturbations. Let M_j denote the mass of jth particle and Z_j "electrical charge parameter" of this particle. If we let $\hat{x}_{3j-2}$, $\hat{x}_{3j-1}$, $\hat{x}_{3j}$ denote the spatial coordinates of jth particle and if we adopt the ordering of particles according to the magnitude of their masses, i.e., $M_1 \leq \cdots \leq M_{N+1}$, then we can write the time-independent Schrödinger's equation of the system as follows:

$$-\sum_{j=1}^{N+1}\frac{\hbar^2}{2M_j}\nabla_j^2\hat{\psi}+\sum_{j=1}^{N}\sum_{k=j+1}^{N+1}\frac{Z_jZ_ke^2}{r_{jk}}\hat{\psi}=\hat{E}\hat{\psi}. \tag{2.1}$$

Here r_{jk} represents the distance between jth and kth particles. "Electrical charge parameter" Z_j's are defined as the ratio of electrical charge of jth particle to that of proton. If we transform the coordinate system to a frame at the mass center[9] and separate the variables of mass center, the number of independent variables is reduced to $3N$. By choosing an appropriate diagonalizating transformation on these $3N$ new variables and introducing dimensionless coordinates and parameters we can arrive at the following equation:

$$\left(\tilde{\nabla}^2-\sum_{j=1}^{N}\sum_{k=j+1}^{N+1}\frac{Z_jZ_k}{(\tilde{x}^TV_{jk}\tilde{x})^{1/2}}+E\right)\psi=0. \tag{2.2}$$

Here $\tilde{x}$, $\tilde{\nabla}$ are used for $3N$-dimensional position and gradient vectors respectively and V_{jk}'s stand for some $3N$-dimensional square matrices depending only on mass ratios such that each of them has only one triple nonvanishing eigenvalue. E is used for dimensionless energy, and T denotes the transposition.

Since we only investigate the discrete spectrum of the above equation, and discrete spectrum (bound states) corresponds to negative energy values, we can write the following equation

$$E=-\mathcal{E}^2, \tag{2.3}$$

where without loss of generality $\mathcal{E}$ is taken as a positive scalar.

By making use of a new coordinate frame defined below

$$x=2\mathcal{E}\tilde{x} \tag{2.4}$$

we obtain the following equation:

$$\left(\nabla^2-\frac{1}{4}-\frac{1}{2\mathcal{E}}\sum_{j=1}^{N}\sum_{k=j+1}^{N+1}\frac{Z_jZ_k}{(x^TV_{jk}x)^{1/2}}\right)\psi=0. \tag{2.5}$$

3. HYPERSPHERICAL COORDINATES AND HYPERRADIAL EQUATION

To solve (2.5), let us change the coordinates into the hyperspherical ones as follows

$$x_{3N}=r\cos\theta_{3N-1}, \qquad 0\leq r\leq\infty,$$

Copyright © 1977 American Institute of Physics

$$x_{3N-1}=r\sin\theta_{3N-1}\cos\theta_{3N-2},\quad 0\le\theta_2,\ldots,\theta_{3N-1}\le\pi,$$
$$\cdots$$
$$x_1=r\sin\theta_{3N-1}\cdots\sin\theta_1,\quad 0\le\theta_1\le 2\pi. \tag{3.1}$$

Employing these coordinates we obtain from (2.5)

$$\left[\frac{1}{r^{3N-1}}\frac{\partial}{\partial r}\left(r^{3N-1}\frac{\partial}{\partial r}\right)+\frac{1}{r^2}L - \frac{1}{r}\left(\sum_{j=1}^{N}\sum_{k=j+1}^{N+1}\frac{1}{2\mathcal{E}}\frac{Z_jZ_k}{(\xi^T V_{jk}\xi)^{1/2}}\right)-\frac{1}{4}\right]\psi=0. \tag{3.2}$$

Here ξ is used for the unit vector x/r which can be expressed in terms of θ's only. The operator L appearing above may be called "Total hyper angular momentum operator" and defined explicitly in the following form:

$$L=\sum_{j=0}^{3N-2}\frac{1}{(\prod_{k=1}^{j}\sin^2\theta_{3N-k})\sin^{3N-2-j}\theta_{3N-1-j}} \times\frac{\partial}{\partial\theta_{3N-1-j}}\left(\sin^{3N-2-j}\theta_{3N-1-j}\frac{\partial}{\partial\theta_{3N-1-j}}\right). \tag{3.3}$$

The eigenfunctions of this operator, which are finite and continuous on the closed domain of θ's, have been thoroughly investigated. They are known as "hyperspherical surface harmonics" (or briefly hyper harmonics) in literature.[10] These eigenfunctions and corresponding eigenvalues are given below:

$$LY=lY, \tag{3.4}$$
$$l=-l_{3N-1}(l_{3N-1}+3N-2),\quad l_1,\ldots,l_{3N-1}\text{ integers},$$
$$0\le l_2\le l_3\le\cdots\le l_{3N-1},\qquad -l_2\le l_1\le l_2, \tag{3.5}$$

$$Y_l(\theta)=\frac{\exp(il_1\theta_1)}{(2\pi)^{1/2}}\left[\frac{(2l_2+1)(l_2-|l_1|)!}{2(l_2+|l_1|)!}\right]^{1/2}\frac{\Gamma(|l_1|+\frac{1}{2})}{\sqrt{\pi}} \times 2^{|l_1|}\sin^{|l_1|}\theta_2 C_{l_2-|l_1|}^{|l_1|+1}(\cos\theta_2) \times\prod_{j=1}^{3N-3}\left\{\frac{(2l_{3N-j}+3N-j-1)(l_{3N-j}-l_{3N-j-1})!}{2(l_{3N-j}+l_{3N-j-1}+3N-j-2)!}\right\}^{1/2} \times\frac{\Gamma(l_{3N-j-1}+(3N-j-1)/2)}{\sqrt{\pi}}2^{[(3N-j-2)/2]+l_{3N-j-1}} \times\sin^{l_{3N-j-1}}\theta_{3N-j}C_{l_{3N-j}-l_{3N-j-1}}^{l_{3N-j-1}+(3N-j-1)/2}(\cos\theta_{3N-j}). \tag{3.6}$$

In the last equation, and henceforth, we use the notation l and θ for the set of indices $l_1,\ldots,l_{3N-1}$ and the set of angles $\theta_1,\ldots,\theta_{3N-1}$, respectively.[11] The functions denoted by C_n^λ are "Gegenbauer polynomials."[12]

The harmonics given above are (l_{3N-1})th order homogeneous functions of ξ. Let us employ the symbol Y_{jk} for a hyperspherical harmonic which has an order of homogeneity j. In this notation k is used for an ordering in the set of hyperspherical harmonics which are homogeneous of order j. In addition we denote the maximum value of k with ν_j.

Now, the theory of hyperspherical harmonics makes it possible to express a function which is continuous and square integrable over the whole domain of θ's, in terms of the Y_{jk}'s. Since the admissible solutions of (3.2) also have these properties, we can assume the following expansion:

$$\psi=\sum_{j,k}f_{jk}(r)\,Y_{jk}. \tag{3.7}$$

Here the f_{jk}'s are unknown functions of r.

The hyperspherical harmonics with their structures given by (3.6) form an orthonormal set on unit hypersphere S; hence

$$\int_S\bar{Y}_{jk}Y_{lm}\,dS=\delta_{jl}\delta_{km}. \tag{3.8}$$

Now let us define the following matrices:

$$(L_{jk})_{lm}=\int_S\bar{Y}_{jl}LY_{km}\,dS,\quad l=1,\ldots,\nu_j,\ m=1,\ldots,\nu_k, \tag{3.9}$$

$$(P_{jk})_{lm}=-\sum_{n=1}^{N}\sum_{q=n+1}^{N+1}\frac{Z_nZ_q}{2}\int_S\frac{\bar{Y}_{jl}Y_{km}}{(\xi^T V_{nq}\xi)^{1/2}}\,dS \qquad l=1,\ldots,\nu_j,\ m=1,\ldots,\nu_k. \tag{3.10}$$

By using $j+1$ and $k+1$ as row and column numbers, respectively, we can construct infinite square matrices L and P which have the block elements L_{jk} and P_{jk}, respectively. In fact these matrices are none other than the matrix representations of the differential operator L and algebraic potential operator appearing in (3.10) in the space spanned by hyper harmonics.

If we define an infinite-dimensional vector f composed by f_{jk}'s, after some intermediate steps, we can arrive at the following vector differential equation in infinite-dimensional Euclidean space:

$$f''+\frac{3N-1}{r}f'+\frac{1}{r^2}Lf+\frac{1}{\mathcal{E}r}Pf-\tfrac{1}{4}f=0. \tag{3.11}$$

Now, our problem is to find the solutions of this equation which are finite and continuous for all values of r and vanish as r increases unboundedly.

If we exclude all blocks f_{jl}, L_{jk}, P_{jk} $(j,k>n)$ in (3.11), we obtain a system of coupled differential equations of finite number which can be called "truncated hyperradial equation." By denoting these truncated entities by f_n, L_n, and P_n, (3.11) can be brought into the following form:

$$f_n''+\frac{3N-1}{r}f_n'+\frac{1}{r}L_nf_n+\frac{1}{\mathcal{E}r}P_nf_n-\frac{1}{4}f_n=0. \tag{3.12}$$

We can easily see that L_n and P_n are d-dimensional square matrices (d being equal to $\nu_0+\cdots+\nu_n$) and f_n is a d-dimensional vector.

In the limiting case $n\to\infty$ this last equation coincides with (3.11), but it is somewhat easier to study this equation rather than (3.11).

By using the following transformations,

$$f_n=e^{-r/2}F^{(1)},\quad rf_n'=F^{(2)},\quad F^T=[F^{(1)},F^{(2)}] \qquad (2d\text{-dimensional vector}), \tag{3.13}$$

the order of (3.12) can be reduced to one. Hence from (3.12) we obtain the following equation studied by Okubo[8]:

$$F'=[(A/r)+B]F, \tag{3.14}$$

where A and B are $2d$-dimensional square matrices and are defined as follows:

$$A=\begin{bmatrix} 0 & I_d \\ -L_n & (2-3N)I_d \end{bmatrix},$$

$$B=\begin{bmatrix} 0 & 0 \\ \frac{3N-1}{2}I_d - \frac{P_n}{\mathcal{E}} & I_d \end{bmatrix}, \tag{3.15}$$

where I_d represents the d-dimensional unit matrix.

(3.14) has a regular singular point located at the origin ($r=0$) and an irregular singular point at infinity.

The $2d$ linearly independent solutions of (3.14) expanded at the origin are obtained in the usual way with the aid of Frobenius method. These solutions are given below:

$$F^{(+)}_{jk}(r)=\left\{\frac{\partial^{n-j}}{\partial\sigma^{n-j}}\left[(\sigma-j)^{n-j}\Phi\right]\right\}_{\sigma=j}(a^+_R)_{jk},$$
$$j=0,\ldots,n,\quad k=1,\ldots,\nu_j, \tag{3.16}$$

$$F^{(-)}_{jk}(r)=\left\{\frac{\partial^{n+j+1}}{\partial\sigma^{n+j+1}}\left[(\sigma+j+3N-2)^{n+j+1}\Phi\right]\right\}_{\sigma=2-j-3N}(a^-_R)_{jk}$$
$$j=0,\ldots,n,\quad k=1,\ldots,\nu_j. \tag{3.17}$$

Here Φ denotes a generating matrix-valued function defined as follows:

$$\Phi(r;\sigma)=\sum_{k=0}^{\infty} r^{\sigma+k}[(\sigma+k)I_{2d}-A]^{-1}B\cdots[(\sigma+1)I_{2d}-A]^{-1}B \tag{3.18}$$

and $(a^+_R)_{jk}$ and $(a^-_R)_{jk}$ are right eigenvectors of A.

From (3.4), (3.5), (3.8), and (3.9) we can obtain

$$L_{jk}=0,\quad j\neq k,\quad L_{jj}=-j(j+3N-2)I_{\nu_j}. \tag{3.19}$$

Finally after tedious algebraic manipulations we arrive at the following eigenvectors and eigenvalues for A:

$$A(a^+_R)_{jk}=j(a^+_R)_{jk},$$
$$(a^+_R)^T_{jk}=[e^T_{\nu_0+\cdots+\nu_{j-1}+k}, je^T_{\nu_0+\cdots+\nu_{j-1}+k}]$$
$$j=0,\ldots,n,\quad k=1,\ldots,\nu_j, \tag{3.20}$$

$$A(a^-_R)_{jk}=-(j+3N-2)(a^-_R)_{jk}$$
$$(a^-_R)^T_{jk}=[e^T_{\nu_0+\cdots+\nu_{j-1}+k}, -(j+3N-2)e^T_{\nu_0+\cdots+\nu_{j-1}+k}],$$
$$j=0,\ldots,n\quad k=1,\ldots,\nu_j, \tag{3.21}$$

$$A^T(a^+_L)_{jk}=j(a^+_L)_{jk},$$
$$(a^+_L)^T_{jk}=\frac{1}{(2j+3N-2)}$$
$$\times[(j+3N-2)e^T_{\nu_0+\cdots+\nu_{j-1}+k}, e^T_{\nu_0+\cdots+\nu_{j-1}+k}],$$
$$j=0,\ldots,n,\quad k=1,\ldots,\nu_j, \tag{3.22}$$

$$A^T(a^-_L)_{jk}=-(j+3N-2)(a^-_L)_{jk}$$
$$(a^-_L)^T_{jk}=\frac{1}{2j+3N-2}[je^T_{\nu_0+\cdots+\nu_{j-1}+k}, -e^T_{\nu_0+\cdots+\nu_{j-1}+k}],$$
$$j=0,\ldots,n,\quad k=1,\ldots,\nu_j. \tag{3.23}$$

Here e_ν's denote Cartesian unit vectors in d-dimensional Euclidean space.

4. ASYMPTOTIC SOLUTIONS OF (3.14)

To find the behavior of (3.14) at infinity, we assume a solution in the following form:

$$F=e^{-\beta r}\sum_{k=0}^{\infty} r^{\alpha-k-\xi}C_k, \tag{4.1}$$

where the C_k's are constant vectors. After inserting this into (3.14) we arrive at the following recurrence relation:

$$(B-\beta I_{2d})C_k=-\{A+(\xi+k-1)I_{2d}\}C_{k-1},\quad C_{-1}\equiv 0, \tag{4.2}$$

which gives the indicial equation for $k=0$

$$(B-\beta I_{2d})C_0=0. \tag{4.3}$$

By solving the above eigenvalue problem, the following results are obtained:

$$B(b_{0R})_j=0,\quad (b_{0R})^T_j=\left[e^T_j, e^T_j\left(\frac{P_n}{\mathcal{E}}-\frac{3N-1}{2}I_d\right)\right]$$
$$B(b_{1R})_j=(b_{1R})_j,\quad (b_{1R})^T_j=[0, e^T_j]$$
$$B^T(b_{0L})_j=0,\quad (b_{0L})^T_j=[e^T_j, 0]$$
$$B^T(b_{1L})_j=(b_{1L})_j,\quad (b_{1L})^T_j=\left[e^T_j\left(\frac{3N-1}{2}I_d-\frac{P_n}{\mathcal{E}}\right), e^T_j\right]$$
$$j=1,\ldots,d. \tag{4.4}$$

Now we consider the following transformation in (4.2):

$$C_k=Q_B\hat{C}_k,\quad \hat{C}^T_k=[\hat{C}^{(1)}_k, \hat{C}^{(2)}_k], \tag{4.5}$$

where

$$Q_B=\begin{bmatrix} I_d & 0 \\ \left(\frac{P_n}{\mathcal{E}}-\frac{3N-1}{2}I_d\right) & I_d \end{bmatrix}. \tag{4.6}$$

Hence we have clearly the explicit forms

$$\beta\hat{C}^{(1)}_k=\left[\left(\xi+k-\frac{3N-1}{2}\right)I_d+\frac{P_n}{\mathcal{E}}\right]\hat{C}^{(1)}_{k-1}+\hat{C}^{(2)}_{k-1}, \tag{4.7}$$

$$(\beta-1)\hat{C}^{(2)}_k=-\left[L_n+\left(\frac{P_n}{\mathcal{E}}+\frac{1-3N}{2}I_d\right)\left(\frac{P_n}{\mathcal{E}}+\frac{3N-3}{2}I_d\right)\right]$$
$$-\hat{C}^{(1)}_{k-1}+\left[\left(\xi+k+\frac{1-3N}{2}\right)I_d-\frac{P_n}{\mathcal{E}}\right]\hat{C}^{(2)}_{k-1}, \tag{4.8}$$

$$\hat{C}^{(1)}_{-1}=\hat{C}^{(2)}_{-1}=0.$$

For $k=0$ we find that $\hat{C}^{(2)}_0=0$ and $\hat{C}^{(1)}_0$ remains arbitrary when β vanishes. In the case of $k=1$ (4.7) becomes

$$\left[\left(\xi+\frac{1-3N}{2}\right)I_d+\frac{P_n}{\mathcal{E}}\right]\hat{C}^{(1)}_0=0. \tag{4.9}$$

However, this equation indicates the eigenvalue problem of $P_n/\mathcal{E}$. As can be seen from its definition the matrix $P_n/\mathcal{E}$ is symmetric; hence all of its eigenvalues are real. By depending on the value of $\mathcal{E}$, some eigenvalues may differ from each other by positive integers. In such cases some difficulties appear, and they can be removed by differentiating a few times an appropriate generating matrix function with respect to the exponent ξ. If we exclude these values of $\mathcal{E}$ and denote the eigenvalues and eigenvectors of P_n by π_j and p_j ($j=1,\ldots,d$), respectively we can proceed in the following way.

Let us rearrange Eqs. (4.7) and (4.8) for $\beta=0$ as follows:

$$\begin{bmatrix} \left(\zeta+k+\frac{1-3N}{2}\right)I_d & I_d \\ 0 & I_d \end{bmatrix}\begin{bmatrix} \hat{C}_k^{(1)} \\ \hat{C}_k^{(2)} \end{bmatrix}$$

$$=\begin{bmatrix} 0 \\ L_n+\left(\frac{P_n}{\mathcal{E}}+\frac{1-3N}{2}I_d\right)\left(\frac{P_n}{\mathcal{E}}+\frac{3N-3}{2}I_d\right) \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ \left(\zeta+k+\frac{1-3N}{2}\right)I_d-\frac{P_n}{\mathcal{E}} \end{bmatrix}\begin{bmatrix} \hat{C}_{k-1}^{(1)} \\ \hat{C}_{k-2}^{(1)} \end{bmatrix},$$

$$\hat{C}_{-1}^{(1)}=\hat{C}_{-1}^{(2)}\equiv 0. \tag{4.10}$$

If we suppose that ζ and $\hat{C}_0^{(1)}$ take the values

$$\zeta_j=\frac{3N-1}{2}-\frac{\pi_j}{\mathcal{E}}, \quad \hat{C}_0^{(1)}=P_j, \quad j=1,\ldots,d, \tag{4.11}$$

we can find all solutions of (3.14).

We conclude now that the matrix $\{[\zeta_j+k+(1-3N)/2]I_d+P_n/\mathcal{E}\}$ and also the coefficient matrix in the left-hand side of (4.10) are nonsingular for all values of ζ given above and for all positive k's. Therefore, by utilizing (4.10) we can express the vectors $\hat{C}_k^{(1)}$ and $\hat{C}_k^{(2)}$ in terms of $\hat{C}_0^{(1)}$.

For a more general approach let us define a matrix in the following manner:

$$A_0(k,\zeta)=\begin{bmatrix} A_0^{11}(k,\zeta) & A_0^{12}(k,\zeta) \\ A_0^{21}(k,\zeta) & A_0^{22}(k,\zeta) \end{bmatrix},$$

$$A_0^{11}=\left[\left(\zeta+k+\frac{1-3N}{2}\right)I_d+\frac{P_n}{\mathcal{E}}\right]^{-1}A_0^{21},$$

$$A_0^{12}=-\left[\left(\zeta+k+\frac{1-3N}{2}\right)I_d+\frac{P_n}{\mathcal{E}}\right]^{-1}A_0^{22},$$

$$A_0^{21}=L_n+\left(\frac{P_n}{\mathcal{E}}+\frac{1-3N}{2}I_d\right)\left(\frac{P_n}{\mathcal{E}}+\frac{3N-3}{2}I_d\right),$$

$$A_0^{22}=\left(\zeta+k+\frac{1-3N}{2}\right)I_d-\frac{P_n}{\mathcal{E}}. \tag{4.12}$$

By employing this matrix, after several algebraic manipulations we arrive at the following results:

$$C_k=Q_B\left\{\prod_{l=0}^{k-1}A_0(k-l,\zeta_j)\right\}\begin{bmatrix} p_j \\ 0 \end{bmatrix}, \quad j=1,\ldots,d. \tag{4.13}$$

$$F_j^{(AS,0)}=Q_B\left\{\sum_{k=0}^{\infty}r^{-k-\zeta_j}\prod_{l=0}^{k-1}A_0(k-l,\zeta_j)\right\}\begin{bmatrix} p_j \\ 0 \end{bmatrix},$$

$$j=1,\ldots,d. \tag{4.14}$$

In the case of $\beta=1$, we conclude after a similar argument that

$$F_j^{(AS,1)}=Q_B\left\{\sum_{k=0}^{\infty}r^{-k-(3N-3)/2-\pi_j/\mathcal{E}}\right.$$

$$\left.\times\prod_{l=0}^{k-1}A_1\left(k-l,\frac{\pi_j}{\mathcal{E}}+\frac{3N-3}{2}\right)\right\}\begin{bmatrix} 0 \\ p_j \end{bmatrix}, \tag{4.15}$$

where the matrix A_1 is composed of the following four blocks

$$(A_1(k,\zeta))_{11}=\left(\zeta+k-\frac{3N-1}{2}\right)I_d+\frac{P_n}{\mathcal{E}},$$

$$(A_1(k,\zeta))_{12}=I_d,$$

$$(A_1(k,\zeta))_{21}=(A_1(k,\zeta))_{22}(A_1(k,\zeta))_{11}$$

$$(A_1(k,\zeta))_{22}=\left[\left(\zeta+k+\frac{3-3N}{2}\right)I_d-\frac{P_n}{\mathcal{E}}\right]^{-1}$$

$$\times\left[L_n+\left(\frac{P_n}{\mathcal{E}}+\frac{1-3N}{2}I_d\right)\left(\frac{P_n}{\mathcal{E}}+\frac{3N-3}{2}I_d\right)\right]. \tag{4.16}$$

Therefore, all solutions of (3.14) at any point in the whole complex space of r can be evaluated in terms of functions given by (3.16) and (3.17). The last functions given by (4.14) and (4.15) have an asymptotic character. But they can be employed in evaluating the value of any function given by (3.16) and (3.17) for large values of r. To this end, it is necessary to find the expression of any regular function (in the neighborhood of zero) in terms of the asymptotic ones.

5. MATCHING BY LAPLACE TRANSFORMATION

From the properties of Laplace transformation,[13] Eq. (3.14) can be put into the following form:

$$(B-sI_{2d})\frac{dg}{ds}=(A+I_{2d})g. \tag{5.1}$$

Here, $g(s)$ denotes the transform of F,

$$g(s)=\int_0^{\infty}\exp(-sr)F(r)\,dr. \tag{5.2}$$

We can easily see that the singular points of (5.1) are located at $s=0$, $s=1$, $s=\infty$. But all of these points are regular singular points; hence, any solution of (3.1) at any one of these points are convergent in a disk centered at this point in the complex plane of s, and the radius of this disk is equal to the distance between this point and nearest singular point. This situation is illustrated in Fig. 1.

If we denote the domain of convergence of the solutions expanded at the points 0, 1, ∞, respectively by R_0, R_1, R_∞, we can deduce that some binary intersections are not empty

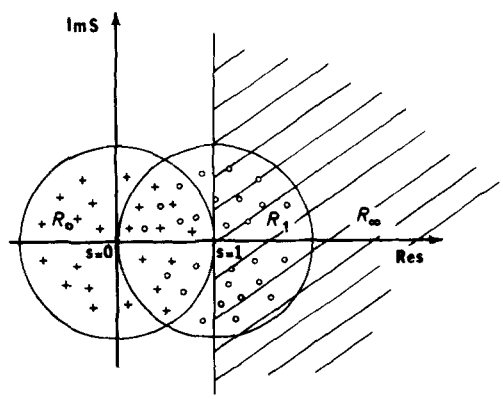


FIG. 1. Domains of convergence of the solutions of (5.1).

$$R_0 \cap R_1 \neq \phi, \quad R_1 \cap R_\infty \neq \phi. \tag{5.3}$$

Hence the solutions expanded at the point $s=0$ can be analytically continued to the domain of the solutions expanded at $s=1$, and these in turn can be analytically continued to the domain R_∞. However, this implies that each of the solutions at any point of singularity can be expressed in terms of those expanded at another singular point. But, to accomplish this, it is necessary to find the totality of solutions expanded at these points of singularity.

At the points $s=0$, $s=1$, $s=\infty$, the solutions of (5.1) can be found with the aid of the method of series solution which was also employed for determination of the solutions of (3.14). By omitting intermediate discussion we can write the general solution of (5.1) for large values of s as follows:

$$g_{kl}=\left\{\frac{\partial^{n-k}}{\partial\sigma^{n-k}}[(\sigma-k)^{n-k}\varphi_R]\right\}_{\sigma=k}(a_R^+)_{kl},$$
$$k=0,\ldots,n, \quad l=1,\ldots,\nu_k,$$
$$h_{kl}=\left\{\frac{\partial^{n+k+1}}{\partial\sigma^{n+k+1}}[(\sigma+k+3N-2)^{n+k+1}\varphi_R]\right\}_{\sigma=2-k-3N}(a_R^-)_{kl},$$
$$k=0,\ldots,n, \quad l=1,\ldots,\nu_k. \tag{5.4}$$

Here φ_R stands for a generating matrix-valued function defined below:

$$\varphi_R=\sum_{k=0}^{\infty}\frac{\Gamma(\sigma+k+1)}{s^{\sigma+k+1}}\left(\prod_{l=0}^{k-1}[(\sigma+k-l)I_{2d}-A]^{-1}B\right). \tag{5.5}$$

These vectors g_{kl}, h_{kl} are called the right solutions of (5.1), and its left solutions satisfy the following adjoint equation of (5.1):

$$\frac{d\gamma}{d(1/s)}=-s(A^T+I_{2d})\left(I_{2d}-\frac{B^T}{s}\right)^{-1}\gamma. \tag{5.6}$$

If we introduce $\bar{\gamma}$ through

$$\gamma=\left(I_{2d}-\frac{1}{s}B^T\right)\bar{\gamma}, \tag{5.7}$$

we obtain

$$\left(I_{2d}-\frac{1}{s}B^T\right)\frac{d\bar{\gamma}}{d(1/s)}=[B^T-s(A^T+I_{2d})]\bar{\gamma}. \tag{5.8}$$

The series solution of this equation in terms of $1/s$ can be obtained in the usual manner, and it follows from (5.7) that

$$\gamma_{kl}=\left(I_{2D}-\frac{B^T}{s}\right)\left\{\frac{\partial^{n+k+1}}{\partial\sigma^{n+k+1}}[(\sigma+k)^{n+k+1}\varphi_L]\right\}_{\sigma=-k}(a_L^+)_{kl},$$
$$\chi_{kl}=\left(I_{2d}-\frac{B^T}{s}\right)\left\{\frac{\partial^{n-k}}{\partial\sigma^{n-k}}[(\sigma+2-3N-k)^{n-k}\varphi_L]\right\}_{\sigma=k+3N-2}(a_L^-)_{kl},$$
$$k=0,\ldots,n, \quad l=1,\ldots,\nu_k, \tag{5.9}$$

where φ_L denotes a generating matrix-valued function

$$\varphi_L=\sum_{k=0}^{\infty}\frac{\Gamma(\sigma+k)}{\Gamma(\sigma)s^{\sigma+k-1}}\left(\sum_{l=0}^{k-1}[(\sigma+k-l)I_{2d}+A^T]^{-1}B^T\right). \tag{5.10}$$

We can easily prove that these functions have the following properties:

$$\gamma_{jk}^T g_{lm}=W_{jklm}^{(1)}, \quad \gamma_{jk}^T h_{lm}=W_{jklm}^{(2)},$$
$$\chi_{jk}^T g_{lm}=W_{jklm}^{(3)}, \quad \chi_{jk}^T h_{lm}=W_{jklm}^{(4)}, \tag{5.11}$$
$$j,l=0,\ldots,n, \quad k=1,\ldots,\nu_j, \quad m=1,\ldots,\nu_l,$$

where the W's represent some constants.

These properties and the others which will be obtained later prove to be crucial when we attempt to match a solution of (5.1) expanded at a point of singularity with those expanded at another singular point.

6. SOLUTIONS OF (5.1) AT $s=1$

Let us assume a solution of (5.1) in the form

$$g=\sum_{k=0}^{\infty}(s-1)^{k+\zeta-1}g_k. \tag{6.1}$$

Substituting (6.1) into (5.1), we get

$$(k+\zeta-1)(B-I_{2d})g_k=[A+(k+\zeta-1)I_{2d}]g_{k-1}, \quad g_{-1}\equiv 0. \tag{6.2}$$

If we introduce[14]

$$g_k=(-1)^k\frac{\pi}{\sin\pi\zeta}\frac{1}{\Gamma(\zeta+k)}\hat{g}_k, \tag{6.3}$$

the above recurrence relation becomes

$$(B-I_{2d})\hat{g}_k=-[A+(k+\zeta-1)I_{2d}]\hat{g}_{k-1}, \quad \hat{g}_{-1}\equiv 0. \tag{6.4}$$

However, this has the same form as the one given by (4.2) with $\beta=1$, and so we can proceed in the same way pursued in Sec. 4 and obtain the following results:

$$\eta_{Rj}=Q_B\left\{\sum_{k=0}^{\infty}\frac{(-1)^k}{\sin\pi\zeta}\frac{\pi}{\Gamma(\zeta+k)}(s-1)^{k+\zeta}\left(\prod_{l=0}^{k-1}A_1(k-l,\zeta)\right)\right\}\begin{bmatrix}0\\ p_j\end{bmatrix},$$
$$\zeta=\frac{3N-5}{2}+\frac{\pi_j}{\mathcal{C}}, \quad j=1,\ldots,d. \tag{6.5}$$

Thus d number of solutions of (5.1) are found and the rest remains unknown. To find them, we consider (6.2) again with $k=0$. Thus we obtain

$$(\zeta-1)(B-I_{2d})g_0=0. \tag{6.6}$$

But this implies that ζ can be taken equal to one, and so we can write

$$k(B-I_{2d})g_k=(A+kI_{2d})g_{k-1}, \quad g_{-1}\equiv 0. \tag{6.7}$$

To solve this, let us make the following transformation:

$$g_k=Q_B\bar{g}_k, \quad \bar{g}_k^T=[\bar{g}_k^{(1)},\bar{g}_k^{(2)}], \tag{6.8}$$

where Q_B is the orthonormal matrix of B and it is defined by (4.6). Therefore, instead of (6.7) we have

$$k\bar{g}_k^{(1)}=-\left[\left(k+\frac{1-3N}{2}\right)I_d+\frac{P_n}{\mathcal{C}}\right]\bar{g}_{k-1}^{(1)}-\bar{g}_{k-1}^{(2)},$$
$$0=-\left[L_n+\left(\frac{P_n}{\mathcal{C}}+\frac{1-3N}{2}I_d\right)\left(\frac{P_n}{\mathcal{C}}+\frac{3N-3}{2}I_d\right)\right]\bar{g}_{k-1}^{(1)}$$
$$+\left[\left(k+\frac{3-3N}{2}\right)I_d-\frac{P_n}{\mathcal{C}}\right]\bar{g}_{k-2}^{(2)}. \tag{6.9}$$

Upon rearranging we find

$$(kI_{2d}-\hat{G}_1)\bar{g}_k=-(k\hat{G}_2+\hat{G}_3)\bar{g}_{k-1}, \quad \bar{g}_{-1}\equiv 0, \tag{6.10}$$

or, returning to g_k's,

$$(kI_{2d}-G_1)g_k=(kG_2+G_3)g_{k-1}, \quad g_{-1}\equiv 0. \tag{6.11}$$

Here the matrices G_1, G_2, G_3 are defined in the following manner:

$$G_j = Q_B \hat{G}_j Q_B^{-1}, \quad j=1,2,3, \tag{6.12}$$

$$\hat{G}_1 = \begin{bmatrix} 0 & 0 \\ \left\{L_n + \left(\frac{P_n}{\mathcal{E}} + \frac{1-3N}{2} I_d\right)\left(\frac{P_n}{\mathcal{E}} + \frac{3N-3}{2} I_d\right)\right\} & \left(\frac{P_n}{\mathcal{E}} + \frac{3N-5}{2} I_d\right) \end{bmatrix}, \tag{6.13}$$

$$\hat{G}_2 = \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}, \quad \hat{G}_3 = \begin{bmatrix} \frac{1-3N}{2} I_d + \frac{P_n}{\mathcal{E}} & I_d \\ 0 & 0 \end{bmatrix}.$$

As we can see from its structure, d number of eigenvalues of G_1 are zero and the others may be equal to some positive integers or zero, depending on the value of $\mathcal{E}$. For some values of $\mathcal{E}$ which make $(kI_{2d} - G_1)$ nonsingular for all k's, g_k can be expressed in terms of g_0 by iterating (6.11); therefore, as a solution of (5.1) we arrive at the following result:

$$g = \sum_{k=0}^{\infty} (1-s)^k \left(\prod_{l=0}^{k-1} [(k-l)I_{2d} - G_1]^{-1}[(k-l)G_2 + G_3]\right) g_0,$$
$$G_1 g_0 = 0. \tag{6.14}$$

The last equation, obtained from (6.11) by taking $k=0$, indicates that the number of solutions defined above is equal to the difference between the dimension of G_1 and its rank. By investigating its structure in some details, we conclude that the rank of G_1 is d; hence there are d number of linearly independent solutions in the form given by (6.14). They are given explicitly below:

$$\rho_{Rj} = \sum_{k=0}^{\infty} (1-s)^{-1} \left(\prod_{l=0}^{k-1} [(k-l)I_{2d} - G_1]^{-1}[(k-l)G_2 + G_3]\right) Q_B \bar{g}_0^{(j)},$$

$$\bar{g}_0^{(j)} = \begin{bmatrix} e_j \\ \left(\frac{5-3N}{2} I_d - \frac{P_n}{\mathcal{E}}\right)^{-1} \left\{L_n + \left(\frac{P_n}{\mathcal{E}} + \frac{1-3N}{2} I_d\right) \times \left(\frac{P_n}{\mathcal{E}} + \frac{3N-3}{2} I_d\right)\right\} e_j \end{bmatrix}, \quad j=1,\ldots,d \tag{6.15}$$

For some values of $\mathcal{E}$, such that the matrix $(kI_{2d} - G_1)^{-1}$ has simple poles at some points in the domain of k, some constraints must be imposed on g_0. But also in this case, without loss of generality, d number of analytic functions can be obtained with some modifications on coefficients. Consequently, we can say that half of the solutions of (5.1) are analytic and the rest are not.

The whole discussion so far remains also true for left solutions of (5.1). Hence we can sketch the following results by excluding intermediate steps:

$$\eta_{Lj} = (B^T - sI_{2d})(Q_B^{-1})^T \left\{\sum_{k=0}^{\infty} (s-1)^{k+\zeta} \frac{\Gamma(\zeta+1)}{\Gamma(\zeta+k+1)} \times \left[\prod_{l=0}^{k-1} \hat{A}_1(k-l,\zeta)\right]\right\} \begin{bmatrix} 0 \\ p_j \end{bmatrix},$$

$$\zeta = -\frac{3N-3}{2} - \frac{\pi_j}{\mathcal{E}}, \quad j=1,\ldots,d,$$

$$\hat{A}_1^{11} = \left(\frac{3-3N}{2} - k - \zeta\right) I_d + \frac{P_n}{\mathcal{E}},$$
$$\hat{A}_1^{12} = -\left[L_n + \left(\frac{P_n}{\mathcal{E}} + \frac{1-3N}{2} I_d\right)\left(\frac{P_n}{\mathcal{E}} + \frac{3N-3}{2} I_d\right)\right],$$
$$\hat{A}_1^{21} = \left[\left(k+\zeta+\frac{3N-3}{2}\right) I_d + \frac{P_n}{\mathcal{E}}\right]^{-1},$$
$$\hat{A}_1^{22} = \left[\left(k+\zeta+\frac{3N-3}{2}\right) I_d + \frac{P_n}{\mathcal{E}}\right]^{-1} \hat{A}_1^{12}, \tag{6.16}$$

$$\rho_{Lj} = (B^T - sI_{2d})(Q_B^{-1})^T \sum_{k=0}^{\infty} (1-s)^k \left(\prod_{l=0}^{k-1} [(k-l)I_{2d} + \tilde{G}_1]^{-1} \times [(k-l)\hat{G}_2 + \tilde{G}_3]\right) \begin{bmatrix} \left(\frac{3N-3}{2} I_d + \frac{P_n}{\mathcal{E}}\right) e_j \\ e_j \end{bmatrix},$$
$$j=1,\ldots,d,$$

$$\tilde{G}_1 = \begin{bmatrix} 0 & 0 \\ -I_d & \frac{P_n}{\mathcal{E}} + \frac{3N-3}{2} I_d \end{bmatrix},$$

$$\tilde{G}_3 = \begin{bmatrix} \frac{3N-3}{2} I_d - \frac{P_n}{\mathcal{E}} & L_n + \left(\frac{P_n}{\mathcal{E}} + \frac{1-3N}{2} I_d\right)\left(\frac{P_n}{\mathcal{E}} + \frac{3N-3}{2} I_d\right) \\ 0 & 0 \end{bmatrix}. \tag{6.17}$$

Now, the Ω's being some constants, we can state the following properties:

$$\eta_{Lj}^T \eta_{Rk} = \Omega_{jk}^{(1)}, \quad \eta_{Lj}^T \rho_{Rk} = \Omega_{jk}^{(2)},$$
$$\rho_{Lj}^T \eta_{Rk} = \Omega_{jk}^{(3)}, \quad \rho_{Lj}^T \rho_{Rk} = \Omega_{jk}^{(4)}, \tag{6.18}$$
$$\text{for all } s \in R_1, \quad j,k=1,\ldots,d.$$

The solutions of (5.1) at the point $s=0$ are not included since they are superfluous for the determination of the discrete spectrum.

7. DETERMINATION OF THE DISCRETE SPECTRUM

We know that the conditions imposed on the probability function are such that it is finite, continuous, and square integrable over the whole space. But this implies that the vector-valued function F is finite, continuous, and square integrable with the weight r^{3N-1} on the domain of r. A careful investigation of (3.16), (3.17), and (3.18) shows that the solutions of (3.14), which remains finite or vanishes at $r=0$, are expressed as (3.16).

Now, if we take Laplace transform of each side of (3.16), we obtain

$$\int_0^{\infty} \exp(-sr) F_{jk}^{(*)} dr = \left\{\frac{\partial^{n-j}}{\partial \sigma^{n-j}} \left[(\sigma - j)^{n-j} \int_0^{\infty} \exp(-sr)\Phi \, dr\right]\right\}_{\sigma=j} (a_R^*)_{jk},$$
$$j=0,\ldots,n, \quad k=1,\ldots,\nu_j. \tag{7.1}$$

By comparing the explicit expression of Φ with φ_R, given by (5.5), we obtain

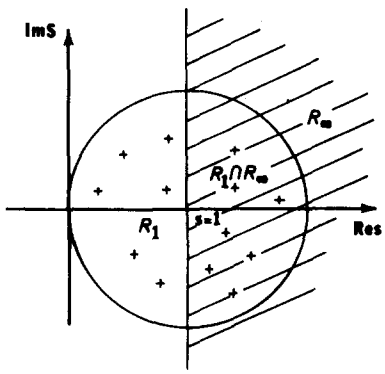


FIG. 2.

$$\int_0^\infty \exp(-sr)F_{jk}^{(+)}\,dr=\left\{\frac{\partial^{n-j}}{\partial\sigma^{n-j}}\left[(\sigma-j)^{n-j}\varphi_R\right]\right\}_{\sigma=j}(a_R^+)_{jk}=g_{jk},$$
$$j=0,\ldots,n,\quad k=1,\ldots,\nu_j. \tag{7.2}$$

We know from the theory of linear differential equations that we can express any solution of (3.14), which remains finite or vanishes when r goes to zero, as follows:

$$F(r)=\sum_{j=0}^{n}\sum_{k=1}^{\nu_j} C_{jk}F_{jk}^{(+)}, \tag{7.3}$$

where the C_{jk}'s are constant. The Laplace transform of $F(r)$ can be written in terms of g_{jk}'s as

$$g(s)=\int_0^\infty \exp(-sr)F(r)\,dr=\sum_{j=0}^{n}\sum_{k=1}^{\nu_j} C_{jk}g_{jk}. \tag{7.4}$$

The theory of Laplace transformation dictates that any function of s, expanded at $s=\infty$ in terms of negative powers of s, is analytic in the half-plane to the right of a vertical line which passes through the point of singularity with largest real part. However, the singularity of a function at any point $s=s_0$ means that its inverse transform behaves as $\exp(s_0 r)$ when r increases unboundedly. If we recall (3.13), we can conclude that the function $F(r)$ defined above must increase more slowly than $\exp(r/2)$ for increasing values of r, in order to vanish at infinity. This implies that its Laplace transform $g(s)$ must be analytic for all s which have a greater real part than $\frac{1}{2}$. As we deduce from previous discussion in this paper, $g(s)$ is analytic for all values of s whose real parts are greater than either 1 or 0, depending on the values of C_{jk}'s. Hence, to have a vanishing $F(r)$ when r increases unboundedly, the coefficients C_{jk}'s must be chosen such that the function is to be analytic for all s with positive real parts. To accomplish this, we seek for the relation between each of the solutions of (5.1) expanded at infinity and those at the point of singularity $s=1$.

We assume the following relations between the two sets of solutions:

$$g_{jk}=\sum_{l=1}^{d}(\tau_{jkl}^{(1)}\eta_{Rl}+\tau_{jkl}^{(2)}\rho_{Rl}), \tag{7.5}$$

$$h_{jk}=\sum_{l=1}^{d}(\tau_{jkl}^{(3)}\eta_{Rl}+\tau_{jkl}^{(4)}\rho_{Rl}), \tag{7.6}$$

where the τ's stand for some constants. If we take s equal to s_0 in the domain defined by $R_1\cap R_\infty$ (see Fig. 2), all functions at each sides of above equation remain convergent; hence their values can be evaluated by using their explicit forms. Then we premultiply each side of above equation by resorting to (6.18) to obtain

$$\eta_{Li}^T g_{jk}=\sum_{m=1}^{\infty}(\tau_{jkm}^{(1)}\Omega_{lm}^{(1)}+\tau_{jkm}^{(2)}\Omega_{lm}^{(2)})=\Lambda_{jkl}^{(1)}, \tag{7.7}$$

$$\eta_{Li}^T h_{jk}=\sum_{m=1}^{d}(\tau_{jkm}^{(3)}\Omega_{lm}^{(1)}+\tau_{jkm}^{(4)}\Omega_{lm}^{(2)})=\Lambda_{jkl}^{(2)}, \tag{7.8}$$

$$\rho_{Li}^T g_{jk}=\sum_{m=1}^{d}(\tau_{jkm}^{(1)}\Omega_{lm}^{(3)}+\tau_{jkm}^{(2)}\Omega_{lm}^{(4)})=\Lambda_{jkl}^{(3)}, \tag{7.9}$$

$$\rho_{Li}^T h_{jk}=\sum_{m=1}^{d}(\tau_{jkm}^{(3)}\Omega_{lm}^{(3)}+\tau_{jkm}^{(4)}\Omega_{lm}^{(4)})=\Lambda_{jkl}^{(4)}, \tag{7.10}$$

$$j=1,\ldots,n,\quad k=1,\ldots,\nu_j,\quad l=1,\ldots,d,$$

where the Λ's denote some constants independent from s_0.

By employing (7.5), $g(s)$ can be expressed in terms of the solutions of (5.1) at $s=1$ as follows:

$$g(s)=\sum_{j=0}^{n}\sum_{k=1}^{\nu_j}\sum_{l=1}^{d}(C_{jk}\tau_{jkl}^{(1)}\eta_{Rl}+C_{jk}\tau_{jkl}^{(2)}\rho_{Rl}). \tag{7.11}$$

As previously discussed, the functions η_{Rl} are not analytic at $s=1$. On the contrary, they possess algebraic or logarithmic branch cuts. However, the functions ρ_{Rl} are analytic in R_1. Hence to render $g(s)$ analytical properties, the coefficients of the functions η_{Rl} in (7.11) must be taken equal to zero. Therefore,

$$\sum_{j=0}^{n}\sum_{k=1}^{\nu_j}\tau_{jkl}^{(1)}C_{jk}=0,\quad l=1,\ldots,d. \tag{7.12}$$

The above equation can be put into the matrix form

$$\tau^{(n)}(\mathcal{E})C_n=0, \tag{7.13}$$

where the elements of $\tau^{(n)}$ and C_n are the terms $\tau_{jkl}^{(1)}$'s and C_{jk}'s respectively.

Nontrivial solutions of (7.13) occur only in the case where determinant of coefficient matrix vanishes. Thus we obtain

$$\det\tau^{(n)}(\mathcal{E})=\Delta_n(\mathcal{E})=0. \tag{7.14}$$

We call the last two equations, (7.13), (7.14), as truncated energy equations of the previously mentioned system. The determinant Δ_n can be called the energy determinant and its roots (in $\mathcal{E}$ domain) determine the discrete energy spectrum of system, which corresponds to the bounded states.

In the limiting case $n\to\infty$ the exact solution of hyperradial equation and therefore the exact solution of Schrödinger's equation of the system are obtained. Hence the equations

$$\{\tau^{(n)}(\mathcal{E})C_n\}_{n\to\infty}=0,\quad \Delta_\infty(\mathcal{E})=0 \tag{7.15}$$

give the negative energy values of the system and the corresponding probability functions.

8. EVALUATION OF THE ENERGY MATRIX, HYDROGENIC SPECTRUM

To evaluate the elements of energy matrix we can utilize Eqs. (7.7)–(7.10). If we use j, k for vertical

ordering and l for horizontal one in these equations, we obtain the following matrix equations:

$$\tau^{(1)}\Omega^{(1)}+\tau^{(2)}\Omega^{(2)}=\Lambda^{(1)},\quad \tau^{(3)}\Omega^{(1)}+\tau^{(4)}\Omega^{(2)}=\Lambda^{(2)},$$
$$\tau^{(1)}\Omega^{(3)}+\tau^{(2)}\Omega^{(4)}=\Lambda^{(3)},\quad \tau^{(3)}\Omega^{(3)}+\tau^{(4)}\Omega^{(4)}=\Lambda^{(4)}. \tag{8.1}$$

From these we can extract $\tau^{(1)}$ in terms of Ω's and Λ's:

$$\tau^{(1)}=[\Lambda^{(1)}-\Lambda^{(3)}(\Omega^{(4)})^{-1}\Omega^{(2)}][\Omega^{(1)}-\Omega^{(3)}(\Omega^{(4)})^{-1}\Omega^{(2)}]^{-1}. \tag{8.2}$$

A careful examination of energy matrix shows that

$$\tau^{(n)}(\mathcal{E})=(\tau^{(1)})^{T}. \tag{8.3}$$

Therefore, we find the expression of energy matrix in terms of Λ's and Ω's.

The evaluation of Ω's and Λ's is a difficult problem in the practical sense because of their complex structures. However, we can find an extreme case in which they have a simpler structure. The commutativity of L with P, in the case of hydrogen atom shows the origin of this simplification. If we replace P by a diagonal matrix D_p whose elements are obtained from the eigenvalues of P by ordering them in increasing values and by noting the simplifications on the structure of products appeared at the previous formulas, we arrive at the following results after some tedious manipulations:

$$g_{jk}=j!\begin{bmatrix}\dfrac{u_1}{s^{j+1}}e_l\\ -\dfrac{d}{ds}\left(\dfrac{u_1}{s^j}e_l\right)\end{bmatrix},\quad u_1(j,k)={}_2F_1\left(\begin{matrix}j+1, & j+\alpha\\ 2j+3N-1\end{matrix}\middle|\,\frac{1}{s}\right),$$

$$\xi_{Rjk}=\Gamma(-q)Q_B\begin{bmatrix}(s-1)^q u_2 e_l\\ s^\alpha\dfrac{d}{ds}\{(s-1)^q s^{1-\alpha}u_2\}e_l\end{bmatrix},$$

$$u_2={}_2F_1\left(\begin{matrix}1-\alpha-q & q+j+1\\ q+1\end{matrix}\middle|\,1-s\right),$$

$$\rho_{Rjk}=Q_B\begin{bmatrix}u_3 e_l\\ -\{su_3'+(1-\alpha)u_3\}e_l\end{bmatrix},$$

$$u_3={}_2F_1\left(\begin{matrix}j+1, & -j-3N+3\\ 1-q\end{matrix}\middle|\,1-s\right),$$

$$l=\nu_0,\ldots,\nu_j+k,\quad k=1,\ldots,\nu_j,$$

$$\alpha=\frac{3N-1}{2}-\frac{\pi_l}{\mathcal{E}},\quad q=\frac{\pi_l}{\mathcal{E}}+\frac{3N-3}{2}, \tag{8.4}$$

where ${}_2F_1$ is the hypergeometric function. From the theory of hypergeometric functions,[15] we can write

$$\frac{1}{s^{j+1}}u_1=\frac{\Gamma(q)(2j+3N-2)!}{\Gamma(q+j+1)(j+3N-3)!}u_3+\frac{\Gamma(-q)(2j+3N-2)!}{\Gamma(j+\alpha)j!}(s-1)^\alpha u_2. \tag{8.5}$$

However, this relation implies that

$$g_{jk}=\frac{(2j+3N-2)!}{\Gamma(j+\alpha)}\eta_{Rjk}+\frac{(2j+3N-2)!}{(j+3N-3)!}\frac{\Gamma(q)}{\Gamma(k+q+1)}\rho_{Rjk}. \tag{8.6}$$

Therefore, by comparing with (7.5) we can obtain the energy matrix and energy functional equation as follows:

$$(\tau^{(n)})_{j_1k_1j_2k_2}=\frac{(2j+3N-2)!}{\Gamma(j+(3N-1)/2-\pi_l/\mathcal{E})}\delta_{j_1j_2}\delta_{k_1k_2},$$
$$l=\nu_0+\cdots+\nu_{j-1}+k,\quad j=0,\ldots,n,\quad k=1,\ldots,\nu_j, \tag{8.7}$$

$$\Delta_n=\prod_{j=0}^{n}\prod_{k=1}^{\nu_j}\frac{(2j+3N-2)!}{\Gamma(j+(3N-1)/2-\pi_l/\mathcal{E})},$$
$$l=\nu_0+\cdots+\nu_{j-1}+k,\quad k=1,\ldots,\nu_j. \tag{8.8}$$

This yields the following series of roots:

$$\mathcal{E}_{jkl}=\frac{\pi_{\nu_0+\cdots+\nu_{j-1}+k}}{[l+j+(3N-3)/2]},$$

$$E_{jkl}=-\frac{\pi^2_{\nu_0+\cdots+\nu_{j-1}+k}}{[l+j+(3N-3)/2]^2},$$
$$j=0,\ldots,n,\quad k=1,\ldots,\nu_j,\quad l=1,2,3,\ldots,\infty. \tag{8.9}$$

This forms of energy values are similar to those in the case of hydrogen atom; therefore, we can call them the "hydrogenic spectrum." However, in the last equation, the constants π's are not yet been evaluated. Their evaluation necessitates the solution of the eigenproblem of the matrix P. For finite-dimensional investigation these can be obtained in the familiar manner.

Of course, the truncation of original hyperradial equation gives an approximate spectrum, and some series of energy are excluded. To obtain all of the series of energy, the truncation of the original equation must be carried out by retaining only the terms corresponding to the hyperharmonic interactions with an order of homogeneity between n_1 and n_2 ($n_2>n_1$), instead of the one between 0 and n. By changing n_1 and n_2 in a convenient manner and observing the alterations in the energy values, we can obtain the approximate spectrum in the desired error limits.

9. CONCLUSIONS

In this paper we have proposed a general formulation for a system of electrically charged particles. By employing hyperspherical coordinates we have made an analogy to the case of hydrogen atom and obtained a similar spectrum.

Of course, we have not carried out a rigorous mathematical treatment, but we have elaborated a fairly effective method in the sense of numerical computation. We hope to offer the several applications of our approach to various systems in the forthcoming papers.

By taking hydrogenic spectrum as a first approximation, several approximation methods of variational or perturbational origin can be developed, and we are presently working in this direction.

Perhaps the most striking feature of our method is the obtaining of the whole energy spectrum in a single step. The modification of hydrogenic spectrum leads to the

well-known Rayleigh series in atomic and molecular spectra. This will be the subject of a future paper.

Another approach to the problem is via the energy matrix mentioned above. Its explicit form in terms of $\mathcal{E}$ may be useful for obtaining the approximate values of the energy spectrum.

We can also add that this method does not lose its effectiveness in the existence of time-independent (or not) external perturbations. However, some modifications are necessary when time dependence appears, due to the great difference in the nature of problem.

Also it seems that it is possible to extend this method to cover relativistic cases.

We hope to report some of our results in forthcoming papers.

ACKNOWLEDGMENT

The authors are indebted to Professor M. T. Erben for several valuable discussions.

[1]M. Demiralp, Doctoral dissertation, Technical University of Istanbul (in Turkish) (1974).
[2]M. Demiralp and E. Şuhubi., "Quantum Mechanics of Electrically Charged Particles," Applied Mathematics Division, Marmara Scientific and Industrial Research Institute, Istanbul, Report No. 35 (1976).
[3]D. L. Knirk, J. Chem. Phys. 60, 66 (1974).
[4]R. C. Whitten and J. S. Sims, Phys. Rev. A 9, 1586 (1974).
[5]H. Mayer, J. Phys. A 8 (10), 1562 (1975).
[6]R. C. Whitten, J. Math. Phys. 10, 1631 (1969).
[7]G. O. Morrel and D. L. Knirk, Theoret. Chim. Acta 37, 345 (1975).
[8]K. Okubo, J. Math. Soc. Japan 15 (3), 268 (1963).
[9]A. Messiah, *Quantum Mechanics* (North-Holland, Amsterdam, 1970), Vol. I, p. 361.
[10]A. Erdélyi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Chap. XI.
[11]A term $\Sigma_{j=m}^{n} \alpha_j$ appearing in any equation will be taken equal to 1 if m is greater than n.
[12]W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics* (Springer-Verlag, Berlin, 1966), Chap. V.
[13]G. Doetsch, *Introduction to the Theory and Application of the Laplace Transformation* (Springer-Verlag, Berlin, 1970).
[14]The term $\pi/\sin\pi\zeta$ is so chosen as to make these solutions as the Laplace transforms of asymptotic solutions of (3.14).
[15]Y. L. Luke, *The Special Functions and Their Approximations* (Academic, New York, 1969), Vol. 1.

On the semiclassical expansion in quantum mechanics for arbitrary Hamiltonians

Maurice M. Mizrahi

Center for Naval Analyses of The University of Rochester, 1401 Wilson Boulevard, Arlington, Virginia 22209
(Received 2 April 1976)

It is shown that if the quantum-mechanical propagator, satisfying the n-dimensional Schrödinger equation $(\mathbf{H}-i\hbar\partial/\partial t_b)\langle q_b,t_b|q_a,t_a\rangle=0$, with $\langle q_b,t_b|q_a,t_b\rangle=\delta(q_b-q_a)$ and arbitrary classical Hamiltonian H_c, admits a semiclassical (WKB) approximation, then the latter is of the form $K_{\rm WKB}=\text{const}\times\hbar^{-n/2}(\det M)^{1/2}\exp(iS_c/\hbar)$, where S_c is the classical action, $M_{ij}\equiv-\partial^2S_c/\partial q_a^i\partial q_b^j$, and $K_{\rm WKB}^{-1}(\mathbf{H}-i\hbar\partial/\partial t_b)K_{\rm WKB}=O(\hbar^2)$. The restrictions on the correspondence rule chosen to pass from H_c to the operator $\mathbf{H}$ are spelled out in detail, and it is found that one has a considerable amount of leeway in choosing such a rule. Differential equations for higher-order corrections can be generated at will. This generalizes previously known partial results to arbitrary Hamiltonians. The arbitrariness of the Hamiltonian makes the method useful as a general tool in the theory of partial differential equations.

INTRODUCTION

The purpose of this paper is to extend to a very wide class of Hamiltonians the range of validity of a well-known and often used expression for the semiclassical (WKB) approximation to the quantum-mechanical propagator $K(q_b,t_b;q_a,t_a)\equiv\langle q_b,t_b|q_a,t_a\rangle$, or probability amplitude that a particle at q_a at time t_a will arrive at q_b at time t_b. This formula is, in n dimensions,

$$K_{\rm WKB}(q_b,t_b;q_a,t_a)=\text{const}\times\hbar^{-n/2}[\det M(q_b,t_b;q_a,t_a)]^{1/2}\times\exp[(i/\hbar)S_c(q_b,t_b;q_a,t_a)], \tag{1}$$

where the constant is determined by the boundary condition

$$K(q_b,t_b;q_at_b)=\delta(q_b-q_a), \tag{2}$$

S_c is the classical action,

$$\int_{t_a}^{t_b}L(q_c,\dot q_c,t)\,dt=\int_{t_a}^{t_b}[p_c\cdot\dot q_c-H_c(p_c,q_c,t)]\,dt,$$

where p_c and q_c are the classical momentum and position, satisfying Hamilton's equations and $q_c(t_a)=q_a$, $q_c(t_b)=q_b$, and $M_{ij}\equiv-\partial^2S_c/\partial q_a{}^i\partial q_b{}^j$ is the Van Vleck–Morette matrix (or "Hessian of the action"). This asymptotic approximation hinges on the smallness of $\hbar$, as the error is of order $\hbar^2$. $K_{\rm WKB}$ is an approximation to the solution of the Schrödinger equation

$$\left(\mathbf{H}-i\hbar\frac{\partial}{\partial t_b}\right)K\equiv\left[H\left(q_b,-i\hbar\frac{\partial}{\partial q_b},t_b\right)-i\hbar\frac{\partial}{\partial t_b}\right]K=0 \tag{3}$$

such that

$$K_{\rm WKB}^{-1}\left(\mathbf{H}-i\hbar\frac{\partial}{\partial t_b}\right)K_{\rm WKB}=O(\hbar^2). \tag{4}$$

Boundary condition (2) is also satisfied by $K_{\rm WKB}$.

Equation (1) was introduced by Van Vleck in 1928 in his famous paper on the correspondence principle.[1] In the case where t_b-t_a is infinitesimally small ("short-time" propagator), it was related to the "normalization factor" of Lagrangian path integrals by Morette[2] in 1951, and later (1957) extended to n-dimensional curved spaces by B.S. DeWitt.[3] In the case of the classical Hamiltonian

$$H_c=(2m)^{-1}g^{ij}(q)\left[p_i-\frac{e}{c}A_i(q)\right]\left[p_j-\frac{e}{c}A_j(q)\right]+V(q),$$

one must choose the corresponding quantum operator to be ($g\equiv-\det g_{ij}$)

$$\frac{1}{2m}g^{-1/4}(\mathbf{Q})\left[\mathbf{P}_i-\frac{e}{c}A_i(\mathbf{Q})\right]g^{1/2}(\mathbf{Q})g^{ij}(\mathbf{Q})\times\left[\mathbf{P}_j-\frac{e}{c}A_j(\mathbf{Q})\right]g^{-1/4}(\mathbf{Q})+V(\mathbf{Q}) \tag{5}$$

[covariant when operating on a scalar density of weight $\frac{1}{2}$, which is what the ket $|q,t\rangle$ is if $|\psi(q,t)|^2\,dq\equiv|\langle\psi|q,t\rangle|^2\,dq$ is to be an invariant], and add a term proportional to $\hbar^2R(\mathbf{Q})$, R being the Ricci scalar.[3] In all these cases the constant is $(2\pi i)^{-n/2}$.

A simple proof by induction shows that (1) is valid for $\mathbf{H}=\mathbf{P}^n$, but not for $\mathbf{H}=f(\mathbf{Q})\mathbf{P}^n$.

All this makes one wonder: For what class of Hamiltonian operators is the approximation to the propagator given by (1) valid to first order in $\hbar$ [i.e., is such that (4) is satisfied]? This is the question that we propose to answer here.

Notation: The Einstein summation convention over repeated indices is used. All integrals are over $\mathbb{R}^p$ for suitable p unless otherwise specified. We operate in n dimensions, $dp=dp_1\cdots dp_n$, $p\cdot q=p_1q^1+\cdots+p_nq^n$. The p's and q's are canonical coordinates, ranging from $-\infty$ to ∞. Boldface capital letters denote operators.

Correspondence rules: We will proceed from the most general Hamiltonian operator $\mathbf{H}\equiv H(\mathbf{P},\mathbf{Q},t)$ one can extract from a classical Hamiltonian $H_c(p,q,t)$, in n dimensions. This operator is of the form given by Cohen[4]:

$$\mathbf{H}=(2\pi\hbar)^{-2n}\int dp\,dq\,du\,dv\,F(u,v,\hbar)H_c(p,q,t)\times\exp\{(i/\hbar)[(q-\mathbf{Q})\cdot u+(p-\mathbf{P})\cdot v]\}, \tag{6}$$

where $F(u,v,\hbar)$ is the transformation function. For example,

$F=1$ gives Weyl ordering,[5]

$$q^np^m\to\frac{1}{2^n}\sum_{l=0}^{n}\mathbf{Q}^{n-l}\mathbf{P}^m\mathbf{Q}^l\binom{n}{l}, \tag{7a}$$

$F=\cos(u\cdot v/2\hbar)$ gives symmetrized ordering,

$$q^np^m\to\tfrac{1}{2}[\mathbf{P}^m\mathbf{Q}^n+\mathbf{Q}^n\mathbf{P}^m], \tag{7b}$$

Copyright © 1977 American Institute of Physics

$F=(u\cdot v/2\hbar)^{-1}\sin(u\cdot v/2\hbar)$ gives Born–Jordan ordering,[6]

$$q^n p^m \to \frac{1}{m+1}\sum_{l=0}^{m}\mathbf{P}^{m-l}\mathbf{Q}^n\mathbf{P}^l. \tag{7c}$$

More compactly, the normal-ordered H is[4]:

$$\mathbf{H}=\left[F\left(i\hbar\frac{\partial}{\partial p},\, i\hbar\frac{\partial}{\partial q},\hbar\right)\exp\left(-\frac{i\hbar}{2}\frac{\partial}{\partial p}\cdot\frac{\partial}{\partial q}\right)H_c(p,q,t)\right]_{\substack{q\ \text{before}\ p\\ p=\mathbf{P},\,q=\mathbf{Q}}}, \tag{8}$$

i.e., after evaluating the function inside the bracket, order the terms such that the q factors precede the p factors (normal-ordering), then replace p and q by the operators P and Q, respectively. If $-i\hbar/2$ in the exponent is replaced by $i\hbar/2$, then the p factors must precede the q factors.

Equation (6) is a kind of function-to-operator inverse Fourier transform. Unfortunately, it is not covariant. The direct transform is given by[7]:

$$H_c(p,q,t)=(2\pi\hbar)^{-n}\int dz\,dy\,du\,\exp\{(i/\hbar)[z\cdot p+y\cdot(q-u)]\}\times F^{-1}(z,y,\hbar)\langle u-\tfrac{1}{2}z\,|\,\mathbf{H}\,|\,u+\tfrac{1}{2}z\rangle, \tag{9}$$

where the bra and the ket are position states.

If H is to be Hermitian with H_c real, this forces the condition $F^*(-u,-v,\hbar)=F(u,v,\hbar)$. If $f(\mathbf{P})$ and $g(\mathbf{Q})$ are to correspond to $f(p)$ and $g(q)$, then $F(0,v,\hbar)=F(u,0,\hbar)=1$. Finally,[7] if the operator corresponding to p^mq^n is to be a linear combination of permutations of products of P m times and Q n times, then F is a function of the one variable $u\cdot v/\hbar$. In fact, $u\cdot v/\hbar$ is the only dimensionless combination one can form with u, v, and $\hbar$.

We will assume that these conditions are satisfied, and that no totally arbitrary operator proportional to $\hbar$ is added to the Hamiltonian operator so constructed—an ever-present possibility when traveling the road from classical to quantum mechanics.

1. SEMICLASSICAL SOLUTION OF THE SCHRÖDINGER EQUATION

In configuration space, at the point q_b, $\mathbf{Q}^i$ is represented by q_b^i and $\mathbf{P}_i$ by $-i\hbar\,\partial/\partial q_b^i$. Substituting this in (6), and using the commutation relation $[\mathbf{Q}^i,\mathbf{P}_j]=i\hbar\delta_j^i\,\mathbf{I}$, together with the formula

$$e^{\mathbf{A}+\mathbf{B}}=e^{\mathbf{A}}e^{\mathbf{B}}e^{-[\mathbf{A},\mathbf{B}]/2}, \tag{10}$$

valid for any operators A and B which commute with their commutator, we have the normal-ordered H,

$$\mathbf{H}=(2\pi\hbar)^{-2n}\int dp\,dq\,du\,dv\,F(u,v,\hbar)H_c(p,q,t)\times\exp\left[\frac{i}{\hbar}u\cdot\left(q-q_b+\frac{v}{2}\right)\right]\exp\left(\frac{i}{\hbar}p\cdot v\right)\exp\left(-v\cdot\frac{\partial}{\partial q_b}\right). \tag{11}$$

Let us now expand the transformation function F about $u=0$, assuming it is analytic at that point,

$$F(u,v,\hbar)=\sum_{l=0}{}'\frac{1}{l!}\,\varphi_{i_1\cdots i_l}(v,\hbar)u^{i_1}\cdots u^{i_l}. \tag{12}$$

Here $\sum'$ is the sum over all $i_j=\{1,\ldots,n\}$ $(i_0=0)$ and over all l from 0 to ∞, and

$$\varphi_{i_1\cdots i_l}(v,\hbar)\equiv\frac{\partial^l F(u,v,\hbar)}{\partial u^{i_1}\cdots\partial u^{i_l}}\bigg|_{u=0}=\hbar^{-l}F^{(l)}(0)v^{i_1}\cdots v^{i_l},\quad\text{if } F=F(u\cdot v/\hbar). \tag{13}$$

Substituting (12) in (11), and using

$$(2\pi\hbar)^{-n}\int v^{i_1}\cdots v^{i_m}\exp(ip\cdot v/\hbar)\,dv=(-i\hbar)^m\frac{\partial^m\delta(p)}{\partial p_{i_1}\cdots\partial p_{i_m}} \tag{14}$$

we have

$$\mathbf{H}=(2\pi\hbar)^{-n}\int dp\,dq\,dv\,H_c(p,q,t_b)\exp\left(\frac{ip\cdot v}{\hbar}\right)\times\sum_{l=0}^{\infty}{}'\frac{(-i\hbar)^l}{l!}\varphi_{i_1\cdots i_l}(v,\hbar)\frac{\partial^l\delta(x)}{\partial x^{i_1}\cdots\partial x^{i_l}}\bigg|_{x=q-q_b+v/2}\times\exp\left(-v\cdot\frac{\partial}{\partial q_b}\right)$$
$$=(2\pi\hbar)^{-n}\int dp\,dv\,\exp\left(\frac{ip\cdot v}{\hbar}\right)\sum_{l=0}^{\infty}{}'F^{(l)}(0)\frac{i^l}{l!}\times v^{i_1}\cdots v^{i_l}\eta_{i_1\cdots i_l}\left(p,q_b-\frac{v}{2},t_b\right)\exp\left(-v\cdot\frac{\partial}{\partial q_b}\right), \tag{15}$$

where

$$\eta_{i_1\cdots i_l}(p,q,t)\equiv\frac{\partial^l}{\partial q^{i_1}\cdots\partial q^{i_l}}H_c(p,q,t). \tag{16}$$

Since $\exp[k(d/dx)]f(x)=f(x+k)$, when H operates on $\psi(q_b,t_b)$ we have

$$\mathbf{H}\psi(q_b,t_b)=(2\pi\hbar)^{-n}\int dp\,dv\,\exp\left(\frac{ip\cdot v}{\hbar}\right)\psi(q_b-v,t_b)\times\sum_{l=0}^{\infty}{}'\frac{1}{l!}F^{(l)}(0)i^l v^{i_1}\cdots v^{i_l}\eta_{i_1\cdots i_l}\left(p,q_b-\frac{v}{2},t_b\right). \tag{17}$$

We look for an asymptotic solution of the Schrödinger equation (3) for small $\hbar$. Let us assume one such solution of the form

$$K(q_b,t_b;q_a,t_a)=a(\hbar)C(q_b,t_b;q_a,t_a)\exp[(i/\hbar)B(q_b,t_b;q_a,t_a)], \tag{18}$$

where $a(\hbar)$ is a constant, B and C are real. The next step is to expand $K^{-1}(\mathbf{H}-i\hbar\,\partial/\partial t_b)K$ in a power series in $\hbar$. For this, we expand some terms inside the integrand of (17) in a power series in v. To simplify the writing, the arguments of B, C, and K will not be written; primes and dots will denote differentiation with respect to q_b and t_b, respectively. Thus, $B''_{ij}\equiv\partial^2B(q_b,t_b;q_a,t_a)/\partial q_b^i\,\partial q_b^j$. We have

$$K^{-1}\left(\mathbf{H}-i\hbar\frac{\partial}{\partial t_b}\right)K=(2\pi\hbar)^{-n}C^{-1}\exp\left(-\frac{iB}{\hbar}\right)\left\{\int dp\,dv\,\exp\left[\frac{i}{\hbar}v\cdot(p-B')\right]\times C\exp\left(\frac{iB}{\hbar}\right)\left[1+\frac{i}{2\hbar}v^iv^jB''_{ij}-\frac{1}{8\hbar^2}v^iv^jv^kv^sB''_{ij}B''_{ks}\right.\right.$$

$$- \frac{i}{6\hbar} v^i v^j v^k B'''_{ijk} + \cdots \Big] \Big[1 - v^s \frac{C'_s}{C} + \frac{v^i v^j}{2C} C''_{ij} + \cdots \Big]$$
$$\times \sum_{l=0}^{\infty}{}' \frac{i^l}{l!} F^{(l)}(0) v^{i_1} \cdots v^{i_l} \Big[\eta_{i_1 \cdots i_l}(p, q_b, t_b) - \frac{v^k}{2}$$
$$\times \eta'_{i_1 \cdots i_l k}(p, q_b, t_b) + \tfrac{1}{8} v^i v^j \eta''_{i_1 \cdots i_l ij}(p, q_b, t_b) + \cdots \Big] \Big\}$$
$$+ \dot{B} - i\hbar \frac{\dot{C}}{C} . \tag{19}$$

We can now do the v integral. From (14), we know that if the integrand contains the kth power of v, the corresponding term will be of order $\hbar^k$. Thus, all the terms not explicitly written in the expansions of (18) are of order $\hbar^{l+m}$ where $m \geq 3$. Since we want to exhibit explicitly terms of no higher order in $\hbar$ than 2, we need not be concerned with them.

It is straightforward to carry out a typical integral in (18),

$$(2\pi\hbar)^{-n} \int dp\, dv \exp\Big[\frac{i}{\hbar} v \cdot (p - B')\Big] \eta^{(m)}_{i_1 \cdots i_l j_1 \cdots j_m}(p, q_b, t_b)$$
$$\times v^{i_1} \cdots v^{i_l} v^{j_1} \cdots v^{j_m} v^{k_1} \cdots v^{k_r}$$
$$= \Big(\frac{\hbar}{i}\Big)^{l+m+r} \int dp\, \eta^{(m)}_{i_1 \cdots i_l j_1 \cdots j_m}(p, q_b, t_b)$$
$$\times \frac{\partial^{l+m+r} \delta(x)}{\partial x_{i_1} \cdots \partial x_{i_{l+m+r}}} \Big|_{x = p - B'}$$
$$= (i\hbar)^{l+m+r} \Big[\frac{\partial^{l+m+r}}{\partial p_{i_1} \cdots \partial p_{i_{l+m+r}}} \eta^{(m)}_{i_1 \cdots i_l j_1 \cdots j_m}(p, q_b, t_b) \Big]_{p = B'}, \tag{20}$$

where the subscripts $j_1 \cdots j_m$ on $\eta_{i_1 \cdots i_l}$ denote derivatives with respect to $q_b^{j_1}, \ldots, q_b^{j_m}$.

Thus, to order $l+2$ in $\hbar$, (19) gives

$$K^{-1}\Big(\mathrm{H} - i\hbar \frac{\partial}{\partial t_b}\Big) K$$
$$= \dot{B} - i\hbar \frac{\dot{C}}{C} + \sum_{l=0}^{\infty}{}' \frac{(-\hbar)^l}{l!} F^{(l)}(0) \Big\{ \frac{\partial^{2l}}{\partial p_{i_1} \cdots \partial p_{i_l} \partial q_b^{i_1} \cdots \partial q_b^{i_l}}$$
$$\times \Big[1 - \frac{i\hbar}{2} \frac{\partial^2}{\partial p_i \partial q_b^i} - i\hbar \frac{C'_i}{C} \frac{\partial}{\partial p_i} - \frac{\hbar^2}{8} \frac{\partial^4}{\partial p_i \partial p_j \partial q_b^i \partial q_b^j}$$
$$- \frac{\hbar^2 C'_i}{2C} \frac{\partial^3}{\partial p_i \partial p_k \partial q_b^k} - \frac{\hbar^2 C''_{ij}}{2C} \frac{\partial^2}{\partial p_i \partial p_j} - \frac{i\hbar}{2} B''_{ij} \frac{\partial^2}{\partial p_i \partial p_j}$$
$$- \frac{\hbar^2}{4} B''_{ij} \frac{\partial^4}{\partial p_i \partial p_j \partial p_k \partial q_b^k} - \frac{\hbar^2}{2C} C'_k B''_{ij} \frac{\partial^3}{\partial p_i \partial p_j \partial p_k}$$
$$- \frac{\hbar^2}{8} B''_{ij} B''_{ks} \frac{\partial^4}{\partial p_i \partial p_j \partial p_k \partial p_s} - \frac{\hbar^2}{6} B'''_{ijk} \frac{\partial^3}{\partial p_i \partial p_j \partial p_k} \Big]$$
$$\times H_c(p, q_b, t_b) \Big\}_{p = B'}. \tag{21}$$

To order $\hbar^2$, only parts of the first three terms of the l series survive, and we have

$$K^{-1}\Big(\mathrm{H} - i\hbar \frac{\partial}{\partial t_b}\Big) K$$
$$= \dot{B} + F(0) H_c(B', q_b, t_b) + \hbar \Big[-i \frac{\dot{C}}{C}$$
$$- \Big(\frac{\partial^2}{\partial p_i \partial q_b^i} H_c(p, q_b, t_b) \Big)_{p = B'} \Big(F'(0) + \frac{i}{2} F(0) \Big)$$
$$- \frac{iF(0)C'_i}{C} \Big(\frac{\partial H_c}{\partial p_i}(p, q_b, t_b) \Big)_{p = B'} - i \frac{F(0)}{2} B''_{ij}$$
$$\times \Big(\frac{\partial^2 H_c}{\partial p_i \partial p_j}(p, q_b, t_b) \Big)_{p = B'} \Big] + \hbar^2 \Big\{ \Big[-\frac{1}{8} F(0) \frac{\partial^2}{\partial q_b^i \partial q_b^j}$$
$$- F(0) \frac{C'_i}{2C} \frac{\partial}{\partial q_b^j} - F(0) \frac{C''_{ij}}{2C} - \frac{1}{4} F(0) B''_{ij} \frac{\partial^2}{\partial p_k \partial q_b^k}$$
$$- F(0) \frac{C'_k}{2C} B''_{ij} \frac{\partial}{\partial p_k} - \frac{1}{8} F(0) B''_{ij} B''_{ks} \frac{\partial^2}{\partial p_k \partial p_s}$$
$$- \frac{1}{6} F(0) B'''_{ijk} \frac{\partial}{\partial p_k} + \frac{i}{2} F'(0) \frac{\partial^2}{\partial q_b^i \partial q_b^j} + iF'(0) \frac{C'_i}{C} \frac{\partial}{\partial q_b^j}$$
$$+ \frac{i}{2} F'(0) B''_{ij} \frac{\partial^2}{\partial p_k \partial q_b^k} + \frac{1}{2} F''(0) \frac{\partial^2}{\partial q_b^i \partial q_b^j} \Big]$$
$$\times \frac{\partial^2}{\partial p_i \partial p_j} H_c(p, q_b, t_b) \Big\}_{p = B'} + O(\hbar^3). \tag{22}$$

The semiclassical approximation consists in dropping terms of order $\hbar^2$ in (22). Thus, assuming F is real, we equate the real and imaginary parts of both the constant term and the term proportional to $\hbar$ to zero, and obtain the differential equations which must be simultaneously satisfied by B and C for K in (18) to be K_{WKB}:

$$F(0) H_c\Big(\frac{\partial B}{\partial q_b}, q_b, t_b\Big) + \frac{\partial B}{\partial t_b} = 0, \tag{23}$$

$$F(0) \frac{\partial}{\partial q_b^j} \Big[C^2 \frac{\partial H_c}{\partial p_j}(p, q_b, t_b) \Big|_{p = \partial B / \partial q_b} \Big] + \frac{\partial}{\partial t_b}(C^2) = 0, \tag{24}$$

$$F'(0) \Big[\frac{\partial^2}{\partial p_i \partial q_b^i} H_c(p, q_b, t_b) \Big]_{p = \partial B / \partial q_b} = 0, \tag{25}$$

where the second equation has been written in a different form.

If $F(0) = 1$, (23) is recognized as being the Hamilton–Jacobi equation, yielding $B = S_c$, the classical action.

By virtue of Hamilton's equations, one has, at the classical path $q_c(t)$,

$$\frac{\partial H_c}{\partial p_i}(p, q_b, t_b) \Big|_{p = \partial S_c / \partial q_b} = \dot{q}_c^i(t_b), \tag{26}$$

the classical velocity at the point (q_b, t_b). The continuity equation satisfied by the Van Vleck–Morette determinant $\det M \equiv \det(-\partial^2 S_c / \partial q_a^i \partial q_b^j)$ is precisely[8]

$$\frac{\partial}{\partial q_b^j} [\dot{q}_c^j(t_b) \det M] + \frac{\partial}{\partial t_b} (\det M) = 0. \tag{27}$$

Therefore, (24) implies that $C = (\det M)^{1/2}$.

Equation (25) is never satisfied for arbitrary classical Hamiltonians H_c unless $F'(0) = 0$.

Conclusion: The WKB approximation (1) to the propagator is valid for all Hamiltonian operators obtained from an arbitrary classical Hamiltonian by the very general formula (6), where the transformation function $F(u, v, \hbar)$ is a real function of $u \cdot v / \hbar$ alone, such that

$$F(0) = 1, \quad F'(0) = 0. \tag{28}$$

Note that the three ordering examples (7a)–(7c) do satisfy these conditions.

The higher-order terms in (22) contain higher-order derivatives of H_c. Thus, for a Hamiltonian quadratic

in the momenta, such as,

$$H_c = (2m)^{-1} g^{ij}(q)\left[p_i - \frac{e}{c}A_i(q)\right]\left[p_j - \frac{e}{c}A_j(q)\right] + V(q), \tag{29}$$

the "miss" factor is simply

$$K_{\mathrm{WKB}}^{-1}\left(\mathbf{H} - i\hbar\frac{\partial}{\partial t_b}\right)K_{\mathrm{WKB}}$$
$$= -\frac{\hbar^2}{2m}\left(\left[\tfrac{1}{4} - F''(0)\right]g^{ij}_{,ij}(q_b)\right.$$
$$+ (\det M)^{-1/2} g^{ij}_{,j}(q_b)\frac{\partial}{\partial q_b^i}(\det M)^{1/2}$$
$$\left. + g^{ij}(q_b)(\det M)^{-1/2}\frac{\partial^2}{\partial q_b^i \partial q_b^j}(\det M)^{1/2}\right), \tag{30}$$

whether the corresponding quantum operator H is given by (5) (which seems to correspond to Weyl's rule,[9] $F = 1$) or by another ordering of the factors corresponding to a different F satisfying the required conditions. If H_c is quadratic in both the p's *and* the q's, then S_c is a quadratic function of q_a and q_b, $\det M$ is independent of q_a and q_b, and the WKB approximation is exact, a well-known result.

2. HIGHER-ORDER CORRECTIONS

Higher-order corrections to the semiclassical approximation can be easily generated by this method. For example, suppose we want the exponent in (1) to be of the form $iU/\hbar$, where U is a series in $i\hbar$ in which the classical action is, of course, the zero-order term. The first-order term $i\hbar U_1(q_b, t_b; q_a, t_a)$ would be lumped into $(\det M)^{1/2}$ and need not be considered. The second-order term, yielding a trial propagator of the form

$$\bar{K} = a(\hbar)(\det M)^{1/2}\exp[(i/\hbar)(S_c + \hbar^2 D)] \tag{31}$$

can be readily analyzed. Substituted in (19), the extra term contributes a factor

$$\exp[i\hbar D(q_b - v, t_b)]$$
$$= \exp[i\hbar D(q_b, t_b)]\left(1 - i\hbar v^i\frac{\partial D}{\partial q_b^i} + \cdots\right) \tag{32}$$

in the integrand, with $\hbar^2 \dot{D}$ added at the end. The terms not shown above contribute only to order $\hbar^{1+3}$. The net result is to add two extra terms to the coefficient of $\hbar^2$ in (22). Setting this coefficient equal to zero yields a differential equation for the extra function $D(q_b, t_b; q_a, t_a)$,

$$\frac{\partial D}{\partial t_b} + \varphi^i\frac{\partial D}{\partial q_b^i} + \psi = 0, \tag{33}$$

where φ^i and ψ are the known functions:

$$\varphi^i \equiv \left[\frac{\partial H_c(p, q_b, t_b)}{\partial p_i}\right]_{p=B'},$$
$$\psi \equiv -\left\{\left[\frac{C_i'}{2C}\frac{\partial}{\partial q_b^i} + \frac{1}{2C}C_{ij}''\right.\right.$$
$$+ \frac{1}{4}B_{ij}''\frac{\partial^2}{\partial p_k \partial q_b^k} + \frac{C_k' B_{ij}''}{2C}\frac{\partial}{\partial p_k} + \frac{1}{8}B_{ij}''B_{ks}''\frac{\partial^2}{\partial p_k \partial p_s}$$
$$\left.\left. + \frac{1}{6}B_{ijk}'''\frac{\partial}{\partial p_k} + \frac{1}{2}\left(\frac{1}{4} - F''(0)\right)\frac{\partial^2}{\partial q_b^i \partial q_b^i}\right]\frac{\partial^2 H_c(p, q_b, t_b)}{\partial p_i \partial p_j}\right\}_{p=B'},$$

and $B = S_c$, $C = (\det M)^{1/2}$, $B_i' \equiv \partial S_c/\partial q_b^i$, etc., $F(0) = 1$, $F'(0) = 0$. The new propagator is then exact to order $\hbar^2$, i.e.,

$$\bar{K}^{-1}\left(\mathbf{H} - i\hbar\frac{\partial}{\partial t_b}\right)\bar{K} = O(\hbar^3). \tag{34}$$

Similarly, differential equations for higher-order corrections can be generated at will. It is remarkable that all these terms are built entirely from classical quantities[10] once the (quantum) transformation function is fixed.

3. THE BOUNDARY CONDITION

We have found B and C such that (18) approximately satisfies the Schrödinger equation at (q_b, t_b), and hence is an approximate wavefunction at (q_b, t_b). In order for K to be the true propagator, there must also exist a constant $a(\hbar)$ such that when $t_b \to t_a$, $K_{\mathrm{WKB}} \to \delta(q_b - q_a)$, i.e., the following equation must be satisfied:

$$I(\hbar; q_a, q_b, t_b) \equiv \lim_{T\to 0^+}\int \varphi(q_a)(\det M)^{1/2}\exp[(i/\hbar)S_c]\,dq_a$$
$$= a^{-1}(\hbar)\varphi(q_b), \tag{35}$$

where φ is any C^∞ function and $a(\hbar)$ is a constant.

Assume that there exists a Lagrangian $L(q, \dot{q}, t) \equiv p\cdot\dot{q} - H_c(p, q, t)$, i.e., p can be obtained explicitly in terms of q and $\dot{q}$ by solving Hamilton's equation $\dot{q} = \partial H/\partial p$, and that, as $t_b - t_a \equiv T$ approaches 0,

$$S_c = \int_{t_a}^{t_b} L(q, \dot{q}, t)\,dt \to L\left(q_b, \frac{q_b - q_a}{T}, t_b\right)T, \tag{36}$$

$$M_{ij} \equiv \frac{-\partial^2 S_c}{\partial q_a^i \partial q_b^j}$$
$$\to \left(\frac{\partial^2 L}{\partial \dot{q}^i \partial q^j} + \frac{1}{T}\frac{\partial^2 L}{\partial \dot{q}^i \partial \dot{q}^j}\right)\left(q_b, \frac{q_b - q_a}{T}, t_b\right). \tag{37}$$

If we substitute (36) and (37) in (35), and make the change of variable $q_a = q_b - uT$, we get

$$I = \lim_{T\to 0^+} T^n \int du\,\varphi(q_b - uT)\exp\left\{\frac{i}{\hbar}L(q_b, u, t_b)T\right\}$$
$$\times \det{}^{1/2}\left[\frac{\partial^2 L}{\partial u^i \partial q_b^j}(q_b, u, t_b) + \frac{1}{T}\frac{\partial^2 T}{\partial u^i \partial u^j}(q_b, u, t_b)\right]. \tag{38}$$

In the limit, the second term in the argument of the determinant dominates the first, and so the latter can be dropped. Further, $\varphi(q_b - uT) \to \varphi(q_b)$, and we can pull the latter outside the integral. What remains to be shown is that

$$a^{-1}(\hbar) = \lim_{T\to 0^+} T^{n/2}\int\left(\det\frac{\partial^2 L}{\partial u^i \partial u^j}(q_b, u, t_b)\right)^{1/2}$$
$$\times \exp\left[\frac{iT}{\hbar}L(q_b, u, t_b)\right]du \tag{39}$$

is indeed a constant independent of q_b and t_b. If we let $T = \hbar\epsilon$, we readily obtain the $\hbar$ dependence, $a(\hbar) = \text{const} \times \hbar^{-n/2}$. This, incidentally, is the only $\hbar$ dependence which would give the propagator K the dimensions $(\text{length})^{-n}$, which the boundary condition (or the completeness relation for the $|q, t\rangle$ states) implies that it must have. Thus, we should have, when the system

allows a WKB approximation,

$$\lim_{\epsilon\to 0^+}\epsilon^{n/2}\int\left[\det\frac{\partial^2 L(q_b,u,t_b)}{\partial u^i\partial u^j}\right]^{1/2}\exp[i\epsilon L(q_b,u,t_b)]\,du = \text{const} \tag{40}$$

for all Lagrangians with sufficient properties of continuity and differentiability.

We do not know of any general proof of (40), valid for all Lagrangian functions, nor do we know of a counterexample. Some general properties of this integral can be worked out, and (40) can be directly verified for many large classes of Lagrangians:

(1) In the case of the Lagrangian

$$L=\tfrac{1}{2}mg_{ij}(q,t)\dot q^i\dot q^j+(e/c)A_i(q,t)\dot q^i-V(q,t), \tag{41}$$

corresponding to the Hamiltonian (29), use of the well-known integral,

$$\int du\exp\left(\frac{i}{\hbar}b\cdot u+\frac{i}{2\hbar}S_{ij}u^iu^j\right) = \frac{(2\pi i\hbar)^{n/2}}{(\det S)^{1/2}}\exp\left(\frac{i}{2\hbar}(S^{-1})^{ij}b_ib_j\right), \tag{42}$$

readily verifies (39), for $a(\hbar)=(2\pi i\hbar)^{-n/2}$.

(2) Consider the one-dimensional case where the Lagrangian $L(q,u,t)$ is asymptotically (small u) of the form $g(q,t)+f(q,t)u^l$ for $l>0$ and $(-1)^l\neq -i$. In the same approximation, the Hamiltonian is

$$H(p,q,t)=[(l-1)/l]p^{l/(l-1)}[f(q,t)l]^{-1/(l-1)}-g(q,t).$$

The change of variable $u^l=x^2$ readily yields[11]

$$a^{-1}(\hbar)=2\left[\frac{\pi i\hbar(l-1)}{l}\right]^{1/2}. \tag{43}$$

This can be generalized to n dimensions, for Lagrangians which are asymptotically (small u) of the form

$$L(q,u,t)=f_1(q,t)u_1^{l_1}+\cdots+f_n(q,t)u_n^{l_n} \tag{44}$$

for $l_i>0$ and $(-1)^{l_i}\neq -i$. The integral in (39) breaks up into the product of n identical one-dimensional integrals, and the result is

$$a^{-1}(\hbar)=2^n(\pi i\hbar)^{n/2}\left[\frac{(l_1-1)\cdots(l_n-1)}{l_1\cdots l_n}\right]^{1/2}. \tag{45}$$

(3) In one dimension, one can show that if the limit in (39) exists for two Lagrangians L_1 and L_2, it exists for their sum L_1+L_2.

Proof: Standard inequalities are used to show that the limit for L_1+L_2 is bounded from above by a combination of the limits for L_1 and L_2. Setting $\partial^2L(q_b,u,t_b)/\partial u^2\equiv L''$, we have

$$\begin{aligned}&\left|\sqrt{\epsilon}\int du(L_1''+L_2'')^{1/2}\cos\epsilon(L_1+L_2)\right|\\ &\le\sqrt{\epsilon}\int\left|(L_1''+L_2'')^{1/2}\right|\cdot\left|\cos\epsilon(L_1+L_2)\right|du\\ &\le\sqrt{\epsilon}\int\left|L_1''^{1/2}\right|\cdot\left|\cos\epsilon(L_1+L_2)\right|du\\ &\quad+\sqrt{\epsilon}\int\left|L_2''^{1/2}\right|\cdot\left|\cos\epsilon(L_1+L_2)\right|du\\ &\le\sqrt{\epsilon}\int\left|L_1''^{1/2}\right|\cdot\left|\cos\epsilon L_1\right|du+\sqrt{\epsilon}\int\left|L_1''^{1/2}\right|\cdot\left|\sin\epsilon L_1\right|du\\ &\quad+\sqrt{\epsilon}\int\left|L_2''^{1/2}\right|\cdot\left|\cos\epsilon L_2\right|du\\ &\quad+\sqrt{\epsilon}\int\left|L_2''^{1/2}\right|\cdot\left|\sin\epsilon L_2\right|du=\text{const as }\epsilon\to 0.\end{aligned} \tag{46}$$

and similarly for the imaginary part with $\sin\epsilon(L_1+L_2)$. By virtue of (2) above, this implies that the theorem is true for all Lagrangians in one dimension which can be expanded in powers of the velocity $\dot q$.

[1] J.H. Van Vleck, Proc. Nat. Acad. Sci. (USA) **14**, 178 (1928).

[2] C. Morette, Phys. Rev. **81**, 848 (1951). See also W. Pauli, Zürich lecture notes, 1950—1951, published in C.P. Enz, Ed., *Pauli's Lectures on Physics* (M.I.T. Press, Cambridge, Mass., 1973), Vol. 6 ("Selected Topics in Field Quantization"), and Ph. Choquard, Helv. Phys. Acta **28**, 89—157 (1955).

[3] B.S. DeWitt, Rev. Mod. Phys. **29**, 377 (1957).

[4] L. Cohen, J. Math. Phys. **7**, 781 (1966). Our notation here is slightly different, in that we have introduced $\hbar$ explicitly in F and in the Fourier transforms, and changed i into $-i$ in the exponent.

[5] H. Weyl, Z. Phys. **46**, 1 (1927); *Theory of Groups and Quantum Mechanics* (Dover, New York, 1950), p. 275; see also S.R. DeGroot and L.G. Suttorp, *Foundations of Electrodynamics* (North-Holland, Amsterdam, 1972), Chap. VI and Appendix.

[6] M. Born and P. Jordan, Z. Phys. **34**, 873 (1925).

[7] I.W. Mayes, Ph.D. thesis, The University of Manchester, 1971. The result is expressed therein in the form of a trace which, when evaluated in configuration space, readily yields the stated formula.

[8] Note that $\partial^2S_c/\partial\alpha^i\partial q_b^j$, where α is any other constant of integration, satisfies the continuity equation just as well. Here we have chosen $\alpha=q_a$ because of the context.

[9] It can be shown [see M. Mizrahi, J. Math. Phys. **16**, 2201 (1975)] that the inverse Weyl transform [i.e., (6) with $F=1$] of (29) is the covariant operator (5) plus the two terms $\hbar^2R(Q)/8m$ and $\hbar^2\Gamma^m_{l_i}(Q)\Gamma^{l_i}_{mj}(Q)g^{ij}(Q)/8m$. This is essentially the "quantum-mechanical potential" of B.S. DeWitt in Phys. Rev. **85**, 653 (1952). The second additional term is not covariant, reflecting the fact that the whole correspondence scheme as presented here is not manifestly covariant. By requiring that his Schrödinger equation be covariant, B.S. DeWitt in Ref. 3 does not get this second term, and his formula (1) further has $(\det M)^{1/2}$ replaced by $g^{-1/4}(q_a)$ $(\det M)^{1/2}g^{-1/4}(q_b)$ for proper weight. See also I.W. Mayes and J.S. Dowker, J. Math. Phys. **14**, 434 (1973) for more discussion of this quadratic Hamiltonian.

[10] For some common Lagrangians the higher-order terms can be expressed and evaluated directly as path integrals. See C. DeWitt-Morett, Ann. Phys. (N.Y.) **97**, 367—399 (1976) and M. Mizrahi, "WKB Expansions by Path Integrals, With Applications to the Anharmonic Oscillator," preprint.

[11] By the method of stationary phase, which can be made applicable here by expressing $\sqrt{\epsilon}$ as $(\pi i)^{-1/2}\int_{-\infty}^{\infty}\exp(iy^2/\epsilon)dy$ and doing the u integral first, one can show [*Handbook of Applied Mathematics*, edited by Carl E. Pearson (Van Nostrand, Princeton, N.J., 1974), p. 656] that if $p(t)\sim p(0)+Pt^m$ and $q(t)\sim Qt^{n-1}$ as $t\to 0^+$, where P, m, and n are real and positive constants and Q can be complex, and other conditions on p and q expressed therein, then, as $\epsilon\to 0^+$,

$$\int_0^\infty\exp\left(\frac{i}{\epsilon}p(t)\right)q(t)\,dt\sim\exp\left(\frac{n\pi i}{2m}\right)\frac{Q}{m}\Gamma\left(\frac{n}{m}\right)\exp\left(\frac{i}{\epsilon}p(0)\right)\left(\frac{\epsilon}{P}\right)^{n/m}.$$

This result is then used to yield (43).

Second-quantized representation for a model system with composite particles*

J. Dennis Gilbert

Institute of Theoretical Science, University of Oregon, Eugene, Oregon 97403
(Received 24 March 1976)

A second-quantized representation for nonrelativistic systems with composite particles is given in which composite particles of two-fermion constituents are described as elementary bosons. The Tani transformation of Girardeau is used to introduce elementary composite fields. An equivalent linear equation of motion method—introducing a "d matrix"—is used in place of the usual nonlinear equation of motion method. This and a diagram formalism are developed to calculate a Hamiltonian and subsidiary condition for this model system and to derive some properties of the Tani transformation.

1. INTRODUCTION

Systems of atoms, nuclei, ions, molecules, and localized quasiparticles like Cooper pairs and Wannier excitons can be meaningfully regarded as systems of composite particles. Some systems, plasmas for example, can be meaningfully regarded as containing both "free" constituents and those "bound" as composite particles. This is justified at the theoretical as well as observational level. The existence of composite particle fields must be included in the theory to get the correct qualitative results. The binding of constituents is not a perturbation. Historically, the theory of such systems has included composite particle fields, but in an essentially phenomenological way. An important theoretical task, then, is the unification of the theory of the constituents with the theory of systems with composite particles. Considerable progress appears to have been made recently with two basic approaches, though no nontrivial calculations of observables have been completed yet.

The first approach is to regard composite particles as elementary with their field operators satisfying elementary bose or fermi commutation relations. The task, then, separates into two parts. One is that of the many-body problem. The other is that of determining the interactions among the composite particles and "free" constituent particles, and possibly, subsidiary conditions. Transformation of representation methods are used and, because of the separation, work is focused on the second part of the task. Work in developing this approach includes that by Girardeau,[1-3] Brittin and Stolt,[4] Sakakura,[5] Brittin and Sakakura,[6] Nishigori,[7] and Vorob'ev, Khomkin, and Gleizer.[8]

The second approach is to regard composite particles as being other than elementary. This approach, which must take care of the nonelementary properties in the formulation of the statistical mechanics, uses Green's functions methods. Work in developing this approach includes that by Sahlin and Schwartz[9] and Goldberg and Puff.[10] A full understanding of the relationship between both approaches has apparently not been developed.

The work reported here follows the first approach. Separate from the difficulties of the usual many-body problem, two theoretical obstacles must be overcome. The first is a problem of representation. If both constituent and composite particles are to be identified, there are problems of exchange symmetry and overcompleteness. There is the problem of handling the symmetry or antisymmetry coming from the identity of constituent particles whether "bound" or "free." Also, there is the overcompleteness problem of making a representation that does not count "bound" constituent particles once as composite particles and again as "free" constituent particles. The second obstacle is a problem of specifying the interactions among "free" constituents and composite particles. These ought to be derivable from what is usually known, the interaction among constituents. It is important to notice that even the "free" constituent particle interaction is expected to be different from the known interaction between constituents. The interaction between a "free" proton and "free" electron, for example, will reflect the process of the formation of hydrogen atoms.

The solution presented here is based on work begun by Girardeau and given in more detail by Gilbert.[11] It uses a redundant mode approach and an equation of motion technique for calculating the solution. The solution developed will be for a general system consisting of boson composite particles made of two fermions, and any other noncomposite particles being fermions. A more general system is not treated because it would obscure the approach which is still valid for more general systems. A more specific system was not chosen because this is the simplest general system containing the hydrogen plasma—although idealized to neglect H^-, H_2^+, H_2, H_2^-, etc.—as a specific case. This general system is actually simpler in form than the hydrogen plasma system. Other systems which might be well represented as special cases of this general model are mixed hydrogen and tritium plasma neglecting nuclear interactions, mixed hydrogen and antihydrogen plasma, Cooper pairs, and certain nuclear reactions.

In general, the interactions among composite particles are not known, but the interactions of constituent particles are known or have a more fundamental meaning. Even though it may not be the most useful description, a system can be well defined in terms of constituent particles only. It is on this solid ground that the solution begins. The usual Hamiltonian expressed in terms of constituent particle field operators is

$$H=\sum_{\mu}\int dx\,\psi^{\mu\dagger}(x)\,T^{\mu}(x)\,\psi^{\mu}(x)+\sum_{\mu\nu}\int dx\,dy\,\psi^{\mu\dagger}(x)\,\psi^{\nu\dagger}(y)\,V^{\mu\nu}(xy)\,\psi^{\nu}(y)\,\psi^{\mu}(x).$$

In the notation assumed throughout, the constituent

 Copyright © 1977 American Institute of Physics

particles of kind μ are associated with fermion field operators $\psi^{\mu}_{(x)}$. Field operators for fermions are taken to anticommute, otherwise field operators commute (i.e., fermion—boson and boson—boson). The coordinate variable is meant to include both spin and position coordinates.

In this description of the system, composite particles are usually represented by fields A^{σ}_{α} defined by

$$A^{\sigma\dagger}_{\alpha} = \int dx\,dy\, F^{\sigma\mu\nu}_{\alpha}(xy)\,\psi^{\nu\dagger}(x)\,\psi^{\mu\dagger}(y),$$

$$A^{\sigma}_{\alpha} = (A^{\sigma\dagger}_{\alpha})^{\dagger} = \int dx\,dy F^{\sigma\mu\nu *}_{\alpha}(xy)\,\psi^{\mu}(y)\,\psi^{\nu}(x).$$

The coordinates or momentum and internal state of the composite particle is given by α, and the kind of particle is given by σ. $F^{\sigma\mu\nu}_{\alpha}(xy)$ is the boundstate wavefunction of the σ composite particle in state α with constituent particle kinds μ and ν, with coordinates y and x, respectively. A standard ordering of μ, ν is assumed; for only one μ, ν ordering, $\mu \neq \nu$, will $F^{\sigma\mu\nu}_{\alpha}(xy) \neq 0$. For hydrogen, for example,

$$F_{(rnlm\sigma_1\sigma_2)}(xy) = \varphi_{(rnlm\sigma_1\sigma_2)}(xy)$$
$$= \delta(\mathbf{r} - \mathbf{r}_{cm})\,\varphi_{nlm}(\mathbf{x} - \mathbf{y})\,\delta_{\sigma_1\sigma_p}\,\delta_{\sigma_2\sigma_e}.$$

Such composite field operators are exact for a single composite particle system, but in general, because of the identity and symmetry among constituent particles, these operators do not satisfy all the bose commutation relations;

$$A^{\dagger}_{\alpha} = \int dx\,dy\,\varphi_{\alpha}(xy)\,\psi^{p\dagger}(x)\,\psi^{e\dagger}(y), \quad N_{\alpha} = A^{\dagger}_{\alpha}A_{\alpha},$$

$$[A^{\dagger}_{\alpha}, A^{\dagger}_{\beta}] = [A^{\dagger}_{\alpha}, \psi^{p\dagger}(z)] = [A^{\dagger}_{\alpha}, \psi^{e\dagger}(z)] = 0,$$

$$[\psi^{p}_{(x)}, A^{\dagger}_{\alpha}] = \int dy\,\varphi_{\alpha}(zy)\,\psi^{e\dagger}(y), [\psi^{e}_{(z)}, A^{\dagger}_{\alpha}] = -\int dx\,\varphi_{\alpha}(xz)\,\psi^{p\dagger}_{(x)}$$

$$[A_{\alpha}, A^{\dagger}_{\beta}] = \delta_{\alpha\beta} - \int dx\,dy\,dz\,\varphi^{*}_{\alpha}(xy)\,\varphi_{\beta}(xz)\,\psi^{e\dagger}(z)\,\psi^{e}(y)$$
$$- \int dx\,dy\,dz\,\varphi^{*}_{\alpha}(xy)\,\varphi_{\beta}(zy)\,\psi^{p\dagger}(z)\,\psi^{p}(x),$$

$$[N_{\alpha}, N_{\beta}] = \int dx\,dy\,dz\,du\,dv\,dr\,ds\,\{\varphi_{\beta}(uv)\,\varphi^{*}_{\beta}(xy)\,\varphi_{\alpha}(xz)\,\varphi^{*}_{\alpha}(sr)$$
$$- \varphi_{\alpha}(uv)\,\varphi^{*}_{\alpha}(xy)\,\varphi_{\beta}(xz)\,\varphi^{*}_{\beta}(sr)\}$$
$$\times \psi^{p\dagger}(u)\,\psi^{e\dagger}(v)\,\psi^{e\dagger}(z)\,\psi^{e}(y)\,\psi^{e}(r)\,\psi^{p}(s)$$
$$+ \int dx\,dy\,dz\,du\,dv\,dr\,ds\{\varphi_{\beta}(uv)\,\varphi^{*}_{\beta}(xy)\varphi_{\alpha}(xy)\,\varphi^{*}_{\alpha}(sr)$$
$$- \varphi_{\alpha}(uv)\,\varphi^{*}_{\alpha}(xy)\,\varphi_{\beta}(zy)\,\varphi^{*}_{\beta}(sr)\}$$
$$\times \psi^{p\dagger}(u)\,\psi^{p\dagger}(z)\,\psi^{e\dagger}(v)\,\psi^{e}(r)\,\psi^{p}(x)\,\psi^{p}(s).$$

Notice that even the number operators for composite particles of different states do not commute giving a fundamental ambiguity in defining states. This mathematical inadequacy is, in part, a statement of the problem of systems with composite particles. It is true that in some limit the commutation relations are well satisfied. When the composite particles are well enough separated from each other and from constituent particles, the indistinguishability of the constituent particles and resultant wavefunction symmetry has little effect. There is no exchange. The particles and the operators behave like elementary bosons and fermions.[12] When this is not the condition, the effect appears in the commutation relations. It is clearly in limiting conditions which are only approximated in real systems that the A's are a good representation of the composite particles.

2. METHOD OF SOLUTION

The redundant mode approach used here involves two steps. First, the system is considered as consisting of the original fermions, and then to this Hilbert space is added one of ideal composite particles. The Hilbert space originally used is enlarged with no difficulty to include the ideal composite particle field operators.[13] The part of the space of interest, the "physical part", is then where there are no ideal composite particles. The ideal composite particle field operators correspond identically in quantum numbers to the previously defined A^{σ}_{α} and are written a^{σ}_{α}. But the a's are different in that they obey all the bose commutation relations by definition, and thus represent elementary bosons of some sort.

Second, a unitary transformation is made. The object of the transformation is to put the composite particle properties of the A's into the a's. Of course, the A's can't be transformed exactly into the a's since a unitary transformation preserves commutation relations and the A's and a's have different commutation relations. The transformation will also give a subsidiary condition coming from the assumption that there were no ideal composite particles in the system before the transformation. The constituent particle fermion fields will also be transformed. Before the transformation such field operators would stand for both "bound" and "free" fermions. After the transformation they are expected to represent "free" fermions only.

Schematically, the transformation is expected to do the following:

$$A \to a + \left\{\begin{array}{l}\text{something small that goes to zero in limits}\\ \text{when the } A\text{'s satisfy bose commutation rela-}\\ \text{tions—when the composite particles are}\\ \text{sufficiently isolated from each other and}\\ \text{other constituents.}\end{array}\right\}$$

The meaning of the transformed field operators comes from three places: their algebraic properties, the transformed Hamiltonian that contains them, and any subsidiary condition that contains them. The commutation relations are defined to give what is desired for algebraic properties. After the new transformed Hamiltonian and the subsidiary condition are shown, it can be seen how the new field operators satisfy their expected meanings.

When the fermion fields are transformed, they could be substituted into the original Hamiltonian giving the transformed Hamiltonian. Or the Hamiltonian could be transformed directly. This Hamiltonian would show explicitly the processes of collisions, breakup, and combination that occur in a general system of composite and constituent particles. The subsidiary condition can be incorporated into the Hamiltonian by defining a projected Hamiltonian. This approach is discussed later. Here, however, the subsidiary condition will be viewed

separately as a constraint on the system.

The transformation used in the redundant mode approach is similar to one used by Tani and for that reason has been called a Tani transformation by Girardeau.[14] The Tani transformation has the form

$$U=\exp(\pi F/2),$$

$$F=\sum_{\alpha\mu} a_\alpha^{\mu\dagger}A_\alpha^\mu - A_\alpha^{\mu\dagger}a_\alpha^\mu \equiv a_\alpha^{\mu\dagger}A_\alpha^\mu - A_\alpha^{\mu\dagger}a_\alpha^\mu,$$

using an Einstein-like summation convention that will be used throughout. (A point of convention: Girardeau has chosen to view the Tani transformation in a formal sense as from the representation with "free" constituent particles and ideal composite particles to the representation where there are only constituent particles. Thus a transformed operator O is $U^{-1}OU$, rather than the usual UOU^{-1}. This convention will be followed.)

The Tani transformation has the desired properties

$$U^{-1}A_\alpha^{\sigma\dagger}|0\rangle = a_\alpha^{\sigma\dagger}|0\rangle, \qquad U^{-1}\psi^{\mu\dagger}(x)|0\rangle = \psi^{\mu\dagger}(x)|0\rangle$$

exactly, and

$$U^{-1}A_{\alpha_1}^{\sigma_1\dagger}\cdots A_{\alpha_n}^{\sigma_n\dagger}\psi^{\mu_1\dagger}(x_1)\cdots\psi^{\mu_m\dagger}(x_m)|0\rangle$$

$$= a_{\alpha_1}^{\sigma_1\dagger}\cdots a_{\alpha_n}^{\sigma_n\dagger}\psi^{\mu_1\dagger}(x_1)\cdots\psi^{\mu_m}(x_m)|0\rangle,$$

in the limit that the α_i's define in composite particles whose wavefunctions are nonzero only in disjoint regions, the x_i's are outside of these regions, and any two fields necessary for a boundstate are outside such regions. As would be expected, F commutes with the total number of constituents whether "bound" or "free" of any particular kind.

Evaluation of the Tani-transformed operators or Hamiltonian is a nontrivial task. Expansion of exponentials and summation of terms after normal ordering is difficult and has not been considered productive because of contributions from an infinite number of increasingly more complicated terms. For this reason, an equation of motion technique is used for carrying out the transformation. The technique comes from the result that

$$X(\epsilon)=\exp(-\epsilon F)\,X\exp(\epsilon F) \text{ implies } \frac{dX(\epsilon)}{d\epsilon}=[X(\epsilon),F].$$

The object is to solve such equations and set $\epsilon=\pi/2$, giving the answer. To accomplish this, a space of operators is defined so an operator $X(\epsilon)$ can be expressed as a sum of operators in this space with appropriate coefficients. The coefficients have all the ϵ dependence and the basis operators contain all the field operators. That such a space exists and can be generated has been shown and is indicated in Sec. 4.

If the set of basis operators is $\{x_i\}$, $X(\epsilon)=\sum c_i(\epsilon)x_i$. It is clear that $\dot{X}(\epsilon)$ is also in the space defined by $\{x_i\}$; and from the equation of motion, the commutator $[X(\epsilon),F]$ is also in the space defined by $\{x_i\}$, so the matrix d can be defined $[x_i,F]=\sum d_{ij}x_j$. Then $\dot{X}(\epsilon)=\sum\dot{c}_i(\epsilon)x_i=\sum c_i(\epsilon)d_{ij}x_j$ and since the x_i are chosen independent, $\dot{c}_i(\epsilon)=\sum c_j(\epsilon)d_{ji}=\sum\tilde{d}_{ij}c_j(\epsilon)$ which has the formal solution $C(\epsilon)=\exp(\tilde{d}\epsilon)C(0)$.

The d matrix is infinite, but it will be shown in Sec. 4 that the basis set $\{x_i\}$ can be chosen such that d is the sum of a block diagonal matrix plus an upper triangular matrix;

$$d=\begin{bmatrix} d_1 & d_{21} & d_{31} & \cdots \\ & d_2 & d_{32} & \cdots \\ 0 & & d_3 & \cdots \end{bmatrix}$$

where the d elements are matrices. (Later expressions have a better form if this reversed matrix elements subscript notation is used here.) Because of the form of d, it is possible to solve for the vector $C(\epsilon)$ in sections corresponding to the diagonal blocks, with later sections depending on earlier ones. To see this, define $C(\epsilon)$ in terms of these sections as

$$C(\epsilon)=\begin{bmatrix} C_I(\epsilon) \\ C_{II}(\epsilon) \\ \vdots \\ C_K(\epsilon) \\ \vdots \end{bmatrix},$$

where the $C_K(\epsilon)$'s are vector sections of $C(\epsilon)$. Then

$$\dot{C}_K=\sum\tilde{d}_{KJ}C_J$$

where the $\tilde{d}_{KJ}$ are the transpose of the matrices in d previously defined. But $\tilde{d}_{KJ}=0$ for $K<J$, so

$$\dot{C}_K=\sum_{J\le K}\tilde{d}_{KJ}C_J.$$

Thus

$$C_K(\epsilon)=\exp(\tilde{d}_K\epsilon)\left(\int_0^\epsilon dZ\exp(-\tilde{d}_KZ)\sum_{J<K}\tilde{d}_{KJ}C_J(Z)+C_K(0)\right).$$

It may also be convenient to solve the equations using the Laplace transformation. In this case, let

$$f_K(s)\equiv\mathcal{L}[C_K(\epsilon)].$$

Then

$$\mathcal{L}[\dot{C}_K(\epsilon)]=sf_K(s)-C_K(0),$$

so

$$sf_K(s)-C_K(0)=\tilde{d}_Kf_K(s)+\sum_{J<K}\tilde{d}_{KJ}f_J(s),$$

$$f_K(s)=(1s-\tilde{d}_K)^{-1}\left(\sum_{J<K}\tilde{d}_{KJ}f_J(s)+C_K(0)\right),$$

$$C_K(\epsilon)=\mathcal{L}^{-1}\left[(1s-\tilde{d}_K)^{-1}\left(\sum_{J<K}\tilde{d}_{KJ}f_J(s)+C_K(0)\right)\right].$$

Transformed operators are, in general, infinite sums. There is, in general, an infinite number of distinct kinds of terms each for the same field operator structure. These correspond to higher orders of possible particle exchanges. And there is, in general, an infinite set of field operator structures. These correspond to higher numbers of particles interacting. The basis set definition and ordering that gives the desired triangular behavior for the d matrix goes roughly from most important to least important in both senses. Thus the solution for a finite number of terms gives a good approximation upon which a many-body calculation can be made.

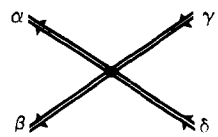
Basis elements can have the form of integrals and sums over creation operators times functions or c-number operators times annihilation operators. The functions and c-number operators represent the internal structure of the process(es) corresponding to the basis element (i.e., what interaction, what exchanges there are). The field operators represent the external structure (i.e., what is going in and out). There are many choices for basis sets, and some are clearly better than others for different purposes. An important thing to realize is that if a basis is chosen independent of the internal structure of the untransformed operator, the d matrix is a general expression of the relations among various operators and the processes they represent.

The d matrix method and the diagram formalism of Sec. 3 enable one to understand the form of the transformed operators and to rigorously make the necessary approximations. Practically, they enable one to solve the equations of motion for all orders that are expected to be important, though the calculation is quite difficult.

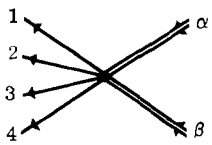
3. DIAGRAM FORMALISM

Besides evaluating the terms of a transformed operator, interpreting them is necessary. Both the complicated and cumbersome algebra are made more clear and manageable by introducing a geometrical representation. This has proven a useful aid in other areas of physics, and this seems true for composite particle systems too.

The transformed Hamiltonian, for example, will have terms for the two-particle interaction of composite particles. The field operator structure of such a term is $a^\dagger_\alpha a^\dagger_\beta a_\delta a_\gamma$. A diagram representation for such a term has been written as[15]



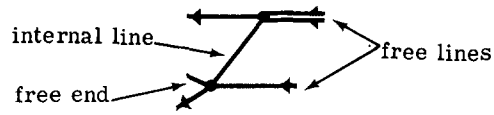
and a term of structure $\psi^\dagger_1 \psi^\dagger_2 \psi^\dagger_3 \psi^\dagger_4 a_\beta a_\alpha$ corresponding to break up as



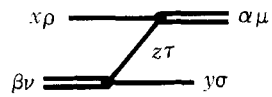
The interaction would be represented by the vertex point. Such a diagram system gives an overview, but not a view of the interaction which is often a complicated series, a central concern, and in need of interpretation.

The diagram formalism introduced here preserves the conventions of the external diagram representation while giving a representation for the internal structure as well. This formalism will be used to carry on the solution of the problem and interpret the results.

The diagrams consist of directed lines, from right to left, and vertex points. Lines coming in from infinity or going to infinity are called free lines. These are the lines of the external representation. Structure besides these lines occurs within the body of the diagram. Vertex points, for example, occur within the body of the diagram. Lines between vertex points are called internal lines. Lines just freely ending in the body of the diagram are called free ends. In the diagrams where there may be some confusion, free ends will not have arrows on them while free lines will; otherwise, arrows will be left off. Here is a diagram, for example,

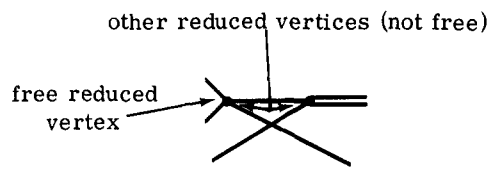


All lines and vertex points have subscripts corresponding to the kind of field (e.g., electron, proton, hydrogen, atom, etc.) and the coordinate, momentum, or state of the field the lines represent [e.g., $x=(\mathbf{r},\sigma)$, $\alpha=(rnlm\,\sigma_1\sigma_2)$, etc.]. Vertex points and composite particle lines are labeled with subscripts giving the kind of field and state. Constituent particle lines are labeled with subscripts giving the kind of field and coordinates. For example,



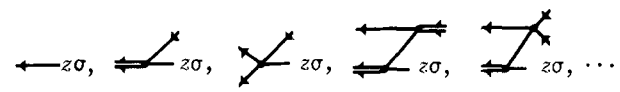
The composite subscripts go with both the vertex points and the composite particle lines shown as double lines.

A reduced vertex (not a vertex point) is defined as the convergence or divergence of lines to or from a vertex point. A reduced vertex with all free lines is called a free reduced vertex. For example,



In the interest of economical writing a reduced vertex may be called a vertex. The vertex point, however, will always be called a vertex point.

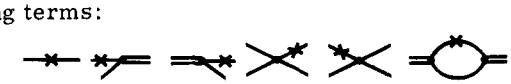
Operators of interest can be expanded in terms of diagrams. The free ends correspond to the net over-all field operator structure of the operator. For example, an expansion of the transform of the operator $\psi^\sigma(z)^\dagger$ will contain diagrams with one free end with the line going to the left and ending on the right. It would contain the following diagrams:



An expansion of the operator transformed from the kinetic energy operator,

$$\sum_\sigma \int dz\, \psi^\sigma(z)^\dagger T^\sigma(z)\, \psi^\sigma(z),$$

would have terms with no free ends since there is no net creation or annihilation. Using ⊣✕⊢ to denote the kinetic energy (or single particle diagonal interaction), the expansion of the kinetic energy contains the following terms:



The diagrams can be visualized as representing processes where certain fields are created and destroyed. The vertices can be thought of as representing intermediate "virtual" creation or dissolution of boundstates. A transformed creation operator for an electron, say, is expressed as an infinite series of diagrams with a free end representing an electron field. Each term can be thought of as a process leading to the creation of an electron. The coefficient of each term would give the magnitude of the contribution of this process. A transformed Hamiltonian is an infinite sum of terms representing processes where the interactions of the untransformed Hamiltonian occur, with each interaction process contributing according to the coefficients of the diagrams.

Following is the translation between diagrams and algebraic expressions. *Labels*:

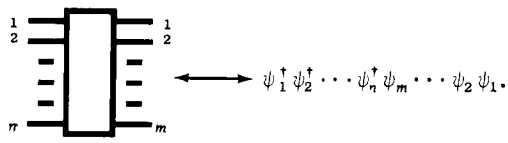
(1) Label every single (i.e., not double) line with $x\mu$. The x stands for coordinates and spin; μ stands for particle kind.

(2) Label every vertex point with $\alpha\nu$. The α stands for coordinates or momentum and spin and other internal quantum numbers; ν stands for particle kind.

(3) Label each double line with the $\alpha\nu$ of the vertex point of the vertex it enters or leaves. In practice this labeling is implicitly assumed and need not be specified.

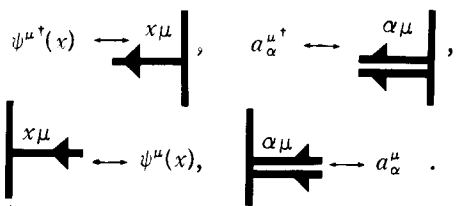
Correspondence:

(1) The field operator structure is given by the lines entering or leaving the body of the diagram, the free lines. The ordering for the (single) fermion lines is



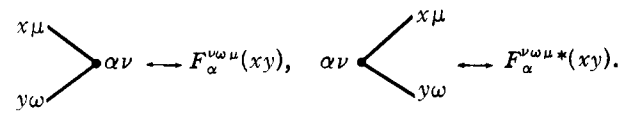
$$\longleftrightarrow \psi_1^\dagger \psi_2^\dagger \cdots \psi_n^\dagger \psi_m \cdots \psi_2 \psi_1 .$$

The composite boson (double) lines may be ordered in the same way by convention, but the ordering is not important for boson operators which commute with each other and fermion operators. The double free lines correspond to field operators denoted with a's. Either the operator or its Hermitian conjugate is used depending on the direction of the line corresponding to the operator. The correspondence for individual free lines is



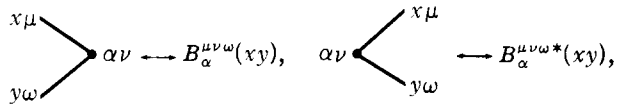
(2) The rest of the diagram structure is given by multiplying algebraic terms corresponding to parts of the body of the diagram. The terms are multiplied in the order from left to right that they appear in the diagram. Where this ordering is not unique because of the two-dimensional nature of the diagrams, it does not matter which specific ordering is used. The following correspondence between parts of the body of the diagram and the algebraic components is made:

a. When both lines entering or leaving a vertex are free, making a free vertex



Recall $F_\alpha^{\nu\omega\mu}(xy)$ is the wavefunction for the boundstate of particle kind ν with coordinates and state α, evaluated for the ω constituent particle at y and the μ constituent particle at x. $F_\alpha^{\nu\omega\mu}(xy)$ is defined to be nonzero for only one ordering of ω, μ.

b. If the vertex is not a free vertex, but linked, the following holds:



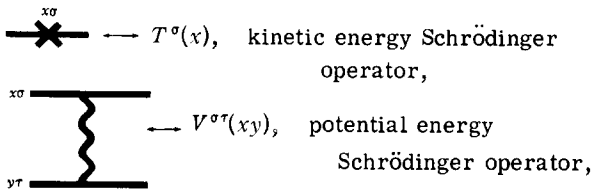
where $B_\alpha^{\mu\nu\omega}(xy) = F_\alpha^{\nu\omega\mu}(xy) - F_\alpha^{\nu\mu\omega}(yx)$.

Geometrically (in the two-fermion boson case), the second F operator is horizontally reflected, or flipped.

$$B = F - F(\text{flipped}) = F - \tilde{F}.$$

When the constituents are the same making up a composite boson, B is twice the F wavefunction since the wavefunction is antisymmetric in indices.

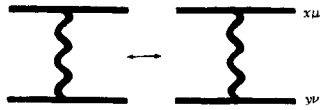
c. Interactions are represented in the following way:



where $T^\sigma(x)$ is the kinetic energy (single particle energy) of particle of kind σ at x, and $V^{\sigma\tau}(xy)$ is the potential energy of particles of kind σ at x and kind τ at y with suitable factors because of double counting in summation over both indices.

(3) Except for the variables of free ends, all indices and variables are summed and integrated over.

A few examples of the application of these rules are given for illustration,



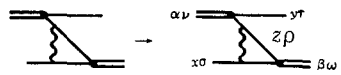
$$\longleftarrow \psi^{\mu\dagger}(x)\, \psi^{\nu\dagger}(y)\, V^{\mu\nu}(xy)\, \psi^\nu(y)\, \psi^\mu(x)$$

(summation and integration is assumed).

For the hydrogen plasma case

$$\begin{aligned} &= \psi^{p\dagger}(x)\, \psi^{p\dagger}(y)\, V^{pp}(xy)\, \psi^p(y)\, \psi^p(x) \\ &+ \psi^{e\dagger}(x)\, \psi^{e\dagger}(y)\, V^{pp}(xy)\, \psi^e(y)\, \psi^e(x) \\ &+ \psi^{p\dagger}(x)\, \psi^{e\dagger}(y)\, V^{pe}(xy)\, \psi^e(y)\, \psi^p(x) \\ &+ \psi^{e\dagger}(x)\, \psi^{p\dagger}(y)\, V^{ep}(xy)\, \psi^p(y)\, \psi^e(x). \end{aligned}$$

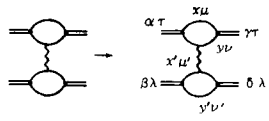
Thus $V^{ep} = V^{pe} = \frac{1}{2}V$ (electron–electron), $V^{pp} = \frac{1}{2}V$ (proton–proton), and $V^{ee} = \frac{1}{2}V$ (electron–electron).



$$\longleftarrow a_\alpha^{\nu\dagger}\psi^{\sigma\dagger}(x)B_\alpha^{\tau\nu\rho*}(yz)\,V^{\rho\sigma}(zx)\,B_\beta^{\rho\omega\sigma}(zx)\,\psi^\tau(y)\,a_\beta^\omega$$

$$= -a_\alpha^\dagger\,\psi^{\rho\dagger}(x)(\varphi_\alpha^*(yz)\,V^{\epsilon\rho}(zx)\,\varphi_\beta(xz))\psi^\rho(y)\,a_\beta$$

$$-a_\alpha^\dagger\,\psi^{\epsilon\dagger}(x)(\varphi_\alpha^*(zy)\,V^{\rho\epsilon}(zx)\,\varphi_\beta(zx))\psi^\epsilon(y)\,a_\beta,$$



$$\longleftarrow a_\alpha^{\tau\dagger}a_\beta^{\gamma\dagger}(B_\alpha^{\mu\tau\nu*}(xy)\,B_\beta^{\mu'\lambda\nu'*}(x'y')\,V^{\nu\mu'}(yx')$$
$$\times B_\gamma^{\mu\tau\nu}(xy)\,B_\delta^{\mu'\lambda\nu'}(x'y'))\,a_\delta^\lambda a_\gamma^\tau.$$

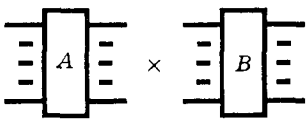
Some simple operations with one diagram follow:

(1) The right–left reflection of a diagram of an operator is the diagram of the hermitian conjugate of that operator. Thus if the diagram is invariant under this reflection it is the diagram of a Hermitian operator.

(2) As long as the free lines and free ends of a diagram remain in the same relative configuration the diagram changes at most by a sign. Each time a vertex is flipped it adds a factor of (-1).

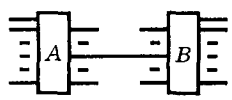
(3) The up–down reflection of a diagram with no free ends is equivalent to the unreflected diagram where free vertices remain unreflected. They are equivalent in the sense that they both correspond to the same operator.

Algebraic multiplication has its counterpart in diagramatic terms. It is more complicated than many diagram representations because this one is noncommutative. Diagrams have been defined so far to correspond only to normally ordered algebraic terms. Below is the correspondence in form between terms multiplied together,



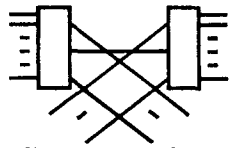
$$\longleftarrow \psi^\dagger\cdots\psi^\dagger A\psi\cdots\psi\psi^\dagger\cdots\psi^\dagger B\psi\cdots\psi.$$

It is easy to see that the contraction of two fermion field operators corresponds to the joining of two of the free field lines, one going out from the right-hand term and one going in from the left-hand term. Since the variables on a line are separately defined and external and internal lines are integrated over, this joining accomplishes setting the contraction equal to a delta function and integrating or summing. The correspondence is illustrated,



$$\longleftarrow \psi^\dagger\psi^\dagger\cdots\psi^\dagger A\psi\cdots\overbrace{\psi\cdots\psi\psi^\dagger\cdots\psi^\dagger}\cdots\psi^\dagger B\psi\cdots\psi\psi.$$

After lines are joined the rest go below the body of the diagram by convention as shown.

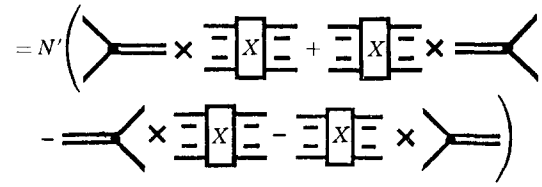


Two diagrams multiplied together can be expressed as a sum of normally ordered diagrams. These diagrams are formed by joining the outgoing free lines of the right diagram with the incoming free lines of the left diagram in all possible combinations from no lines joined to all possible lines joined. All possible combinations are in the sum except where free reduced vertices are joined. When only one line of the free reduced vertex is being joined, only the upper line is joined; when one line each from two free reduced vertices are being joined, only the upper lines are joined; and when both lines from two free reduced vertices are being joined, only the top with the top and bottom with bottom lines are joined. The remaining lines are put in the standard position. The sign is $(-1)^s$. The sign factor, s, is equal to the number of times the original free lines cross. Crossing of lines that were subsequently joined are included in this counting.

4. THEORETICAL FOUNDATIONS

The diagram formalism can be used to view the commutator $[X,F]$, where $F=a_\alpha^{\mu\dagger}A_\alpha^\mu-A_\alpha^{\mu\dagger}a_\alpha^\mu$ is the Tani transformation generator. Since a_α^μ and A_α^μ commute—as opposed to anticonnute—with other operators, commutators like $[X,a_\alpha^{\mu\dagger}]$, $[X,A_\alpha^{\mu\dagger}]$, $[a_\alpha^\mu,X]$, and $[A_\alpha^\mu,X]$ are the normal order expressions of $Xa_\alpha^{\mu\dagger}$, $XA_\alpha^{\mu\dagger}$, $a_\alpha^\mu X$, and $A_\alpha^\mu X$ without the part with no contractions. In terms of diagrams, it is all the linked diagrams in the normal order expansion. This incomplete normal order expansion will be denoted by N'. Thus

$$\begin{aligned}[X,F]&=[X,a_\alpha^{\mu\dagger}A_\alpha^\mu-A_\alpha^{\mu\dagger}a_\alpha^\mu]\\&=A_\alpha^{\mu\dagger}[a_\alpha^\mu,X]+[X,a_\alpha^{\mu\dagger}]A_\alpha^\mu\\&\quad-a_\alpha^{\mu\dagger}[A_\alpha^\mu,X]-[X,A_\alpha^{\mu\dagger}]a_\alpha^\mu\\&=N'(A_\alpha^{\mu\dagger}a_\alpha^\mu X+Xa_\alpha^{\mu\dagger}A_\alpha^\mu-a_\alpha^{\mu\dagger}A_\alpha^\mu X-XA_\alpha^{\mu\dagger}a_\alpha^\mu)\end{aligned}$$



The first two terms break up incoming and outgoing composite bosons. The last two combine fermions into composite bosons or add a new fermion joined with a fermion already present to form a composite boson.

Consider now the set of all operators that can be used to expand X and $[X,F]$. They can be generated in theory by considering one member to be X, the next several, $\{X'\}$, to come from $[X,F]$, the next from $[X',F]$, and so on.

Diagramatically, the set comes from joining vertices to the preceding diagrams. Three things can be concluded from this. Since reduced vertices are the simplest building blocks of a diagram, all diagrams which contain the body of the diagram X will be included in the basis set, and the set will contain only such linked diagrams. The particle order of a diagram is a measure of the number of incoming and outgoing free lines (i.e.,

no. fermion lines $+2\times$ no. boson lines). For either the incoming or outgoing lines, the particle order of members of the basis set will be equal or greater than that of the untransformed operator. This is true since the commutator operation adds pieces to X to form the set of basis elements. The exchange order of a diagram is a measure of the number of vertex points in a diagram. As the set of basis operators is generated, the exchange order increases or stays the same. Thus the exchange order of members of the basis set are equal or greater than that of the untransformed operator.

These conclusions are just the assumptions needed to take the approach outlined in Sec. 2. They show that an expansion can be made of transformed operators and show how to construct one kind of expansion. They show that the d matrix will be of the desired form. More rigorously, on this last point, let the untransformed operator be expanded in the set $\{x_0\}$. One such set is easy to define: the set of all connected diagrams that make up the untransformed operator. Then the commutator of each element with F defines a new set

$$\{[x_0, F]\}=\{x_0\}\cup(\cup_k\{x_k\}).$$

Then since $[X, F]$ produces operators of equal or greater particle order and exchange order,

$$\{[x_l, F]\}=\{x_l\}\cup(\cup_{k>l}\{x_k\}),$$

and so on with the sets being disjoint. The basis is ordered in increasing exchange or particle order as $\{x_0\}$, $\{x_1\}$, $\{x_2\}$, $\cdots$. This gives the triangular character of d. The block diagonal character comes from the size of the sets $\{x_k\}$. It is clear from what $[X,F]$ does that the sets defining blocks must be big enough to include terms with the same internal structure except for free reduced vertices. Thus the particle order will be the same for all members of the block.

The actual structure of d depends on the choice of basis. When this is specified there may be even more structure.

As stated in Sec. 2, it is expected that only a finite number of terms of the transformed operator will be needed. This truncation produces an approximate transformed operator.

A. Particle order truncation

The particle order is the number of bound and free constituents in the term, the number of internal constituent particle lines. For a particle conserving Hermitian operator, with sums and integrations over all field operators, like the Hamiltonian,

Particle order = no. of constituent particle creation operators

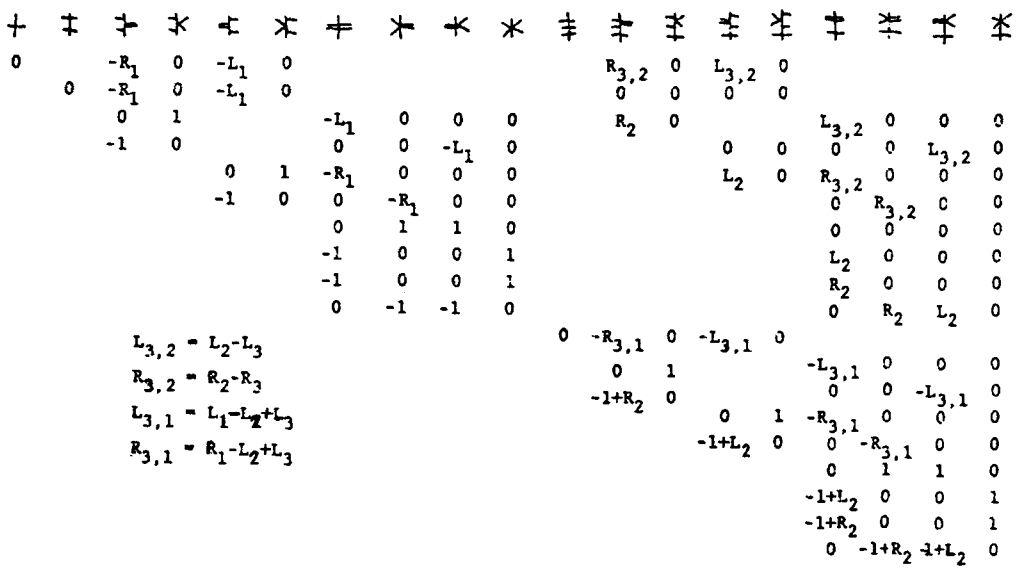
$+2\times$ no. of composite particle creation operators.

= no. of constituent particle annihilation operators $+2\times$ no. of composite particle annihilation operators.

Consider the part, or perhaps it is all, of the diagram connected by internal lines and vertex points, as opposed to only connected by the two particle interaction. The boundstates chosen for the composite particles are those with their constituents existing within a small volume. Their wavefunctions are well localized.

Thus, the diagrams represent significantly large operators only when the coordinates of the internal lines are close together, where the wavefunctions overlap. Further, the coordinates of the incoming and outgoing lines must be close to the internal lines and hence to each other. Assuming that the composite particles are characterized by space coordinates, this means all incoming and outgoing particles must be close together. From dynamical considerations it would seem that a truncation of an expansion is a good approximation after including all diagrams having linked parts with up to two incoming or outgoing lines. Physically this is saying it is a good approximation when it is negligible that more than two particles get within the range of the bound states or when binary collisions dominate.

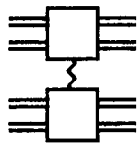
This consideration eliminates all but the kinds of diagrams with pieces with two or less incoming or out-



$L_{3,2} = L_2 - L_3$
$R_{3,2} = R_2 - R_3$
$L_{3,1} = L_1 - L_2 + L_3$
$R_{3,1} = R_1 - L_2 + L_3$

FIG. 1. d matrix for particle orders 1 through 3, using basis introduced in Sec. 6.

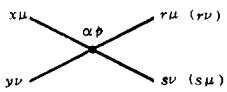
going lines, though two pieces may be joined by an interaction. These terms include, for example, the four-particle interaction term of the form



It is important to realize that by the previous argument the terms of more than two particles entering and leaving, connected by the two-particle interaction, are significant only when the interaction is long range. This, of course, characterizes the Coulomb interaction. But this case has another property which must be considered. It can be shown that the sum over the kinds of constituent particles with opposite charges makes the net interaction fall off more rapidly than $1/r$ like that, for example, between dipoles. Thus it is plausible that all the terms with other than two incoming or outgoing lines can be neglected.

B. Exchange order truncation

The exchange order is the number of vertex points. For a particular field operator structure which specifies the free lines of a diagram, there are many possible particle exchange factors which specify the possible internal diagram structures. As the exchange order increases for a particular field operator structure, the added reduced vertices occur in pairs. One way they occur is this:



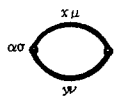
$$\longleftarrow +(-)\sum_{\alpha} B_{\alpha}^{\mu n\nu}(xy)\, B^{\mu n\nu *}(rs).$$

There will also be two new integrations, say over r and s. If α were summed over all bound and unbound states, this would give

$$\int dr\, ds\, \delta(r-x)\,\delta(s-y).$$

This is not the case. In fact, α is only summed over the well localized bound states. As such, it is plausible that the added vertex point gives a factor significantly less than one.

Increased exchange order may come with the addition of two reduced vertices as a "bubble,"

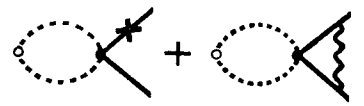


$$\longleftarrow \sum_{\mu\nu} \int dx\, dy\, B_{\alpha}^{\mu\sigma\nu *}(xy)\, B_{\beta}^{\mu\tau\nu}(xy).$$

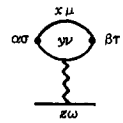
Such a factor has a value $\delta_{\alpha\beta}\,\delta_{\sigma\tau}$ and never occurs in a diagram unless part of it is connected by an interaction. When the interaction occurs inside the bubble it's usually the kinetic plus potential energy diagrams,



If part of the wavefunctions—the relative coordinate part—is an eigenfunction of part of the Hamiltonian, this part would give the eigenvalue and the bubble would collapse down. Thus the result would be exactly the same as if this part of the Hamiltonian were acting on the other side of one of the vertex points, like



For this reason, this kind of bubble addition is included in approximations to the same order as those without the bubble. It is different when the interaction is outside the bubble, like



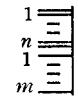
$$\longleftarrow \sum_{\mu\nu} \int dx\, dy\, dz\, B_{\alpha}^{\mu\sigma\nu *}(xy)\, V^{\nu\omega}(yz)\, B_{\beta}^{\mu\tau\nu}(xy).$$

It seems plausible that $V^{\nu\omega}(yz)\, B_{\beta}^{\mu\tau\nu}(xy)$ is no longer expressible in the space spanned by localized bound-states, and as such, the orthonormality condition on the eigenfunctions is no longer an argument for saying this kind of bubble diagram is of the same order as without the bubble. It has more in common with the previous way in which exchange order is increased.

It may be wondered whether the constant coefficients of terms of ever increasing exchange order might compensate for this decreasing magnitude. In Sec. 6 the solution for the third particle order shows that this is not the case, and it is plausible that this is so for other orders as well.

The representation for state vectors is the following:

$$a_{\alpha_1}^{\dagger}\cdots a_{\alpha_n}^{\dagger}\,\psi_1^{\dagger}\cdots\psi_m^{\dagger}\,|0\rangle \longleftarrow$$

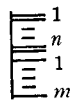


This is consistent with the previous notation. The free end of the $a^{\dagger}$ and $\psi^{\dagger}$ operators have been joined on the left to a vertical line. When such a state is transformed, the particle order is conserved. Terms with a free line going to the right would represent an annihilation operator in the vacuum and are zero. So for a state diagram, no free lines can go across the vertical. This is how diagramatically the representation is consistent with the fact that F commutes—and thus $U=e^{\epsilon F}$ commutes—with the particle order operator. That is,

$$[F,\, 2\sum_{\beta\nu} a_{\beta}^{\nu\dagger} a_{\beta}^{\nu} + \sum_{\sigma} \int dx\, \psi^{\sigma\dagger}(x)\,\psi^{\sigma}(x)] = 0.$$

In the same way, the representation for the conjugate states can be made,

$$\langle 0|\,\psi_m\cdots\psi_1\, a_{\alpha_n}\cdots a_{\alpha_1} \longleftarrow$$



Using the convention that free lines cannot cross the vertical lines of the state or conjugate states and the diagramatic rules for multiplication, the evaluation of matrix elements is defined.

Several conclusions follow easily from the diagramatic representation of states. Consider the transformation of a state like

$$|\psi\rangle = \psi_1^{\dagger}\cdots\psi_n^{\dagger}\,|0\rangle.$$

If increased exchange order is not possible, then the Tani transformed state (denoted with curved bracket) is simply

$$|\psi) \equiv U^{-1}\,|\psi\rangle = |\psi\rangle.$$

This could be the case for a state consisting of all protons or all electrons where the only bound states considered are of both kinds of fermions. It could also be true if the bound states of say particle kinds 1 and 2 only exist in regions such that there is no pair of $\psi^{1\dagger}(x)$ and $\psi^{2\dagger}(y)$ in the definition of $|\psi\rangle$ with x and y in these regions.

The last case may be generalized further by considering the case where there are disjoint regions R_n for which sets of wavefunctions $\{\varphi_{\alpha_n}(xy)\}$ are nonzero only when the coordinates of both constituents fall within the region. Consider a state

$$|X\rangle = X_1^\dagger \cdots X_n^\dagger |0\rangle \longleftarrow \begin{matrix} 1 \\ \\ n \end{matrix},$$

where the $X_k^\dagger$'s are characterized by having only coordinates within R_k. The Tani transform of such a state has a simple form because none of the expansion terms can have links between the $X_k^\dagger$'s. If $X_k^\dagger$ has particle order p_k and the terms of particle order p are denoted $[\ ]_p$, then

$$|X) = [U^{-1} X_1^\dagger U]_{p_1} \cdots [U^{-1} X_n^* U]_{p_n} |0\rangle.$$

A special case is

$$|\alpha\rangle = A_{\alpha_1}^\dagger \cdots A_{\alpha_n}^\dagger |0\rangle,$$

where

$$A_{\alpha_k}^\dagger = \int dx\, dy\, \varphi_{\alpha_k}(xy)\, \psi^\dagger(x)\, \psi^\dagger(y),$$

and

$$\varphi_{\alpha_k}(xy) = 0 \quad \text{if } x, y \notin R_k.$$

Since $[U^{-1} A_{\alpha_k} U]_2 = a_{\alpha_k}$,

$$|\alpha) = a_{\alpha_1}^\dagger \cdots a_{\alpha_n}^\dagger |0\rangle.$$

This just what is expected physically since the Tani transformation should not change the representation of a state of composite particles separated so their wavefunctions do not overlap and there is no possibility of exchange.

A more general conclusion can be made about operators. Consider

$$X = \prod_{k=1}^{n} X_k,$$

where the X_k are normal ordered and are characterized only by coordinates in R_k. Then, no linking is possible between the X_k, so

$$U^{-1} X U = \prod_k (U^{-1} X_k U) = {:}\,\Pi (U^{-1} X_k U)\,{:}',$$

where the normal ordering is done in a particular way: If the terms of the expansion of $U^{-1} X_k U$ are $X_k'^\dagger X_k''$, then the special normal ordering of a product of such terms is defined by

$${:}\prod_k X_k'^\dagger X_k''\,{:}' = X_1'^\dagger X_2'^\dagger \cdots X_n'^\dagger X_n'' \cdots X_2'' X_1'.$$

Diagramatically this means that a product of such operators need not be considered as a horizontal row of diagrams but may be considered as a vertical row of diagrams which are independently transformed.

Such consequences are interesting because they are what is physically expected. Also, they have relevance to some aspects of physical systems like crystals where nuclei are fixed at lattice sites.[17] Finally, they show the aid of the diagrams to developing intuitively understandable proofs which appear more opaque in their algebraic formulation.

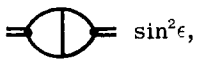
5. THE TRANSFORMED HAMILTONIAN IN LOWEST ORDER

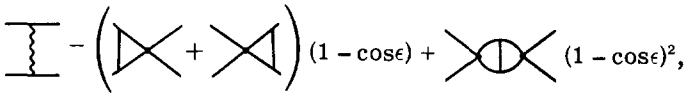
The transformed Hamiltonian in lowest particle and exchange orders has the form

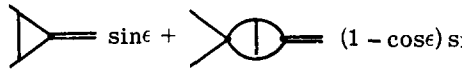
$$\begin{aligned} H = {} & T_\psi + T_a + V_{\psi\psi} + \{U(\psi\psi \leftarrow a) + \text{h.c.}\} + U(a\psi \leftarrow a\psi) \\ & + \{U(\psi\psi\psi \leftarrow a\psi) + \text{h.c.}\} + U(aa \leftarrow aa) \\ & + \{U(a\psi\psi \leftarrow aa) + \text{h.c.}\} + \{U(\psi\psi\psi\psi \leftarrow aa) + \text{h.c.}\}. \end{aligned}$$

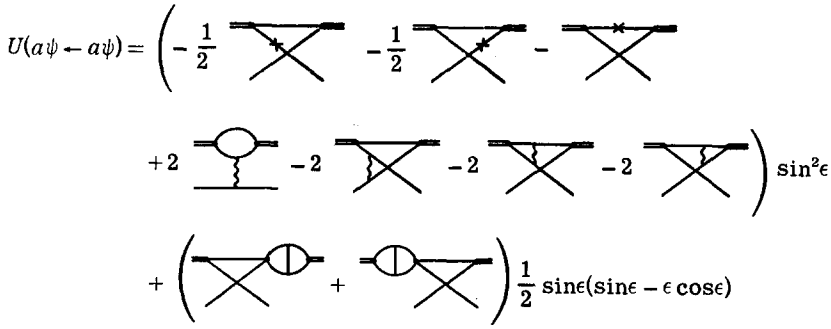
In diagram form with ϵ dependence, it is

$T_\psi =$,

$T_a =$  $\sin^2\epsilon$,

$V_{\psi\psi} =$  ,

$U(\psi\psi \leftarrow a) = -$  ,

$U(a\psi \leftarrow a\psi) =$ 

and so on, with other terms increasing in complexity.[18]

Algebraically it is written for $\epsilon = \pi/2$ as

$$T_\psi = \psi^{\mu\dagger}(x)\, T^\mu(x)\, \psi^\mu(x), \quad T_a = a^{\sigma\dagger}_\alpha h^\sigma_{\alpha\beta} a^\sigma_\beta.$$

Here $V^{\mu\nu}(xy) \to \frac{1}{2} V^{\mu\nu}(xy)$ so that $V^{\mu\nu}(xy)$ is the usual potential

$$[h^\sigma_{\alpha_\beta} \equiv B^{\mu\sigma\nu*}_\alpha(xy)\,(T^\mu(x) + \tfrac{1}{2} V^{\mu\nu}(xy))\, B^{\mu\sigma\nu}_\beta(xy)],$$

$$\begin{aligned} V_{\psi\psi} = {} & \psi^{\mu\dagger}(x)\, \psi^{\nu\dagger}(y) \{\tfrac{1}{2} V^{\mu\nu}(xy)\, \delta_{\mu\mu'}\, \delta_{\nu\nu'}\, \delta(x-x')\, \delta(y-y') \\ & - (T^\mu(x) + \tfrac{1}{2} V^{\mu\nu}(xy))\, B^{\mu\sigma\nu}_\alpha(xy)\, F^{\sigma\nu'\mu'*}_\alpha(x'y') \\ & - F^{\sigma\nu\mu}_\alpha(xy)\, B^{\mu\sigma\nu*}_\alpha(x'y')\, (T^\mu(x') + \tfrac{1}{2} V^{\mu\nu}(x'y')) \\ & + F^{\sigma\nu\mu}_\alpha(xy)\, h^\sigma_{\alpha\beta} F^{\sigma\nu'\mu'*}_\beta(x'y')\} \psi^{\nu'}(y')\, \psi^{\mu'}(x'), \end{aligned}$$

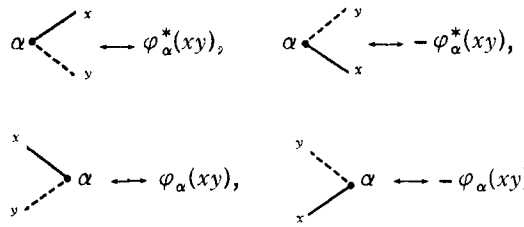
$$\begin{aligned} U(\psi\psi \leftarrow a) = {} & \psi^{\mu\dagger}(x)\, \psi^{\nu\dagger}(y) \{-(T^\mu(x) + \tfrac{1}{2} V^{\mu\nu}(xy)) B^{\mu\sigma\nu}_\alpha(xy) \\ & + F^{\sigma\nu\mu}_\beta(xy)\, h^\sigma_{\beta\alpha}\} a^\sigma_\alpha, \end{aligned}$$

and so on, with other terms increasing in complexity.[18]

For hydrogen, or a similar composite, the transformed Hamiltonian is graphically given next. The dashed lines refer to electrons and the full lines to protons. The bubble has the same meaning as previously. The reduced free vertices are only of the form



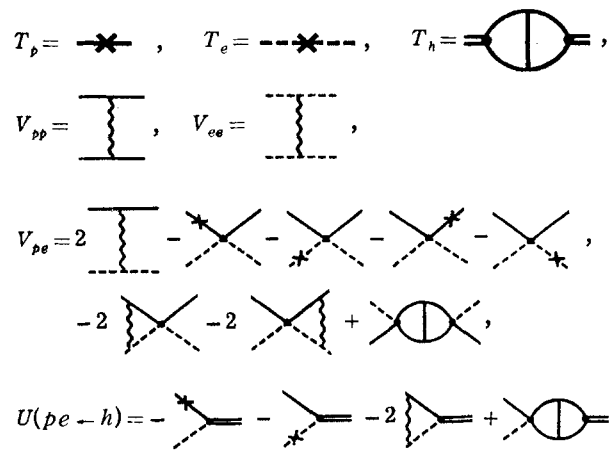
The algebraic correspondence is



The transformed Hamiltonian has the following form in lowest orders:

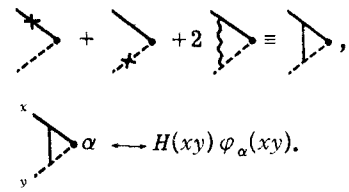
$$\begin{aligned} H = {} & T_p + T_e + T_h + V_{pp} + V_{pe} + V_{ee} + \{U(pe \leftarrow h) + \text{h.c.}\} \\ & + U(hp \leftarrow hp) + U(he \leftarrow he) + \{U(ppe \leftarrow hp) + \text{h.c.}\} \\ & + \{U(pee \leftarrow he) + \text{h.c.}\} + U(hh \leftarrow hh) \\ & + \{U(hpe \leftarrow hh) + \text{h.c.}\} + \{U(ppee \leftarrow hh + \text{h.c.}\}. \end{aligned}$$

The transformed Hamiltonian is given by

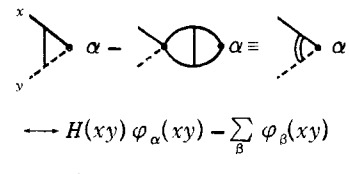


and so on, with other terms increasing in complexity.[18]

Terms containing the kinetic energy operator can be combined with ones with appropriate potential energy interactions to give a two-particle Hamiltonian acting on a reduced vertex wavefunction. Such a combination can be represented as



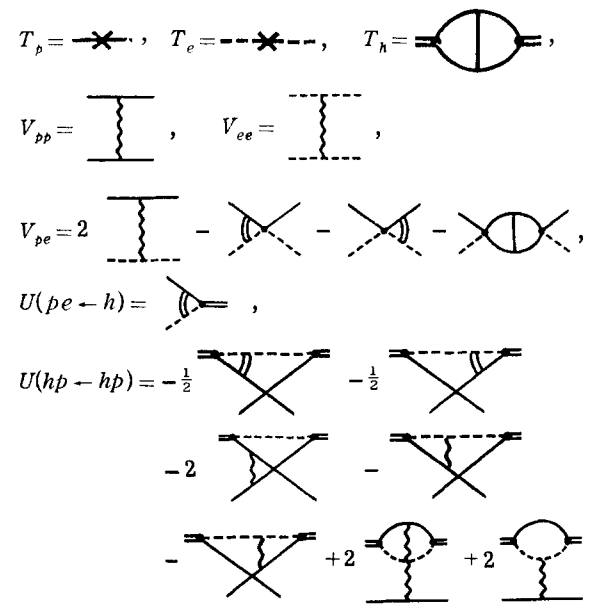
It is also useful to define the combination



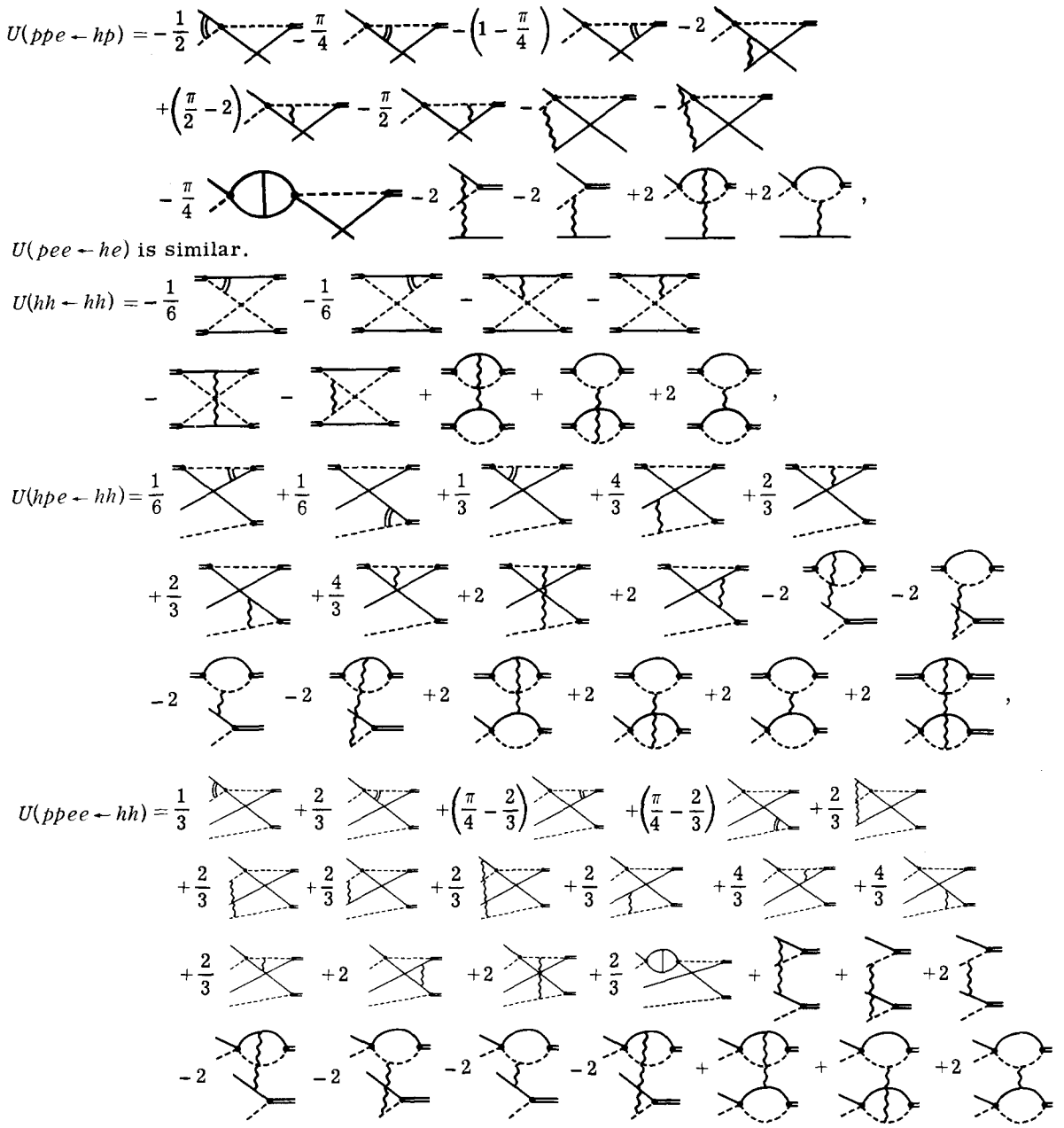
If $H(xy)$ acting on the chosen "bound" state wavefunctions keeps within the space of these wavefunctions, diagrams containing such a combination of vertices are zero.

This is physically reasonable. For example, in a low density gas of composites the correct bound state wavefunctions are expected to be nearly the ground state and perhaps a few excited states of the two-particle Hamiltonian. In such an essentially noninteracting system with these bound states, changes of state of composites or spontaneous break up is not expected, and this corresponding term in the Hamiltonian is zero. On the other hand, the introduction of an external potential into such a system could give rise to changes of state or break up of the composites, and this would accordingly be included in the transformed Hamiltonian.

The Hamiltonian with such terms distinguished is given by the following:



$U(he \leftarrow he)$ is similar.



There is no complete agreement with the other transformation approaches (the Green's functions approach is not easily compared here). There is agreement with the Brittin—Stolt result in many terms. However, their result does not have the bubble terms nor does it have the π factors in some coefficients. Aside from the π factors there is significant disagreement most notably in the spontaneous break up term which goes to zero if the $\varphi_\alpha(xy)$ are energy eigenstates, but remains nonzero in the Brittin—Stolt Hamiltonian. For the reasons already mentioned these terms are expected to vanish in that case according to physical considerations.

The higher particle order terms were not calculated in the Hamiltonian of Nishigori and are not needed for the system treated by him. However, among his low order terms, the spontaneous break up interaction does not include the "bubble" term and thus has the same unphysical property of the Brittin—Stolt breakup interaction.

The transformed Hamiltonian of Sakakura, while similar in many respects, is significantly different in that it contains unlinked terms along with the unphysical break up interaction. The results of a more recent approach of Brittin and Sakakura are in agreement, but are only worked out for noninteracting atoms, protons, and electrons where the boundstate wavefunctions would be eigenfunctions of the two-particle Hamiltonian, and differences would not be apparent.

For a more detailed discussion of the various transformation methods, it is necessary to see to what extent all low order terms as defined here were actually gotten by Brittin and Stolt, Sakakura, and Nishigori. The important "bubble" terms, for example, have an appearance of being a higher order than they are; and in the Tani transformation calculation they are gotten at a later stage.[19] Also, the effect of the subsidiary condition on the Tani-transformed Hamiltonian must be seen.

6. HIGHER EXCHANGE ORDER SOLUTIONS

The Tani transformation is carried out by selecting a basis, forming the d matrix, and solving the matrix differential equation. One choice of basis is the diagrams in the approximate transformed Hamiltonian plus ones of similar exchange and particle orders. The d matrix elements and coefficients are real numbers. Another basis is one that specifies only the external part of the diagram, the free lines and whether they come from free reduced vertices. The solution developed from this basis gives information about all exchange orders, and because the d matrix is independent of the particular operator transformed it gives a general relation among various operators.

In this basis, the coefficients and d matrix elements are not real numbers, but complex integral operators, easily visualized with diagrams. The body of a diagram of particle order n consists of n lines joined in various ways. Operators can be defined, operating on such a set of lines, to join the lines on the left and right in particular ways. Thus besides real number functions of ϵ, coefficients can be represented with terms containing products of such operators.

The operators acting on the right for up to fourth particle order are written R_k, where k means the following is done:

$k=1$: 1 and 2 joined, others follow in relative order;

2 : 1 and 3 joined, others follow in relative order;

3 : 2 and 3 joined, others follow in relative order;

4 : 1 and 4 joined, others follow in relative order;

5 : 2 and 4 joined, others follow in relative order;

6 : 3 and 4 joined, others follow in relative order;

7 : 1 and 2, 3 and 4 joined;

0 : 3 and 4, 1 and 2.

The L_k's which operate on the left are defined similarly. Note that all except $k=0$ operator increases exchange order. In this representation the first three-particle orders are given in Fig. 1.[18]

The Hamiltonian can be transformed easily in the first and second particle orders because there are only a finite number of possible terms. In third order it is more difficult because there are an infinite number of terms of different exchange orders. In low order the solution was found with both the Laplace and integration methods. In fourth particle order only the lowest exchange order solution has been found and this required the substantial use of algebraic computer programming. This is apparently due to the greater difficulty of dealing with noncommuting operators and the more complicated structure of the fourth order diagrams.

The complete first and second exchange order and the low exchange order solutions for the third and fourth orders have been given in Sec. 5. The complete third order solution follows:

$$C_6(\epsilon)=\begin{bmatrix}\sin\epsilon-(1/\phi)\sin\phi\epsilon\\ 1-\cos\epsilon-(1/\phi^2)(1-\cos\phi\epsilon)\end{bmatrix}R_1(T+V)+\begin{bmatrix}(1/\phi)\sin\phi\epsilon\\ (1/\phi^2)(1-\cos\phi\epsilon)\end{bmatrix}(R_2-R_3)V,\quad \phi^2=1-R_2.$$

The first component gives the  coefficient,

$$C_7(\epsilon)=\begin{bmatrix}\sin\epsilon-(1/\theta)\sin\theta\epsilon\\ 1-\cos\epsilon-(1/\theta^2)(1-\cos\theta\epsilon)\end{bmatrix}L_1(T+V)+\begin{bmatrix}(1/\theta)\sin\theta\epsilon\\ (1/\theta^2)(1-\cos\theta\epsilon)\end{bmatrix}(L_2-L_3)V,\quad \theta^2=1-L_2.$$

The first component gives the  coefficient. Next, the first three components give the  ,  and  coefficients,

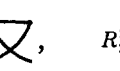
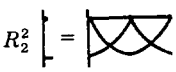
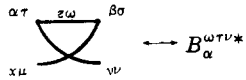
$$C_8(\epsilon)=\begin{bmatrix}\frac{1}{\theta}\sin\theta\epsilon\frac{1}{\phi}\sin\phi\epsilon-\sin^2\epsilon\\ \frac{1}{\theta^2}(1-\cos\theta\epsilon)\frac{1}{\phi}\sin\phi\epsilon-(1-\cos\epsilon)\sin\epsilon\\ \frac{1}{\theta}\sin\theta\epsilon\frac{1}{\phi^2}(1-\cos\phi\epsilon)-(1-\cos\epsilon)\sin\epsilon\\ \frac{1}{\theta^2}(1-\cos\theta\epsilon)\frac{1}{\phi^2}(1-\cos\phi\epsilon)-(1-\cos\epsilon)^2\end{bmatrix}R_1L_1(T+V)+\begin{bmatrix}\frac{1}{\theta}\sin\theta\epsilon\frac{1}{\phi}\sin\phi\epsilon\\ \frac{1}{\theta^2}(1-\cos\theta\epsilon)\frac{1}{\phi}\sin\phi\epsilon\\ \frac{1}{\theta}\sin\theta\epsilon\frac{1}{\phi^2}(1-\cos\phi\epsilon)\\ \frac{1}{\theta^2}(1-\cos\theta\epsilon)\frac{1}{\phi^2}(1-\cos\phi\epsilon)\end{bmatrix}$$

$$\times(-R_2L_1T+(L_2-L_3)(R_2-R_3)V-L_1(R_2-R_3)V-R_1(L_2-L_3)V)$$

$$-\begin{bmatrix} \cos\theta\epsilon\cos\phi\epsilon - 1 + (\theta^2 - \phi^2 - 1)\frac{1}{\theta}\sin\theta\epsilon\frac{1}{\phi}\sin\phi\epsilon \\ \frac{1}{\theta}\sin\theta\epsilon\cos\phi\epsilon - \cos\theta\epsilon\frac{1}{\phi}\sin\phi\epsilon + (1-\phi^2)\frac{1}{\theta^2}(\cos\theta\epsilon - 1)\frac{1}{\phi}\sin\phi\epsilon \\ \frac{1}{\phi}\sin\phi\epsilon\cos\theta\epsilon - \cos\phi\epsilon\frac{1}{\theta}\sin\theta\epsilon + (1-\theta^2)\frac{1}{\phi^2}(\cos\phi\epsilon - 1)\frac{1}{\theta}\sin\theta\epsilon \\ \frac{\theta^2+\phi^2-1}{\theta^2\phi^2}(\cos\theta\epsilon\cos\phi\epsilon - 1) + \frac{1-\phi^2}{\theta^2\phi^2}(\cos\phi\epsilon - 1) + \frac{1-\theta^2}{\theta^2\phi^2}(\cos\theta\epsilon - 1) + \frac{1}{\theta}\sin\theta\epsilon\frac{1}{\phi}\sin\phi\epsilon \end{bmatrix}\frac{1}{L_2R_2}D,$$

$$D = L_2L_3R_1T = R_2R_3L_1T, \quad \theta^2 = 1 - L_2, \quad \phi^2 = 1 - R_2.$$

The powers of L_2 and R_2 in the third particle order have a simple interpretation,

R_2 [=  , R_2^2 [=  , R_2^n [= [[ ]]^n  $\longleftrightarrow B_\alpha^{\omega\tau\nu*}(zy)B_\beta^{\omega\sigma\mu}(zx)$

In the case of hydrogen

$$= \int dz\,\varphi_\alpha^*(zy)\varphi_\beta(zx) + \int dz\,\varphi_\beta^*(yz)\varphi_\beta(xz).$$

This is the sum of the proton plus electron bound state continuum exchange kernels defined by Girardeau.[20]

The R_2 and L_2 operators come in the following ways:

$$\cos(1-x)^{1/2}\epsilon \quad \text{and} \quad \frac{1}{(1-x)^{1/2}}\sin(1-x)^{1/2}\epsilon,$$

Expansion of these operators in terms of x gives the exchange order expansion

$$\cos(1-x)^{1/2}\epsilon = \sum_{n=0}^{\infty}\frac{(-i)^{2n}}{(2n)!}(1-x)^n\epsilon^{2n} = \sum_{k=0}^{\infty}x^k\sum_{n=k}^{\infty}\frac{(-1)^n}{(2n)!}\binom{n}{k}\epsilon^{2n},$$

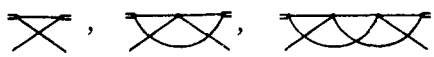
$$(1-x)^{-1/2}\sin(1-x)^{1/2}\epsilon = \sum_{n=0}^{\infty}\frac{(-1)^{2n}}{(2n)!}(1-x)^n\epsilon^{2n+1} = \sum_{k=0}^{\infty}x^k\sum_{n=k}^{\infty}\frac{(-1)^n}{(2n+1)!}\binom{n}{k}\epsilon^{2n+1}.$$

The series in both x and ϵ converge for all ϵ, but for smaller ϵ, the convergence is faster. The first terms illustrate this in the x series,

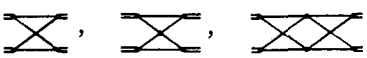
$$\cos(1-x)^{1/2}\epsilon = \cos\epsilon + x\tfrac{1}{2}\epsilon\sin\epsilon + x^2\tfrac{1}{8}(\epsilon\sin\epsilon - \epsilon^2\cos\epsilon) + \cdots,$$

$$(1-x)^{-1/2}\sin(1-x)^{1/2}\epsilon = \sin\epsilon + x\tfrac{1}{2}(\sin\epsilon - \epsilon\cos\epsilon) + x^2\tfrac{1}{8}(3\sin\epsilon - 3\epsilon\cos\epsilon - \epsilon^2\sin\epsilon) + \cdots.$$

There are an infinite number of ϵ values which satisfy the conditions put on the Tani transformation; namely, $\epsilon = \pm(n + \frac{1}{2})(\pi/2)$. However, not all are equally desirable if the exchange order truncation is to be made. Then $\epsilon = \pi/2$ is best. The complete expansion also shows that there are quite complicated nonzero terms. It has been concluded that Tani transformed operators can be expanded in terms of all possible exchange processes. It should be emphasized that this means with all possible intermediate bound states. Even though assumed to be of higher order and neglected, there are terms of all the following forms:

 , etc.,

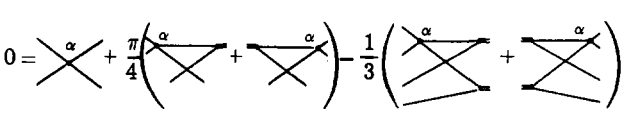
and

 , etc.

Consider the superposition of two atomic proton and electron distributions, for example, and look at its interactions. There are only direct and simple exchange interactions. If each atomic distribution could be uniquely associated with a particular atomic state, one would expect only simple exchange interactions. However, in the introduction it was seen that when such exchange occurs, such distributions cannot be uniquely identified. The expansion over intermediate boundstates is interpreted as a reflection of this fact.

7. THE SUBSIDIARY CONDITION

The subsidiary condition is the transformed statement that there are no ideal composites. In the same lowest orders given in the approximate transformed Hamiltonian, the subsidiary condition consists of 20 diagrams which for $\epsilon = \pi/2$ reduces to

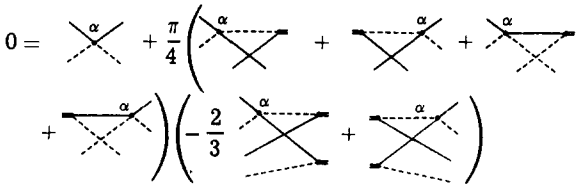
0 = 

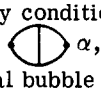
Algebraically it is written

$$0 = \psi^{\mu\dagger}(x)\psi^{\nu\dagger}(y)F_\alpha^{\sigma\nu\mu}(xy)F_\alpha^{\sigma\nu'\mu'*}(x'y')\psi^{\nu'}(y')\psi^{\mu'}(x')$$
$$+ (\pi/4)\{\psi^{\mu\dagger}(x)\psi^{\nu\dagger}(y)\psi^{\omega\dagger}(z)F_\alpha^{\tau\nu\mu}(xy)B^{k\tau\rho*}(sr)$$
$$\times B_{\alpha'}^{k\sigma\omega}(sz)\psi^\rho(r)a_{\alpha'}^\sigma + \text{h.c.}\}$$
$$- \tfrac{1}{3}\{\psi^{\mu\dagger}(x)\psi^{\nu\dagger}(y)\psi^{\omega\dagger}(z)\psi^{\rho\dagger}(r)F_\alpha^{\rho\nu\mu}(xy)B_\alpha^{\pi\sigma\epsilon*}(pq)$$
$$\times B_{\alpha'}^{\pi\sigma\omega}(pz)B_{\beta'}^{\epsilon\tau\rho}(qr)a_{\beta'}^\tau a_{\alpha'}^\sigma + \text{h.c.}\}.$$

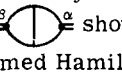
This stands for several subsidiary conditions, each for a boundstate given by the vertex index α.

The subsidiary condition for hydrogen written diagramatically is



If each subsidiary condition is multiplied by the factor corresponding to α  α, the subsidiary condition cancels the diagonal bubble terms in the transformed Hamiltonian expressed with the combined vertices. This subsidiary condition is the transformation of

 $\longrightarrow a_\alpha^\dagger h_{\alpha\alpha} a_\alpha$.

Since $a_\alpha^\dagger a_\alpha |\psi\rangle = 0$ implies $a_\beta^\dagger a_\alpha |\psi\rangle = 0$ where $|\psi\rangle$ is a physical state, the transformation of $a_\beta^\dagger a_\alpha = 0$ gives a subsidiary condition. In particular, the transformation of  shows that all the bubble terms in the transformed Hamiltonian expressed with combined vertices should give zero if the subsidiary condition is satisfied.

In this article the subsidiary condition may be considered as a constraint on the solution states of the transformed Hamiltonian. Satisfaction of it is expected to have no effect on the order of the approximation on the energy of these states in the sense

$$(\langle\psi|U)H(U^{-1}|\psi\rangle) = (\langle\psi|aU)H(U^{-1}a^\dagger|\psi\rangle),$$

since the subsidiary condition commutes with the Hamiltonian. However, satisfying the subsidiary condition directly affects two important variables, the number operators for composites and constituents. The number of constituents, both "bound" and "free" given by the transformed operator

$$N = 2N_a + N_\psi = 2a_\alpha^{\nu\dagger} a_\alpha^\nu + \psi^{\mu\dagger}(x)\psi^\mu(x)$$

remains invariant under the Tani transformation. The value of $\langle N_a\rangle$ in the untransformed states affects $\langle N_a\rangle$ and $\langle N_\psi\rangle$ in the transformed states. Given the total number N as constant, it indirectly affects the energies as well.

A projected Hamiltonian can be formally defined to put the subsidiary condition essentially into the Hamiltonian.[21] The diagrams for such a Hamiltonian are not linked, however, and a generalized expansion of the transform of such operators and an interpretation of them is needed. This and the projected Hamiltonian is the subject of a later article.[22]

One may be fortunate or clever enough in choosing states that it is adequate to check trial solution states to see that the expectation value of the condition is small. In this regard it is thought that to require the expectation value to be exactly zero is too strict since the Hamiltonian and subsidiary condition are approximated. More work will test the adequacy of this treatment.

8. CONCLUSIONS

This work began with the Tani transformation introduced by Girardeau. A linear equation of motion method—using the d matrix—and a diagram formalism were developed. These show the qualitative structure of the transformed operators to be expressed as a sum of terms corresponding to all possible exchange processes and intermediate boundstate involving this operator. This leads to the definitions of an exchange order corresponding to the complexity of such exchange processes and a particle order corresponding to the number of constituent particles involved. Arguments were given showing the plausibility that only low orders in both senses give a good approximation. Specifically, operators with two or less incoming or outgoing particles are thought sufficient in many cases.

The calculation of the transformed Hamiltonian and a subsidiary condition are carried out. The solution for all exchange orders is given for up to third particle order. For fourth particle order the low exchange order solution is given. This Hamiltonian is discussed and compared with the work of others. Though of the same form in many ways, it is found to be significantly different and more physically agreeable in the composite break up interaction. The subsidiary condition is calculated to the same orders, but treated as a constraint. Incorporation of it in a projected Hamiltonian is the subject of a later paper.

The transformation method gives a new Hamiltonian and subsidiary condition, in terms of the constituent interactions and composite wavefunctions. In one sense, these are fundamental parameters for which there is a physical feeling and from which the necessary approximations to make predictions can be made and understood. In turn, these predictions test these parameters; specifically, the existence of particular kinds of composite particles, the demarcation between "bound" and "free," the wavefunctions of boundstates, and the constituent interaction.

This approach is to be contrasted with two others. One is that of starting directly with the correct Hamiltonian expressed only in terms of constituents. Here, in general, going further requires perturbation techniques which neglect the qualitative effects of composite particles. The second is that of using a Hamiltonian expressed in terms of composite and free constituent particle fields, with each interaction a parameter. Qualitatively correct results can be gotten, but there is no derivation of the interactions. Each method fails to link the theory of constituents with the theory of systems with composites.

This link made, new questions arise. What are the results for a nontrivial physical system? What is particularly special, or, as may be the case, common about the Tani transformation?

ACKNOWLEDGMENTS

The author wishes to thank Marvin Girardeau for guidance and collaboration during the development of the work, and Wesley Brittin, Arthur Sakakura, and Mirjana Popović-Božić for helpful discussions.

The work and use of computer facilities was supported in part by National Science Foundation Grant DMR72-03211 and the University of Oregon.

*Supported in part by National Science Foundation grant DMR72-03211.

[1]M. D. Girardeau, J. Math. Phys. **4**, 1096 (1963); **11**, 681 (1970); **12**, 1799 (1971).
[2]M. D. Girardeau, Lect. Theor. Phys. **XIVB**, 147 (1973).
[3]M. D. Girardeau, Phys. Rev. Lett. **27**, 1416 (1971); J. Math. Phys. **16**, 1901 (1975).
[4]R. H. Stolt and W. E. Brittin, Phys. Rev. Lett. **27**, 616 (1971); Lect. Theor. Phys. **XIVB**, 55 (1973).
[5]A. Y. Sakakura, Phys. Rev. Lett. **27**, 822 (1971); Lect. Theor. Phys. **XIVB**, 483 (1973).
[6]W. E. Brittin and A. Y. Sakakura, in *Physical Reality and Mathematical Description*, edited by C. P. Ewz and J. Mehra (Reidel, 1974).
[7]T. Nishigori, Prog. Theor. Phys. **51**, 1387 (1974).
[8]V. S. Vorob'ev and A. L. Khomkin, Teor. Mat. Fiz. **8**, 109 (1971); Teplofiz. Vys. Temp. **10**, 939 (1972); V. S. Vorob'ev and A. I. Gleizer, Teplofiz. Vys. Temp. **12**, 1137 (1974). This work became known to the author after completion of this paper. They independently developed the "Tani" transformation, but carried it out only for the lowest particle order terms, and derived the transformation parameter (ϵ) in a less straightforward way. Their work centers around using these low order terms for making several calculations.
[9]H. L. Shalin and J. L. Schwartz, Phys. Rev. **138**, B267 (1965).
[10]A. Goldberg and R. D. Puff, Phys. Rev. Lett. **30**, 869 (1973); Phys. Rev. A **10**, 323 (1974).
[11]J. D. Gilbert, Ph. D. dissertation, University of Oregon (1974).
[12]P. Ehrenfest and J. R. Oppenheimer, Phys. Rev. **37**, 333 (1931).
[13]M. D. Girardeau, private communication and Ref. 3.
[14]S. Tani, Phys. Rev. **117**, 252 (1960).
[15]Such diagrams were used by Girardeau, Brittin, and Sakakura in Refs. 2—5.
[16]Details are given in Ref. 11.
[17]M. Popović-Božić, Phys. Status Solidi B **65**, 305 (1974); Magnetism and Magnetic Materials—1974 AIP Conference Proceedings No. 24 (American Institute of Physics, New York, 1975), p. 331—32; and to be published. This work uses the approach of Ref. 1. Using the Tani transformation approach, M. Popović-Božić and M. D. Girardeau, to be published.
[18]Details of the other terms and the complete d-matrix through fourth order may be obtained from the author through the University of Oregon.
[19]This is clearly apparent in the iteration method of Ref. 3. Without being careful, it is possible to define these "bubble" terms as being beyond the low orders needed.
[20]In Ref. 2.
[21]This is discussed in Refs. 3 and 11.
[22]J. D. Gilbert, work in progress.

Bernoulli sequences and trajectories in the anisotropic Kepler problem

Martin C. Gutzwiller

IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598
(Received 12 October 1976)

The anisotropic Kepler problem is investigated in order to establish the one-to-one relation between its trajectories and the binary Bernoulli sequences. The Hamiltonian has a quadratic kinetic energy with an anisotropic mass tensor and a spherically symmetric Coulomb energy. Only trajectories in two dimensions with a negative energy (bound states) are discussed. The previous study of this system was based on extensive numerical computations, but the present work uses only analytical arguments. After a review of the earlier results, their relevance to the understanding of the relation between classical and quantum mechanics is emphasized. The main new result is to show the existence of at least one trajectory corresponding to each binary Bernoulli sequence. The proof employs a number of unusual mathematical tools, although they are all elementary. In particular, the virial as a function of the momenta (rather than the action as a function of the position coordinates) plays a crucial role. Also, different kinds of limiting trajectories, mostly involving a collision with the center of the Coulomb attraction are treated in some detail.

INTRODUCTION AND GENERAL MOTIVATION

Among the methods for solving Schrödinger's equation, the oldest one uses the trajectories of the classical mechanical system. Indeed, this method was found by Bohr in 1912 and applied to the H atom before Schrödinger wrote down his equation in 1925. Einstein[1] pointed out in 1917 that Bohr's method and its various refinements had only a very limited range of applicability because they were unable to cope with an ergodic system, or more precisely, with a system whose classical trajectories in phase space do not lie on tori of dimension equal to the number of degrees of freedom.

The use of classicial mechanics to solve a quantum-mechanical problem is important in several instances such as determining asymptotically the energy spectrum for large quantum numbers, or approximating scattering cross sections. Also, one can hardly claim that the relations between classical and quantum mechanics are understood, if one has to exclude all those systems whose behavior in phase space does not fit Einstein's restrictions. Very little effort has gone into overcoming them, although a lot of work has been done which ignores this difficulty.[2] It is not known at present whether it can be surmounted.

The present paper makes an attempt to get closer to a solution of this puzzle by studying a particular example, the anisotropic Kepler problem. This investigation is based on the author's earlier work[3] and may seem to have many points in common with the efforts of other authors who have studied the relations between classical and quantum mechanics in recent years. However, it is important to stress a number of points which are not sufficiently well emphasized elsewhere.

A large and significant body of knowledge has been accumulated by mathematicians who are interested in ergodic systems with a view toward statistical mechanics and the behavior of ordinary differential equations.[4] There has been only one good example to provide complete intuitive understanding, the geodesics on a space of constant negative curvature which were first studied by Hadamard[5] in 1898. The corresponding quantum system is represented by the Laplace operator on this space, and a number of exact results[6] are known which relate the eigenvalues of the Laplacian (kinetic energy) to the closed geodesics (periodic orbits). There are many possible geometries because a closed, compact space of constant negative curvature can be realized in many ways. Each corresponds essentially to introducing reflecting walls at certain well-chosen places so that the particle is constrained to move inside a finite volume. This is exactly Hadamard's billiard ball on a surface of constant negative curvature.[7]

The closed, compact spaces of constant, negative curvature are related to the torus with a flat metric (Euclidean box) because in both cases the universal covering space has a large group of isometries and the Laplace operator can be diagonalized explicitly. However, a general understanding of the relations between classical and quantum mechanics has to be based on more typical examples. Efforts in this direction have recently been started. The spectrum of the Laplacian and the periodic geodesics on a compact Riemannian manifold are being investigated.[8]

The main idea is to study the singularities of the so called elementary solution for the heat equation or the wave equation. Exactly the same approach was used independently (and earlier) by Balian and Bloch,[9] and by the present author to investigate Schrödinger's equation. The spectrum can be viewed in two different ways. One may be interested in the total number $N(\lambda)$ of eigenvalues below some (large) absolute value λ, and obtain an asymptotic expansion for $N(\lambda)$ in decreasing powers of λ.[10] The so called density of states, i.e., the derivative of $N(\lambda)$ with respect to λ, in this approximation looks smooth and presumably monotonic. There is no hint of the quantization which is the most obvious difference between classical and quantum mechanics.

A more detailed view of the spectrum would reveal a density of states with, at the least, some maxima and minima which approximate the set of δ functions of the Schrödinger operator. These oscillations are closely related to the periodic orbits as was pointed

 Copyright © 1977 American Institute of Physics

out first by the author, and subsequently by Balian and Bloch. The formal expression which gives the approximate density of states in terms of the set of all periodic orbits at a given energy, shows that an eigenvalue of the energy is not necessarily associated with a particular periodic orbit. This was demonstrated explicitly for a two-dimensional and three-dimensional flat box and a three-dimensional spherically symmetric potential of arbitrary shape (cf. III). Subsequently, in a preliminary discussion of the anisotropic Kepler problem, a particular set of eigenvalues was directly associated with the simplest periodic orbit available at that time (cf. IV). A number of authors have shown recently that a set of eigenvalues can be associated with a particular periodic orbit provided this orbit is stable against small (linear) deviations.[11] It is not known whether all eigenvalues can be obtained in this way if the periodic orbits of the classical mechanical system are stable, nor is it clear how the results of the author in the case of a flat box and of a spherically symmetric potential fit into this picture.

The partial understanding in the case of stable periodic orbits does not really help to overcome the difficulties which were first pointed out by Einstein. The neighborhood of the stable periodic orbit is surrounded in the linear approximation by invariant tori of dimension equal to the number of degrees of freedom. The Bohr–Sommerfeld rules of quantization can be applied, and the approximate eigenvalues coincide with the more recent results if a number of minor modifications[12] are taken into account. However, when the classical system has ergodic behavior, the basic rules for quasiclassical quantization have still to be established in general and tried out in particular examples.

The development requires three main steps. First, some formal expression has to be found which relates in an approximate manner the density of states for the Schrödinger operator to the set of periodic orbits of the corresponding classical system. Second, a particular mechanical system has to be found where we know the spectrum of Schrödinger's operator (at least numerically) on one hand, and where we can enumerate and effectively compute all the classical periodic orbits. Such a system has to be ergodic in order to be of use for our purpose. Third, the formal expression of the first step is applied to the particular system of the second step so as to study its validity in this special case.

The first step was discussed in IV, although the main idea and some simple examples were treated in the earlier papers. Although the arguments were mostly heuristic, the result is eminently reasonable and turned out to be closely related to expressions which have been proposed and studied by various mathematicians under the name of ζ function.[13]

The second step was made in AK1 with the discussion of the anisotropic Kepler problem in two dimensions. This system had never been considered classically as far as the author is aware. The discussion in AK1 is based partially on the result of extensive numerical calculations, but there is no doubt that this system presents an ideal case of ergodic behavior which can be completely understood. In particular, all its periodic orbits can be obtained explicitly.

The present paper presents an analytical (rather than a numerical) argument for the main claims of AK1. These are concerned with the relationship between binary Bernoulli sequences and the trajectories in the anisotropic Kepler problem. The exact nature of this relationship is explained in Sec. I A, but the main point is a one-to-one continuous mapping (homeomorphism) from the binary sequences onto the trajectories. A proof will be given that there is at least one trajectory for each binary sequence, and, in particular, that there is at least one closed orbit for each periodic binary sequence. The proof consists in demonstrating the existence of a maximum for a certain function. If it can be shown, in addition, that there are no other extrema in the domain of variation, then not only the existence, but also the uniqueness of the trajectory (or closed orbit) will have been established.

The existence proof requires a rather extensive mathematical apparatus, and this is not surprising in view of the utter scarcity of examples for ergodic systems that can be treated analytically. After reviewing the previous work on the anisotropic Kepler problem in Sec. I, some unusual techniques are explained in Sec. II. The most novel of these is the use of the virial rather than the action integral. The author had already pointed out in the earliest papers I and II, that the ordinary Kepler problem can be much better treated semiclassically in momentum space if one is interested in bound states. This turns out to be true for the anisotropic Kepler problem as well, both providing the existence proof and in making stability calculations.

Special attention is given to the trajectories in position space near the origin, i.e., the center of attraction for the particle. The various types of behavior are discussed in Sec. III. Their understanding is important because they provide all the limiting cases for the function whose extrema we are trying to find. This function and the domain of its variables is treated in Sec. IV. It should be noted that we are dealing with a function of a finite number of variables so that our analysis is elementary. Since the number of variables equals the number of conjugate points, the second derivatives near the extrema are directly related to the "interesting" degrees of freedom, i.e., the ones which spoil the minimal property of the virial.

I. REVIEW

A. The equations of motion

An electron which carries the elementary charge $-e_0$ moves around an ion of net charge $+e_0$ in a medium with a large dielectric constant κ_0. Since we are interested in bound (rather than scattering) states, we assume that the energy $-E_0$ of the electron is negative. Its kinetic energy is a quadratic function of the momenta, but the matrix of this quadratic form (inverse mass tensor) is not a multiple of the unit matrix. We chose coordinates in which this matrix is diagonal and the diagonal elements are called $1/m_1$ and $1/m_2$. The ratio m_1/m_2 is typically equal to 5, corresponding to the conditions of a donor impurity in silicon.

All dynamical variables are normalized by taking suitable combinations of e_0, E_0, and $m_0=(m_1 \cdot m_2)^{1/2}$.

Thus, the unit for the position coordinates is $e_0^2/2\kappa_0 E_0$, and the unit for the momentum coordinates is $(2m_0E_0)^{1/2}$. If x is the Cartesian coordinate in the longitudinal direction (large mass), y is the coordinate in the transverse (small mass), and if u, v are the conjugate momenta, then the Hamiltonian (normalized by E_0) becomes

$$H = \frac{u^2}{2\mu} + \frac{v^2}{2\nu} - \frac{1}{(x^2+y^2)^{1/2}}, \tag{1}$$

where $\mu = \sqrt{m_1/m_2}$ and $\nu = \sqrt{m_2/m_1}$. The constant value of the Hamiltonian equals $-\frac{1}{2}$. With the unit of time chosen to be $(m_0e_0^4/2\kappa_0^2E_0^3)^{1/2}$, Hamilton's equations of motion are given by

$$\begin{aligned} \frac{du}{dt} &= -\frac{\partial H}{\partial x} = -\frac{x}{(x^2+y^2)^{3/2}}, & \frac{dx}{dt} &= \frac{\partial H}{\partial u} = \frac{u}{\mu}, \\ \frac{dv}{dt} &= -\frac{\partial H}{\partial y} = -\frac{y}{(x^2+y^2)^{3/2}}, & \frac{dy}{dt} &= \frac{\partial H}{\partial v} = \frac{v}{\nu}. \end{aligned} \tag{2}$$

In our units the action integral $\int(u\,dx + v\,dy)$ would have to be multiplied with $(m_0e_0^4/2\kappa_0^2E_0)^{1/2}$ in order to obtain the ordinary expressions. Apart from the constants m_0, e_0, and κ_0, the energy dependence of the action goes as $E_0^{-1/2}$, just as in the usual Kepler problem. This happens because our Hamiltonian is homogeneous of degree 2 in the momenta, and of degree -1 in the coordinates. This feature allows us to discuss the trajectories at only one particular energy, and to obtain the results for any other energy by simple scaling of each variable.

An alternative formulation of the equations of motion uses polar coordinates in both momentum and position space as follows,

$$\begin{aligned} u &= \sqrt{\mu}\, e^{\chi}\cos\vartheta, \quad & v &= \sqrt{\mu}\, e^{\chi}\sin\vartheta, \\ x &= r\cos\psi, & y &= r\sin\psi, \end{aligned} \tag{3}$$

where $-\infty < \chi < +\infty$ and $0 < r < 2$ are related to each other by the conservation of energy, i.e., $H = -\frac{1}{2}$,

$$r = \frac{2}{1+e^{2\chi}}. \tag{4}$$

The angles ϑ and ψ may be restricted to any interval of length 2π. It is convenient to think of our mechanical system as described by the three variables χ, ϑ, and ψ, or, equivalently, as a vector field in a three-dimensional space with the equations

$$\begin{aligned} \frac{d\chi}{dt} &= -\frac{(1+e^{2\chi})^2}{4e^{\chi}}\,(\sqrt{\nu}\cos\vartheta\cos\psi + \sqrt{\mu}\sin\vartheta\sin\psi), \\ \frac{d\vartheta}{dt} &= -\frac{(1+e^{2\chi})^2}{4e^{\chi}}\,(\sqrt{\mu}\cos\vartheta\sin\psi - \sqrt{\nu}\sin\vartheta\cos\psi), \\ \frac{d\psi}{dt} &= -\tfrac{1}{2}e^{\chi}(1+e^{2\chi})(\sqrt{\nu}\cos\vartheta\sin\psi - \sqrt{\mu}\sin\vartheta\cos\psi). \end{aligned} \tag{5}$$

Obviously, the variables χ, ϑ, and ψ are not canonical.

B. Poincaré maps

Every trajectory is confined to a circle of radius 2 around the origin in position space, $x^2 + y^2 \le 4$, because the energy is normalized to $-\frac{1}{2}$. The (longitudinal) x axis is crossed more often than the (transverse) y axis because of the mass anisotropy. Between any two crossings of the y axis, there is at least one crossing of the x axis. The latter give more information about the trajectory than the former.

When the trajectory crosses the x axis, i.e., when $y = 0$, the coordinates x and u satisfy the inequality

$$|x| \le \frac{2}{1+u^2/\mu}, \quad -\infty < u < +\infty. \tag{6}$$

A plot of all values of x and u which satisfy (6) is somewhat inconvenient. Therefore, we make an area conserving transformation to the coordinates

$$X = x(1+u^2/\mu), \quad U = \sqrt{\mu}\,\arctan(u/\sqrt{\mu}). \tag{7}$$

The inequalities (6) now become simply

$$|X| \le 2, \quad |U| \le \sqrt{\mu}\,\pi/2. \tag{8}$$

To each point in this rectangle of the (X, U) plane corresponds a trajectory which intersects the x axis at

$$x = \frac{X}{1+tg^2(U/\sqrt{\mu})} = X\cos^2\left(\frac{U}{\sqrt{\mu}}\right), \quad u = \sqrt{\mu}\tan\left(\frac{U}{\sqrt{\mu}}\right). \tag{9}$$

In most of the discussion we shall call X and U the position and momentum coordinates at the intersection of the trajectory with the longitudinal (heavy) axis. This slight abuse of nomenclature is not believed to lead to any serious confusion.

The consecutive intersections of a trajectory with the longitudinal axis can be interpreted as mapping the rectangle (8) into itself. This transformation conserves the area and we shall call it the Poincaré map, henceforth. It should be noticed that the boundary of this rectangle and its middle line $X = 0$ correspond to special trajectories. If $X = 0$, the trajectory starts at the origin of position space where the equations of motion are singular. We call this a collision (with the positive ion), and we shall devote a substantial effort to understanding this occurrence.

Similarly, when $|U| = \sqrt{\mu}\pi/2$, the trajectory meets the origin of the position space and we have a collision. However, the boundaries $|X| = 2$ are associated with the special trajectory $y = 0$, $v = 0$, i.e., with the motion of the electron along the longitudinal axis. If $|X|$ is close to 2, the trajectory stays in the neighborhood of the longitudinal axis, its equations of motion can be linearized with respect to y and v. Thus, its intersections with the longitudinal axis are still well defined.

C. Qualitative behavior of the trajectories

In order to gain a complete survey of all possible trajectories in the anisotropic Kepler problem, we have to find a description of their behavior which is at once qualitative and exhaustive. The following scheme has turned out to do exactly what is needed. There is a one-to-one mapping between the qualitative description and the set of all possible initial conditions as represented in the rectangle (8).

Consider the sequence $\cdots, x_{-2}, x_{-1}, x_0, x_1, x_2, \cdots$ of position coordinates at the intersections of the trajectory with the x axis. We associate with this sequence of real numbers a sequence $\cdots, a_{-2}, a_{-1}, a_0, a_1, a_2, \cdots$ of

binaries, $a_i = \pm 1$, according to the rule $a_i = x_i/|x_i|$. The index i on both of these sequences ranges from $-\infty$ to $+\infty$. The binary sequence can be represented with two real numbers, ξ and η, by setting

$$\xi = \sum_{i=1}^{\infty} a_i (1/2)^i, \quad \eta = \sum_{i=0}^{\infty} a_{-i} (1/2)^{i+1}. \tag{9'}$$

If the sequence $\{a_i\}$ is allowed to vary freely over all binaries, the real numbers ξ and η fill out the square

$$-1 \leq \xi \leq 1, \quad -1 \leq \eta \leq 1. \tag{10}$$

The relation between these so-called Bernoulli sequences $\{a_i\}$ and the points (ξ, η) in the square (10) is not one-to-one. E.g., if $\xi = \frac{1}{2}$ we can have either $a_1 = a_2 = 1$ with $a_i = -1$ for $i > 2$, or we can have $a_1 = 1$, $a_2 = -1$ with $a_i = +1$ for $i > 2$. It is the square (10) rather than the Bernoulli sequences $\{a_i\}$ which provides the relevant description of all the trajectories.

The discussion in the preceding paper was based on extensive numerical computations and the detailed consideration of certain special cases. It was found that there is strong circumstantial evidence for the following statement: There is a one-to-one correspondence between the points (X, U) in the rectangle (8) and the points (ξ, η) in the square (10). The exceptions are located along the boundaries and the center lines.

It is important to notice that the correspondence is not made directly with the Bernoulli sequences. Indeed, two sequences which lead to the same numbers ξ and η, represent the same trajectory. E.g., a trajectory which proceeds in time and has the positions $\cdots, x_{-2}, x_{-1}, x_0, x_1$, at its intersections with the x axis (where at least $x_1 \neq 0$), can have a collision with the origin at its second intersection, $x_2 = 0$. In that case it can be considered equally well as the limit of a trajectory with $a_2 = 1$ followed by a long string of $a_i = -1$, as the limit of a trajectory with $a_2 = -1$ followed by a long string of $a_i = +1$.

In the ordinary treatment of Bernoulli sequences these two last sequences would not belong to the same neighborhood. However, it is crucial in our problem to choose a topology where these two sequences are considered to be close to each other. Therefore, we impose the topology of the square (10) on our qualitative description of all trajectories. The mapping between the rectangle (8) and the square (10) is then one-to-one and continuous (i.e., homeomorphic) except at the center lines and the boundaries.

The purpose of this paper is to show that given a sequence $\{a_i\}$ of binaries there is at least one initial condition (X, U) which gives a trajectory with the same sequence of $x_i/|x_i|$.

D. The response function

The qualitative description of all the trajectories is necessary if we attempt to calculate the approximate response function $\tilde{g}(E)$ which was derived in IV. The quantum mechanical response function $g(E)$ is defined as

$$g(E) = \sum_j \frac{1}{E - E_j}, \tag{11}$$

where the summation is extended over all eigenvalues E_j of the Schrödinger operator. The density of states follows from writing

$$(E + i0 - E_j)^{-1} = p(1/(E - E_j)) - i\pi\delta(E - E_j),$$

so that

$$\frac{i}{2\pi}\left[g(E + i0) - g(E - i0)\right] = \sum_j \delta(E - E_j). \tag{12}$$

The energy has to be varied off the real axis. The Cauchy principal part p and the Dirac δ function are being used.

The classical approximation $\tilde{g}(E)$ can be written as the sum over all periodic orbits (p.o.) of the classical mechanical system at the energy E, in the form

$$\tilde{g}(E) = -i \sum_{\text{p.o.}} \frac{T}{h} \exp\left(\frac{i}{h} S(E) - \gamma\right). \tag{13}$$

The letter h represents Planck's constant divided by 2π. For each orbit one has to determine the action integral $S(E) = \oint (p_1\, dq_1 + p_2\, dq_2)$, the period $T = dS/dE$, and half the stability index $\gamma = 1/2(\alpha + i\beta)$. The real part α describes how neighboring trajectories drift away from the particular orbit, and the imaginary part β describes how neighboring trajectories twist around the periodic orbit. A conjugate point gives a contribution π to β.

In the case of the anisotropic Kepler problem we have

$$S(E) = \left(\frac{m_0 e_0}{-2\kappa_0^2 E}\right)^{1/2} \oint (u\, dx + v\, dy), \tag{14}$$

because $E_0 = -E$. The stability index γ is independent of E. There is exactly one conjugate point between any two crossings of the periodic orbit with the x axis. If we define the length of a periodic orbit as the number of intersections with the x axis in one period, then each periodic orbit is characterized by an even number $2l$, and $\beta = 2l\pi$.

So far, we have omitted to mention one small complication which is part of the general formula (13). If a periodic orbit arises because we have traversed a shorter periodic orbit more than once, then the factor T in (13) has to be the period of the shorter orbit while the action $S(E)$ in the exponent is still to be computed including the repeated traversals. Thus, we can add the contributions from all repeated traversals of a particular primitive periodic orbit, by forming the series

$$-i \frac{T}{h} \sum_{n=1}^{\infty} \exp\left[in\left(\frac{1}{h} S(E) - l\pi + i\frac{\alpha}{2}\right)\right]. \tag{15}$$

Furthermore, it is convenient to consider the integrated response function $\int \tilde{g}(E)\, dE$ rather than $\tilde{g}(E)$. Since T vanishes as $(-E)^{-3/2}$ for E going to $-\infty$, the integral can be taken from $-\infty$ to some upper limit E. Its imaginary part represents the number of eigenstates in that interval. Thus, we find that

$$\frac{-i}{h} \int dE \frac{dS}{dE} \cdot \frac{\exp(iS/h - i/\pi - \alpha/2)}{1 - \exp(iS/h - i/\pi - \alpha/2)}$$

$$= \log[1 - \exp(iS/h - i/\pi - \alpha/2)], \tag{16}$$

where we have omitted the contribution $\log[1 - \exp(-i/\pi - \alpha/2)]$ from the lower limit of integration. The sum

over all primitive periodic orbits (p. p. o.) can now be expressed as

$$\int_{-\infty}^{E} \tilde{g}(E)\,dE = \log \prod_{\text{p.p.o.}} [1 - \exp[iS/h - i/\pi - \alpha/2)]. \quad (17)$$

This last expression is almost identical with the "ζ function" which has occasionally been discussed in connection with dynamical systems (Ref. 13).

II. SOME MATHEMATICAL TOOLS

A. The virial

It has become the custom to discuss mainly the propagator in position space when attempts are made to connect classical and quantum mechanics. There are many technical and historical reasons for such an approach which will not be debated in this context. One of the consequences is the use of the action integral $\int(p_1\,dq_1 + p_2\,dq_2)$ which is taken to be a function of its starting position q' and final position q'', in addition to the energy E of the trajectory or the time elapsed t.

On the other hand, it was shown in I that the bound states of the hydrogen atom can be obtained with the help of classical mechanics alone, if we compute the approximate propagator in momentum space. In this approach the crucial role is played by the virial $-\int(q_1\,dp_1 + q_2\,dp_2)$, which is taken to be a function of the starting momentum p' and the final momentum p'', in addition to the energy. If one is interested only in the response function $\tilde{g}(E)$ and, therefore, only in periodic orbits, the difference between the two procedures disappears because the action integral equals the virial. The formulas in the last section could just as well have been written in terms of the virials.

Consider now the action integral S and the virial R along one of our trajectories from an initial point 0 to a final point 1. In our normalized coordinates they are given by

$$S = \int_0^1 \left(u\frac{dx}{dt} + v\frac{dy}{dt}\right) dt, \quad (18)$$

$$R = -\int_0^1 \left(x\frac{du}{dt} + y\frac{dv}{dt}\right) dt. \quad (19)$$

If we define a third integral $T = \int_0^1 dt$, i.e., the time elapsed in going from 0 to 1, we can immediately establish the relation

$$R = \tfrac{1}{2}(S + T). \quad (20)$$

This follows from inserting (2) into (18) and (19), and then using the conservation of energy $H = -\frac{1}{2}$. One can argue intuitively from (20) that the virial varies less dramatically from one trajectory to another.

The choice of the initial point 0 and the final point 1 will now be restricted drastically, and we shall adhere to this choice throughout the remainder of this paper. Both the initial and the final point will always be chosen at an intersection of the trajectory with the longitudinal axis. If we consider two neighboring trajectories and calculate the difference of their action integrals and virials to first order in the difference of their coordinates, we find that

$$\delta S = u_1\delta x_1 - u_0\delta x_0, \quad (21)$$

$$\delta R = -x_1\delta u_1 + x_0\delta u_0. \quad (22)$$

These formulas can be derived from the standard formulas for δS and δR in any textbook, if we remember that we have chosen $y_0 = y_1 = 0$, and, therefore, $\delta y_0 = \delta y_1 = 0$.

These last formulas can be interpreted as follows. If one works with the action integral S, it is natural to consider it as a function of the position coordinates x_0 and x_1. However, if one works with the virial R, it is obvious that the momentum coordinates u_0 and u_1 are to be considered as the independent variables. We shall now try to use U_0 and U_1 rather than u_0 and u_1.

Obviously, one can introduce U_0 and U_1 as independent variables into R because of (9). The interesting consequence of this change in variables is the modified form of (22), i.e.,

$$\delta R = -X_1\delta U_1 + X_0\delta U_0. \quad (23)$$

This follows from $\delta u = \delta U/\cos^2(U/\sqrt{\mu}) = \delta U(1 + u^2/\mu)$. We shall always use R with U_0 and U_1 as the independent variables.

With these preliminaries we will now specify further that the final point 1 has to be the intersection of the trajectory with the x axis which follows immediately upon the starting point 0. In other words, there is no intersection of the trajectory with the x axis between the points 0 and 1. There will be four possibilities for such a trajectory according as the signs of the initial X_0 and final X_1, namely

$$X_0 > 0,\ X_1 > 0, \quad (24a)$$

$$X_0 < 0,\ X_1 > 0, \quad (24b)$$

$$X_0 < 0,\ X_1 < 0, \quad (24c)$$

$$X_0 > 0,\ X_1 < 0. \quad (24d)$$

For each of these possibilities we define a function $R(U_0, U_1)$ and we indicate the relevant signs of X_0 and X_1 as indices, e.g., $R_{+,+}$, $R_{-,+}$, etc.

The domain of each of these functions is contained in the square

$$-\sqrt{\mu}\,\frac{\pi}{2} \leq U_0 \leq \sqrt{\mu}\,\frac{\pi}{2}, \quad \sqrt{\mu}\,\frac{\pi}{2} \leq U_1 \leq \sqrt{\mu}\,\frac{\pi}{2}, \quad (25)$$

but there is no *a priori* guarantee that there exists at most one trajectory of the relevant type (24) which starts with U_0 and ends with U_1. Indeed, if we consider the action integral S as a function of x_0 and x_1, we find two trajectories from x_0 to x_1 for certain pairs of x_0 and x_1. This phenomenon is well known in ballistics where we have a steep and a flat trajectory, both belonging to the same energy.

The various functions $R(U_0, U_1)$ all satisfy a crucial inequality for their mixed second derivative,

$$\frac{\partial^2 R}{\partial U_0\,\partial U_1} > 0. \quad (26)$$

In view of (23) this inequality can be written also in the form

$$\left.\frac{\partial X_0}{\partial U_1}\right|_{U_0} = -\left.\frac{\partial X_1}{\partial U_0}\right|_{U_1} > 0. \tag{27}$$

Before trying to prove (26) or (27) consider its consequences.

The function R will generally be calculated in two steps. First, U_1 is computed for various values of X_0 and U_0 by integrating the equations of motion, and the value of the virial is obtained as a function of the initial coordinates X_0 and U_0. Second, the variable X_0 is replaced by U_1 which requires solving for X_0 as a function of U_1 and U_0. The inequality (27) implies that the second step presents no difficulties, because for constant U_0 the value of X_0 is a monotonically increasing function of U_1. In practice, therefore, we choose some fixed value of U_0 and let X_0 run through one of its two possible ranges, from -2 to 0 and from 0 to 2. The corresponding values of X_1 may also cover part of these two ranges. It is important, however, to be aware of any singular behavior when the trajectory sweeps through the origin in the xy plane, i.e., when there is a collision. Although X_0 may increase in such an instance, U_1 may jump backwards from $+\sqrt{\mu}\pi/2$ to $-\sqrt{\mu}\pi/2$ as the trajectory hits the origin. Thus, the inequality (27) applies to each one of the four regions (24) separately.

The justification of (26) and the detailed shapes of the four domains (24) in the square (25) will be discussed later. The main purpose of this section is to familiarize the reader with the properties of the virial in our problem.

B. The geodesic flow

The response function (13) and its integral (17) with respect to E contains the energy in the most explicit manner, as shown in (14). The main difficulty is really to understand the mechanical system at any fixed energy. That problem is equivalent to investigating the geodesics on a Riemannian surface with the appropriate metric. In such a formulation, our efforts can be compared with the work of a number of mathematicians who have recently studied the spectrum of the Laplacian operator on a Riemannian surface and its relation to periodic geodesics.

The relation between the trajectories of a mechanical system and the geodesics on a Riemannian surface was first pointed out by Jacobi. In this formulation the metric is imposed on the space of position coordinates such that the length of a geodesic is equal to the action integral over the corresponding trajectory. In the case of the Hamiltonian (1) this leads to the metric

$$ds^2 = \left(-1 + \frac{2}{(x^2+y^2)^{1/2}}\right)(\mu\, dx^2 + \nu\, dy^2) \tag{28}$$

in the open disc $x^2 + y^2 < 4$.

It now seems quite natural to generalize Jacobi's idea. As it was shown first in II, the metric is imposed on the space of momentum coordinates such that the length of a geodesic is equal to the virial along the corresponding trajectory. The line element is then given in the open (u, v) plane by

$$ds^2 = 4\left(1 + \frac{u^2}{\mu} + \frac{v^2}{\nu}\right)^{-2}(du^2 + dv^2). \tag{29}$$

It was shown in the preceding paper that this Riemannian surface can be imbedded separately for $u > 0$ and $u < 0$ in a three-dimensional Euclidean space with the induced metric. The two halves can be joined very naturally and smoothly; but if one does so, the whole surface intersects itself, and we get only an immersion. Each half looks like the well-known Nautilus shell. The resulting surface, although of finite volume, is not compact.

If one starts with the metric (29) and writes down the differential equations for the geodesics, one finds in terms of $dt = 2ds/(1 + u^2/\mu + v^2/\nu)$ and $r = 2/(1 + u^2/\mu + v^2/\nu)$ the equations of motion (2). In this interpretation of (2) we can now deduce immediately that the quantity

$$x^2 + y^2 - \frac{4}{(1+u^2/\mu+v^2/\nu)^2} = \frac{16}{(1+u^2/\mu+v^2/\nu)^4}\left[\left(\frac{du}{ds}\right)^2 + \left(\frac{dv}{ds}\right)^2 - \frac{1}{4}\left(1 + \frac{u^2}{\mu} + \frac{v^2}{\nu}\right)^2\right] \tag{30}$$

is constant along the geodesic. Thus, the parameter along the geodesic equals the virial if we make quantity (30) equal to zero. This in turn becomes automatic if we use an angle ψ which is defined by

$$x = 2\cos\psi \Big/ \left(1 + \frac{u^2}{\mu} + \frac{v^2}{\nu}\right), \quad y = 2\sin\psi \Big/ \left(1 + \frac{u^2}{\mu} + \frac{v^2}{\nu}\right), \tag{31}$$

rather than the two variables x and y. The differential equations can now be written in terms of u, v, and ψ rather than in terms of u, v, x, and y. In this way we find that

$$\begin{aligned} \frac{du}{ds} &= -\frac{1}{2}\left(1 + \frac{u^2}{\mu} + \frac{v^2}{\nu}\right)\cos\psi, \\ \frac{dv}{ds} &= -\frac{1}{2}\left(1 + \frac{u^2}{\mu} + \frac{v^2}{\nu}\right)\sin\psi, \\ \frac{d\psi}{ds} &= -\sin\psi\,\frac{u}{\mu} + \cos\psi\,\frac{v}{\nu}. \end{aligned} \tag{32}$$

These are the equations for the geodesic flow. The above derivation has been given in order to show how they are related to the equation of motion (2).

There is a natural metric in the space of coordinates (u, v, ψ) for the geodesic flow. The process of constructing this metric is described in the monograph by Anosov about geodesic flows on closed Riemann manifolds with negative curvature (cf. Ref. 5). In two dimensions the procedure is as follows.

Let the metric tensor be g_{ij} in the coordinates x^1, x^2. The direction of a curve through the point (x^1, x^2) is given by the unit vector (ξ^1, ξ^2). A neighboring curve through a neighboring point $(x^1 + \Delta x^1, x^2 + \Delta x^2)$ is given by the unit vector $(\xi^1 + \Delta\xi^1, \xi^2 + \Delta\xi^2)$. In order to find an expression for the distance $d\sigma$ between these two curve elements which is covariant under coordinate transformations, we introduce the variational quantities $\delta\xi^m = \Delta\xi^m + \Gamma^m_{ke}\xi^k\Delta x^e$. Then we have

$$d\sigma^2 = \sum_{i,j} g_{ij}\,\Delta x^i \Delta x^j + \sum_{i,j} g_{ij}\,\delta\xi^i\delta\xi^j, \tag{33}$$

as the natural metric in the space of the geodesic flow. In our case the metric tensor simplifies to $g_{ij}=\delta_{ij}/f^2$ with $x^1=u$, $x^2=v$, and $f=\frac{1}{2}(1+u^2/\mu+v^2/\nu)$. The unit vector (ξ^1,ξ^2) has the components $\xi^1=-f\cos\psi$, and $\xi^2=-f\sin\psi$, with $\Delta\xi^1=f\sin\psi\,\Delta\psi$ and $\Delta\xi^2=-f\cos\psi\,\Delta\psi$. If we calculate the Riemann–Christoffel symbol Γ^m_{kl} according to the standard formulas, we find with $f_1=\partial f/\partial x^1$ and $f_2=\partial f/\partial x^2$ that

$$d\sigma^2=\frac{1}{f^2}(\Delta x^{1^2}+\Delta x^{2^2})+\left(\Delta\psi+\frac{1}{f}(f_2\Delta x^1-f_1\Delta x^2)\right)^2$$

$$=\frac{1}{f^2}(\Delta u^2+\Delta v^2)+\left[\Delta\psi+\frac{1}{f}\left(\frac{v}{\nu}\Delta u-\frac{u}{\mu}\Delta v\right)\right]^2. \quad (34)$$

This derivation was given explicitly because the result is not trivial.

Now define three mutually orthogonal unit vectors in this metric,

$$\begin{aligned} a&=(a^u,a^v,a^\psi)=(f\sin\psi,\ -f\cos\psi,\ -(u/\mu)\cos\psi\\ &\quad-(v/\nu)\sin\psi),\\ b&=(0,0,1),\\ c&=(-f\cos\psi,\ -f\sin\psi,\ -(u/\mu)\sin\psi+(v/\nu)\cos\psi). \end{aligned} \quad (35)$$

With $\xi^1=u$, $\xi^2=v$, and $\xi^3=\psi$, we can write the metric (34) in the abbreviated form $\sum\gamma_{\kappa\lambda}\Delta\xi^\kappa\Delta\xi^\lambda$. A vector product W_ρ between two vectors U^κ and V^λ is defined by

$$W_\rho=(U\times V)_\rho=\sqrt{\gamma}\sum_{\kappa\lambda}\epsilon_{\rho\kappa\lambda}U^\kappa V^\lambda, \quad (36)$$

where γ is the determinant of $\gamma_{\kappa\lambda}$ and $\epsilon_{\rho\kappa\lambda}$ is the antisymmetric Kronecker symbol. One checks that $c=a\times b$.

Furthermore, we define the divergence of a vector by

$$\mathrm{div}\,V=\frac{1}{\sqrt{\gamma}}\frac{\partial}{\partial\xi^\kappa}(\sqrt{\gamma}\,V^\kappa), \quad (37)$$

as usual, and find that $\mathrm{div}a=\mathrm{div}b=\mathrm{div}c=0$. Finally, we define the curl of a vector through

$$(\mathrm{crul}\,V)^\rho=\frac{1}{\sqrt{\gamma}}\sum\epsilon_{\rho\kappa\lambda}\left(\frac{\partial V_\lambda}{\partial\xi^\kappa}-\frac{\partial V_\kappa}{\partial\xi^\lambda}\right), \quad (38)$$

and find that $\mathrm{curl}a=-a$ and $\mathrm{curl}b=-Kb$. The Gaussian curvature K in the original metric (29),

$$K=(\mu+\nu)f-u^2/\mu^2-v^2/\nu^2, \quad (39)$$

is a remarkably simple function of u and v.

The above geometric quantities have been given in detail because it turns out that calculations concerning the stability of the trajectories in the mechanical system are most easily done with the help of the geodesic flow. Let $u(s)$, $v(s)$, $\psi(s)$ describe a particular geodesic, and consider a neighboring geodesic whose coordinates deviate by the amounts $\delta u(s)$, $\delta v(s)$, $\delta\psi(s)$. The parameter s in each case refers to the virial (27) which is appropriate to the particular geodesic. The deviations are decomposed along the local vectors a, b, and c, i.e.,

$$(\delta u,\delta v,\delta\psi)=\alpha a+\beta b+\gamma c, \quad (40)$$

where α, β, and γ measure the components of the deviation along the three mutually perpendicular directions. After some fairly straightforward but tedious calculations, the equations for α, β, and γ as functions of s can be found,

$$\frac{d\alpha}{ds}=\beta,\quad \frac{d\beta}{ds}=-K\alpha,\quad \frac{d\gamma}{ds}=0. \quad (41)$$

These are the well known equations of Jacobi for the geodesic deviation.

As long as the momenta u and v remain bounded along a particular trajectory, there is no difficulty in solving Eqs. (41). The only opportunity for u or v to become unbounded is a collision with the origin of the xy plane. However, a periodic orbit never has a collision so that Eqs. (41). always apply to a periodic orbit. This holds in particular when the periodic orbit goes through the origin of the uv plane, i.e., when it hits the boundary $x^2+y^2=4$. If one bases the stability calculations on the metric (28), or, equivalently, if one applies the usual procedure for finding the linear deviations at constant energy to the equations of motion (2), he finds singular equations near $x^2+y^2=4$. This follows immediately from the expression for the Gaussian curvature with the metric (28), namely

$$K=\frac{1}{r^3(2-r)^3}\{(2-r)[(\mu-2\nu)x^2+(\nu-2\mu)y^2]$$

$$+2(\mu y^2+\nu x^2)\}, \quad (42)$$

where we have used $r=(x^2+y^2)^{1/2}$.

The stability calculations for periodic orbits are essential for two reasons. First, the second derivatives of the virial or of the action integral can only be obtained in this way, and these in turn yield the extremum properties of the orbit. Second, both the real and the imaginary part of the stability index enter into the expression (13) or (17) for the approximate response function.

C. The autonomous region

The binary sequences which are associated with a particular trajectory can change only when the trajectory sweeps through the origin of the xy plane, i.e., when there is a collision. The investigation of a trajectory near a collision becomes, therefore, important in our understanding of the relation between the trajectories and their binary sequences. This point of view was discussed to some extent in the previous paper, but the arguments can now be simplified, and certain aspects can be emphasized as particularly relevant to the purpose of this paper.

The equations of motion in the form (5) are well suited, because the neighborhood of the origin is given according to (4) if the variable χ is very large. In that limit the shape of the solutions in the $\vartheta\psi$ plane becomes independent of χ. Indeed, if we use χ as the independent parameter instead of t, we can write Eq. (5) as

$$\begin{aligned} \frac{d\vartheta}{d\chi}&=\frac{\sqrt{\mu}\cos\vartheta\sin\psi-\sqrt{\nu}\sin\vartheta\cos\psi}{\sqrt{\nu}\cos\vartheta\cos\psi+\sqrt{\mu}\sin\vartheta\sin\psi},\\ \frac{d\psi}{d\chi}&=\frac{2}{1+e^{-2\chi}}\cdot\frac{\sqrt{\nu}\cos\vartheta\sin\psi-\sqrt{\mu}\sin\vartheta\cos\psi}{\sqrt{\nu}\cos\vartheta\cos\psi+\sqrt{\mu}\sin\vartheta\sin\psi}. \end{aligned} \quad (43)$$

For large values of χ, the first factor on the right-hand side of the second equation simplifies to 2. The Eqs. (43) represent then a vector field in the $\vartheta\psi$ plane which does not depend on χ. The system of equations (43) be-

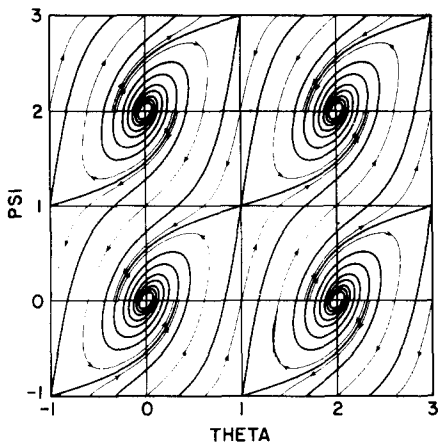


FIG. 1. Projection of the vector field (5) into the (ϑ, ψ) plane in the limit of large χ. The units are multiples of $\pi/2$. The vector field has the period 2π in both ϑ and ψ. The arrows indicate increasing time.

comes autonomous, and its solutions can be discussed with the help of a simple two-dimensional diagram, such as Fig. 1. The arrows indicate the direction of increasing time. Opposite boundaries have to be identified as is usual on a flat torus.

The curves in Fig. 1 were computed for the case $\mu^2 = 5$. But the general behavior can be inferred directly from Eq. (5) or (43). If the three-dimensional vector field (5) is projected into the $\vartheta\psi$ plane for some fixed value of χ, its singularities are immediately obvious. The right-hand sides of the second and third equations vanish simultaneously at two different types of points

(a) $\sin\vartheta = \sin\psi = 0$,

(b) $\cos\vartheta = \cos\psi = 0$.

In both cases the right-hand side of the first equation differs from zero, and χ can be taken as the independent variable rather than t.

As an example of (a) let us consider the neighborhood of $\vartheta = \psi = 0$, and expand the right-hand sides of (43) up to linear terms in ϑ and ψ. We find that

$$\frac{d\vartheta}{d\chi} \cong -\vartheta + \mu\chi, \quad \frac{d\psi}{d\chi} \cong \frac{2}{1+e^{-2\chi}}(-\mu\vartheta + \chi). \tag{44}$$

As an example of (b) consider the neighborhood of $\vartheta = \chi = \pi/2$. In terms of $\vartheta' = \vartheta - \pi/2$ and $\psi' = \psi - \pi/2$ we find similarly that

$$\frac{d\vartheta'}{d\chi} \cong -\vartheta' + \nu\psi', \quad \frac{d\psi'}{d\chi} \cong \frac{2}{1+e^{-2\chi}}(-\nu\vartheta' + \psi'). \tag{45}$$

These equations are to be compared with (37) and (39) of the previous paper, where a different (and more awkward) coordinate system is used to describe an elliptic singularity in the vector field, while Eqs. (45) describe a hyperbolic singularity. Henceforth, we shall always speak about the hyperbolic points or the elliptic points in the $\vartheta\psi$ plane. The detailed discussion will always be made under the assumption of very large χ which simplifies the last equation of (44) and (45).

The singularities (44) and (45) are important because the value of χ can change by a large amount only in their neighborhood. Thus, a trajectory in the neighborhood of the origin (large χ) will stay there unless it gets close to one of the singular points. Actually, a closer look shows that every trajectory in Fig. 1 comes out of the elliptic points at $\vartheta = 0$, $\psi = \pi$ or at $\vartheta = \pi$, $\psi = 0$, and eventually goes into the elliptic points at $\vartheta = 0$, $\psi = 0$ or at $\vartheta = \pi$, $\psi = \pi$. Only those exceptional trajectories which hit any one of the hyperbolic points after coming out of an elliptic point, or which come out of a hyperbolic point to go into an elliptic point, are exempt from this rule.

It can be checked very easily that χ decreases indefinitely as the trajectory gets closer and closer to one of the elliptic points. Since both ϑ and ψ are multiples of π at each elliptic point, the trajectory which approaches one of them, moves away from the origin of the xy plane along the x axis. The spiraling of such a trajectory around the elliptic point generates a long string of points where ψ is a multiple of π (always the same), i.e., it generates a long string of identical binaries in the associated Bernoulli sequence. In other words, such a trajectory has a value of ξ or η according to (9) which is close to a multiple of $(\frac{1}{2})^n$ with some low integer $n > 0$. The length of such a string of identical binaries is determined by the value of χ as the trajectory enters into the spiral around the elliptic point. This initial value of χ determines how many times the trajectory will be able to cross the x axis before it reaches the end at $x = \pm 2$. The trajectory will then turn around and move back toward the origin along the x axis. Eventually, it will reappear in the autonomous region by coming out of an elliptic point.

The detailed discussion of this process will help us to establish simple relations between the rate of change in the binary characteristics ξ and η on one hand, and the rate of change in the initial conditions X and U on the other hand. In this manner, one can show the continuity of the mapping between (ξ, η) and (X, U).

Of more interest for our present purpose is the discussion of the hyperbolic points. Again, the variable χ can change by arbitrarily large amounts provided the trajectory gets close enough. There are two directions of approach, the two asymptotes. Both are in the first or third quadrant, so that we can call one of them horizontal, i.e., nearer to the ϑ direction, and the other vertical, i.e., nearer to the ψ direction. One checks easily that χ always increases when the trajectory approaches the hyperbolic point along the horizontal asymptote, and χ decreases as the hyperbolic point is approached along the vertical asymptote. In both cases the trajectory has both ϑ and ψ near a half-integer multiple of π, i.e., the trajectory moves along the y axis. However, when it gets away from the origin in the xy plane, it does not in general return to the autonomous region. Thus, a trajectory near a hyperbolic point is a collision trajectory.

In later discussions of collision trajectories it will be necessary to look at linear deviations of trajectories in the autonomous region. These follow directly from (43) and are given by

$$\begin{aligned} \frac{d\delta\vartheta}{d\chi} &= \frac{1}{\gamma^2}[-(\mu\sin^2\psi + \nu\cos^2\psi)\,\delta\vartheta + \delta\psi], \\ \frac{d\delta\psi}{d\chi} &= \frac{2}{1+e^{-2\chi}} \cdot \frac{1}{\gamma^2}[-\delta\vartheta + (\mu\sin^2\vartheta + \nu\cos^2\vartheta)\,\delta\psi], \end{aligned} \tag{46}$$

where $\gamma = \sqrt{\nu}\cos\vartheta\cos\psi + \sqrt{\mu}\sin\vartheta\sin\psi$. It should be noted that γ differs from 0 in the neighborhood of each hyperbolic point.

III. THREE TYPES OF LIMITING BEHAVIOR

A. Trajectories near the origin

The autonomous region contains some trajectories which do not go through the singular points, but which correspond to some important limiting cases in the virial $R(U_0, U_1)$. The trajectories can be immediately recognized in Fig. 1. If $X_0 > 0$, then $\psi_0 = 0$. It suffices to let ϑ_0 vary through the interval from 0 to π because we restrict our attention to trajectories in the upper half-plane $y > 0$. The endpoint of any trajectory occurs when its image in the $\vartheta\psi$ plane reaches either $\psi_1 = 0$ or $\psi_1 = \pi$ for the first time.

The relation with the initial coordinates X_0, U_0 and the final coordinates X_1, U_1 follows from (7) and (3),

$$X = 2\cos\psi\,\frac{1+e^{2\chi}\cos^2\vartheta}{1+e^{2\chi}} \cong 2\cos\psi\cos^2\vartheta,$$

$$U = \sqrt{\mu}\,\operatorname{artan}(e^{\chi}\cos\vartheta) \cong \operatorname{sign}(\cos\vartheta)\left(\frac{\pi}{2} - \frac{1}{e^{\chi}|\cos\vartheta|}\right). \tag{47}$$

The expression on the right-hand side can be simplified in the indicated manner provided χ is large and, in the second equation, $\cos\vartheta$ is bounded away from 0. It should be noted that X does not in general vanish, although the position coordinate x has a small absolute value in the autonomous region.

There are a number of special values for ϑ between 0 and π when $\psi = 0$, which will be given special names in order to simplify the future discussion. Call ϑ_v the angle between 0 and $\pi/2$ from which the trajectory in the $\vartheta\psi$ plane goes directly into the hyperbolic point at $\vartheta = \psi = \pi/2$. Call ϑ_h the angle between 0 and $\pi/2$ such that the trajectory starting at $\vartheta = \pi - \vartheta_h$ goes directly into the hyperbolic point at $\vartheta = 3\pi/2$, $\psi = \pi/2$. Notice that the hyperbolic point is approached along the vertical asymptote in the first case, and along the horizontal asymptote in the second case. Thus the names ϑ_v and ϑ_h. Call ϑ_c the angle between 0 and $\pi/2$ from which the trajectory starting at $\vartheta = \pi - \vartheta_c$ goes directly into the point $\vartheta = 3\pi/2$, $\psi = \pi$. Finally, call ϑ_s the angle between 0 and $\pi/2$ such that the trajectory starting at $\pi - \vartheta_s$ goes directly into the point $\vartheta = \pi$, $\psi = \pi/2$. The last trajectory is symmetric with respect to this point, and therefore the name ϑ_s. A little examination of Fig. 1 shows that

$$0 < \vartheta_h < \vartheta_c < \vartheta_s < \vartheta_v < \pi/2. \tag{48}$$

While these inequalities are read off Fig. 1, they can no doubt be derived "purely analytically" from a detailed discussion of the vector field (43) in the limit of large χ.

Consider now the curves with $0 < \vartheta_0 < \vartheta_v$ and $\psi_0 = 0$. Their first intersection with either $\psi = 0$ or $\psi = \pi$ occurs at $\psi_1 = 0$ in the range $-\vartheta_h < \vartheta_1 < 0$. Thus, they belong to the family of (24a). The change in χ is finite along any of these trajectories. If χ_0 is large, so is χ_1. The approximations in (47) apply, and we find that both U_0 and U_1 are near (but below) $\sqrt{\mu}\,\pi/2$. From (4) and (5) we find that

$$R = \int_0^1 \frac{2e^{\chi}}{1+e^{2\chi}} \cdot \frac{d\chi}{\sqrt{\nu}\cos\vartheta\cos\psi + \sqrt{\mu}\sin\vartheta\sin\psi} \tag{49}$$

goes to zero as U_0 and U_1 approach $\sqrt{\mu}\pi/2$. The differences $\sqrt{\mu}\pi/2 - U_0$ and $\sqrt{\mu}\pi/2 - U_1$ depend on the quantities (χ_0, ϑ_0) and (χ_1, ϑ_1) as shown in (47). In particular, we have

$$\frac{\sqrt{\mu}\pi/2 - U_1}{\sqrt{\mu}\pi/2 - U_0} = e^{\chi_0 - \chi_1}\frac{\cos\vartheta_0}{\cos\vartheta_1}. \tag{50}$$

This ratio depends only on ϑ_0 because both $(\chi_0 - \chi_1)$ and ϑ_1 can be calculated from ϑ_0 directly by integrating Eqs. (43) in the limit of large χ. When ϑ_0 is near 0, ϑ_1 is near zero, too, and $\chi_0 - \chi_1$ has a value which can be found from integrating (44). As ϑ_0 increases, the ratio (50) increases and goes to ∞ when ϑ_0 approaches ϑ_v, because the trajectory comes close to the hyperbolic point at $\vartheta = \psi = \pi/2$. Therefore all these trajectories fill out a sector in the upper right-hand corner of the square (25) with both $x_0 > 0$ and $x_1 > 0$. As ϑ_0 approaches ϑ_v, the final value χ_1 can be made as small as we please even if χ_0 is large. This implies that the trajectory leaves the autonomous region moving up along the positive y axis and then out into the xy plane. These trajectories ("coming out of a collision") will be discussed in the next subsection.

In the interval $\vartheta_v < \vartheta_0 < \pi - \vartheta_h$ we have $\psi_1 = \pi$, i.e., $x_0 > 0$ and $x_1 < 0$, so that we are now in family (24d). Again, while ϑ_0 is near ϑ_v there is an arbitrarily large decrease from χ_0 to χ_1, and the trajectory comes out of a collision. As ϑ moves away from ϑ_v and towards $\pi/2$, the decrease in χ remains bounded and the approximation in (47) is valid as long as ϑ_0 remains bounded away from $\pi/2$. Thus, U_0 is near $\sqrt{\mu}\pi/2$ and U_1 is near $-\sqrt{\mu}\pi/2$. However, as ϑ_0 moves into $\pi/2$, the value of U_0 changes very quickly from $\sqrt{\mu}\pi/2$ to 0, even with large χ_0. And as ϑ_0 moves away from $\pi/2$, the value of ϑ_0 goes from 0 to $-\sqrt{\mu}\pi/2$. During these changes in U_0 the value of U_1 remains near $-\sqrt{\mu}\pi/2$. At the same time X_0 equals $2\cos^2\vartheta_0$ except in the neighborhood $\vartheta_0 \sim \pi/2$ where X_0 does not quite go to zero, but reaches a minimum value of $2/(1+e^{2\chi_0})$. The value of X_1 remains close to $-2\cos^2\vartheta_c$. Therefore, as ϑ_0 moves through $\pi/2$, the point (U_0, U_1) moves along the lower boundary of the square (25).

When ϑ_0 goes from $\pi/2$ to $\pi - \vartheta_c$, both U_0 and U_1 are near $-\sqrt{\mu}\pi/2$. As ϑ_0 moves through $\pi - \vartheta_c$, the final value ϑ_1 moves through $3\pi/2$. A similar discussion now shows that U_1 goes very quickly from $-\sqrt{\mu}\pi/2$ to $+\sqrt{\mu}\pi/2$, while U_0 stays near $-\sqrt{\mu}\pi/2$. The point (U_0, U_1) moves along the left-hand boundary of the square (25). The value of X_0 is close to $2\cos^2\vartheta_c$ while the value of X_1 dips almost to zero when $\vartheta = 3\pi/2$. Notice the symmetric trajectory where $\vartheta_0 = \pi - \vartheta_s$, $\vartheta_1 = \pi + \vartheta_s$, $\chi_0 = \chi_1$, and, therefore, $U_0 = U_1$. In the interval $\vartheta_v < \vartheta_0 < \pi - \vartheta_h$, we can use the formula

$$R = \int_0^1 e^{-\chi}\frac{d\psi}{\sqrt{\mu}\sin\vartheta\cos\psi - \sqrt{\nu}\cos\vartheta\sin\psi}, \tag{51}$$

to show that R is small as long as ϑ_0 is bounded away from ϑ_v and $\pi - \vartheta_h$. All this time, we have $X_0 > 0$ and $X_1 < 0$.

When ϑ_0 moves through $\pi - \vartheta_h$, the situation is the reverse of what happened when ϑ_0 moved through ϑ_v. There is now an arbitrarily large increase from χ_0 to χ_1. In other words, if χ_1 is chosen at some large value, χ_0 can be made arbitrarily small provided ϑ_0 is close enough to ϑ_h. The trajectory now goes into a collision, and the details will be discussed in the next subsection.

Finally, in the interval of ϑ_c from $\pi - \vartheta_h$ to π, both ψ_0 and ψ_1 vanish. The trajectories are the time reversed ones with respect to the interval $0 < \vartheta_0 < \vartheta_v$, and belong to the lower left-hand corner in the square (25) with both $X_0 > 0$ and $X_1 > 0$. Again, only a finite sector of this corner is covered with trajectories because the analog of the ratio (50) does not cover the full range from 0 to ∞. The value of R can be shown to approach zero in this corner with the help of the same expression (49).

B. Collision trajectories

The collisions with the origin of the xy plane are essential to the ergodic behavior in the anisotropic Kepler problem. Their discussion is fairly involved and requires some patience. If one refers to Eqs. (5), he can give the following summary of the procedure in this section.

Instead of (X_0, U_0) and (X_1, U_1), the initial and final coordinates are given in terms of (χ_0, ϑ_0) and (χ_1, ϑ_1). Between these two, the trajectory passes close to the hyperbolic point at $\vartheta = \psi = \pi/2$. In its neighborhood two entirely different parameters, A_+ and A_-, are used to characterize the trajectory. Thus, we have a sequence of two mappings, from (χ_0, ϑ_0) to (A_+, A_-), and from (A_+, A_-) to (χ_1, ϑ_1). We are looking for the simplest form of these mappings, but a linear map will not do in either case. However, it is possible to approximate at least one parameter in each map by linearizing Eqs. (5). The first step will be to introduce the parameters A_+ and A_-.

In terms of $z = (1 + \exp(2\chi))^{-1}$, Eqs. (45) become

$$2z(1-z)\frac{d\vartheta'}{dz} \cong \vartheta' - \nu\psi, \quad z\frac{d\psi'}{dz} \cong \nu\vartheta' - \psi'. \tag{52}$$

With the same approximation we can write

$$x = -2z\sin\psi' \cong -2z\psi', \quad y = 2z\cos\psi' \cong 2z.$$
$$u = -\sqrt{\mu}(-1 + 1/z)^{1/2}\sin\vartheta',$$
$$\nu = \sqrt{\nu}(-1 + 1/z)^{1/2}\cos\vartheta' \cong \sqrt{\nu}(-1 + 1/z)^{1/2}. \tag{53}$$

The two first-order equations (52) can be combined into

$$z(1-z)\frac{d^2x}{dz^2} - \frac{1}{2}\frac{dx}{dz} + \frac{\nu^2}{2z}x \cong 0, \tag{54}$$

whose solutions are linear combinations of hypergeometric functions

$$x \cong 2A_+ z^{\alpha_+} F(\alpha_+, \alpha_+ - 1;\ 2\alpha_+ - 1/2;\ z)$$
$$+ 2A_- z^{\alpha_-} F(\alpha_-, \alpha_- - 1;\ 2\alpha_- - 1/2;\ z), \tag{55}$$

where $\alpha_+ = 3/4[1 + (1 + 8\nu^2/9)^{1/2}]$ and $\alpha_- = 3/4[1 - (1 - 8\nu^2/9)^{1/2}]$. The corresponding expressions for ϑ' and ψ' are found from

$$\psi' \cong -\frac{1}{2z}x, \quad \vartheta' \cong \frac{1}{\nu}\frac{d(z\psi')}{dz} \cong -\frac{1}{2\nu}\frac{dx}{dz}. \tag{56}$$

If we restrict ourselves to the leading terms in z, we have

$$\begin{pmatrix}\vartheta'\\ \psi'\end{pmatrix} = A_+\begin{pmatrix}-\mu\alpha_+\\ -1\end{pmatrix}z^{\alpha_+ - 1} + A_-\begin{pmatrix}-\mu\alpha_-\\ -1\end{pmatrix}z^{\alpha_- - 1}. \tag{57}$$

All these equations need a little geometric interpretation to make them better understood. According to (53) the trajectory moves along the positive y axis with small deviations which are essentially described by x and u, or ψ' and ϑ'. If $\nu^2 \ll 1$, we find that $\alpha_+ \cong 3/2$ and $\alpha_- \cong \nu^2/3$. Therefore, the first term in (55) shows a strong dependence on z, i.e., y, while the second term changes very little. On the other hand, (57) resolves the motion near the hyperbolic point into the directions $(-3\mu/2, -1)$ and $(-\nu/3, -1)$, i.e., into the horizontal and vertical asymptote. For small z, the last term in (57) dominates, and describes the approach of the trajectory to the vertical asymptote when χ becomes large. In a simplified manner, the A_- term describes the trajectory near the collision, while the A_+ term is responsible for the trajectory away from the collision.

Let us start the trajectory in the autonomous region with a large χ_0 and ϑ_0 near ϑ_v at $\psi_0 = 0$. Since $\vartheta_0 - \vartheta_v$ is small we can use the linearized equations (46) for large χ in order to connect the neighborhood of $\vartheta = \vartheta_v$ at $\psi = 0$ with the neighborhood of the hyperbolic point at $\vartheta = \psi = \pi/2$. The main solution has some A_{-0} and $A_{+0} = 0$, while the linear variation around it is characterized by the two parameters δA_{+0} and δA_{-0}. Suppose that we start integrating Eqs. (43) for large χ, i.e., we drop the χ dependence on the right-hand side, and that we use the initial conditions χ_0, $\vartheta_0 = \vartheta_v$, and $\psi_0 = 0$. Then, we go directly into the hyperbolic point, and we can obtain the parameter A_{-0} from

$$A_{-0} = \lim_{\chi \to -\infty}(\psi - \pi/2)\exp[2(\alpha_- - 1)\chi]. \tag{58}$$

Similarly, we can integrate Eqs. (43) with the initial conditions χ_0, $\delta\vartheta_0 = \vartheta_0 - \vartheta_v$, and $\delta\psi_0 = 0$, and obtain the parameters δA_{+0} and δA_{-0} in the form

$$\delta A_{+0} = \lim_{\chi \to -\infty}[2/(9\mu^2 - 8)^{1/2}](-\delta\vartheta + \mu\alpha_-\delta\psi)\exp[2(\alpha_+ - 1)\chi],$$

$$\delta A_{-0} = \lim_{\chi \to -\infty}[2/(9\mu^2 - 8)^{1/2}](\delta\vartheta - \mu\alpha_+\delta\psi)\exp[2(\alpha_- - 1)\chi]. \tag{59}$$

The last equations follow directly from (57) if we solve for A_+ and A_- in terms of ϑ' and ψ'.

Obviously, δA_{+0} and δA_{-0} depend linearly on $\delta\vartheta_0$. The dependence on χ_0, however, requires a little more thought. In order to find it, let us consider another solution of (43) and (46) for large χ, and choose the initial condition $\chi_0 = 0$. For the sake of definiteness, let us call the corresponding values of A and δA by the names A_{-00}, δA_{+00}, δA_{-00} (of course, $A_{+00} = 0$). Since χ is an additive parameter in the integration of (43) and (46) as long as the χ dependence on the right-hand side is dropped, we find that

$$A_{-0} = A_{-00}\exp[2(\alpha_- - 1)\chi_0]. \tag{60}$$

From the linear relation between δA_{+00}, δA_{-00}, and $\delta\vartheta_0$,

$$\delta A_{+00} = -k_{+0}\delta\vartheta_0, \quad \delta A_{-00} = k_{-0}\delta\vartheta_0, \tag{61}$$

we find that

$$\delta A_{+0} = -k_{+0}\exp[2(\alpha_+ - 1)\chi_0]\,\delta\vartheta_0,$$
$$\delta A_{-0} = k_{-0}\exp[2(\alpha_- - 1)\chi_0]\,\delta\vartheta_0. \tag{62}$$

The values of A_{-00}, k_{+0}, and k_{-0} have to be obtained by numerical integration of (43) and (46) with the calculation of the limits (58) and (59) for the initial value $\chi_0 = 0$. The signs of k_{+0} and k_{-0} in (61) have been chosen to make k_{+0} and k_{-0} positive. A_{-00}, k_{+0}, and k_{-0} are real numbers which depend only on the mass ratio μ/ν of the anisotropic Kepler problem.

The formulas (61) and (62) allow use to compare different initial conditions. Since $\alpha_- - 1 < 0$, the parameter A_{-0} becomes smaller as we increase χ_0, i.e., the trajectory gets closer to the hyperbolic point according to (57) for comparable values of z. If we let $\delta\vartheta_0 \neq 0$, we can make that statement more interesting, because the trajectory does not go through the hyperbolic point. As we increase χ_0, the parameter $A_- = A_{-0} + \delta A_{-0}$ in (57) becomes smaller, whereas the parameter $A_+ = \delta A_{+0}$ stays constant, if we let $\delta\vartheta_0$ go to zero at the same time (because $\alpha_+ - 1 > 0$). Thus, by choosing the right combination of χ_0 and $\vartheta_0 = \vartheta_v + \delta\vartheta_0$ as initial condition, one can find a trajectory arbitrarily close to the hyperbolic point, and yet, choose some given value of the parameter A_+.

In the limit of $A_- = 0$ and $A_+ \neq 0$, one gets an ideal collision, because according to the foregoing, $x_0 = 2\cos\vartheta_v/[1 + \exp(2\chi_0)]$ vanishes as χ_0 goes to infinity. It now becomes of interest to compute the various trajectories for different values of A_+. According to (53), (55), and (56), such a trajectory is given for small z by the equations $x \cong 2A_+ z^{\alpha_+}$, $y \cong 2z$, and similar expressions for u and v in terms of the parameter z. Given a value of $z \neq 0$, one can use the corresponding values of x, y, u, and v as initial conditions for the integration of (2). However, since Eqs. (2) are singular near the origin of the xy plane, one has to find a better representation for x, y, u, and v in terms of z.

The Appendix gives the first few terms of a power series expansion for x, y, u, and v in terms of z. The result is

$$x = Az^{\alpha}(1 + az + \cdots) + Cz^{3\alpha-2}(1 + \cdots) + \cdots$$
$$y = 2z + Bz^{2\alpha-1}(1 + bz + \cdots) + Dz^{4\alpha-3}(1 + \cdots) + \cdots, \tag{63}$$

where

$$a = \frac{\alpha(\alpha-1)}{2\alpha - 1/2}, \quad B = \frac{3A^2}{4(4\alpha-1)(2\alpha-3)},$$
$$b = \frac{(\alpha + 1/2)(\alpha-1)}{2\alpha + 1/2}, \quad C = \frac{3\nu^2}{16(\alpha-1)(8\alpha-7)}A(4B + A^2),$$
$$D = \tfrac{3}{2}(AC - B^2 - BA^2 - \tfrac{5}{32}A^4)/(8\alpha-5)(4\alpha-5), \tag{64}$$

and A has been written instead of A_+, and α instead of α_+. Obviously, A is the only parameter to determine the trajectory: it will be called the collision parameter, henceforth. With the expressions (63) and similar ones for u and v, one has a good set of initial conditions to find the remainder of the trajectory by the numerical integration of the equations of motion (2).

Let us now find A_+ and A_- in terms of the final coordinates χ_1 and ϑ_1 with $\psi = 0$ or π in the neighborhood of a collision trajectory. Given some value of $A_+ = A_{+1}$ we can find χ_{11} and ϑ_{11} by numerical integration according to the above scheme. If we integrate backwards Eqs. (43), starting with χ_{11} and ϑ_{11} we get

$$A_{+1} = \lim_{\chi \to +\infty}(\psi - \pi/2)\exp[2(\alpha_+ - 1)\chi], \tag{65}$$

in exactly the same manner as (59). Moreover, we can integrate backwards Eqs. (46) along the collision trajectory with A_{+1}, using some arbitrary final coordinates $\chi_1 = \chi_{11} + \delta\chi_1$ and $\vartheta_1 = \vartheta_{11} + \delta\vartheta_1$. In this manner we get δA_{+1} and δA_{-1} by exactly the same formulas as (59) except that the limit of χ going to $+\infty$ is now taken instead of the limit χ going to $-\infty$.

As before, the right-hand side in these modified formulas (59) are linear functions of $\delta\chi_1$ and $\delta\vartheta_1$ so that

$$\delta A_{+1} = k_{+1}\delta\chi_1 + k_{+2}\delta\vartheta_1,$$
$$\delta A_{-1} = k_{-1}\delta\chi_1 + k_{-2}\delta\vartheta_1. \tag{66}$$

The coefficients k_{+1}, k_{+2}, k_{-1}, and k_{-2} can be computed numerically, if necessary. They depend on the collision parameter A_{+1} and the mass ratio μ/ν.

One can now put together the various formulas. First, one gets the relation between (X_0, U_0) and (x_0, ϑ_0). Since U_0 is close to $\sqrt{\mu}\pi/2$, we write $U_0 = \sqrt{\mu}\pi/2 - \delta U_0$ with $\delta U_0 > 0$. Similarly, we write $X_0 = 2\cos^2\vartheta_v - \delta X_0$ because $\vartheta_0 = \vartheta_v + \delta\vartheta_0$ and χ_0 is assumed to be large. Therefore,

$$\delta X_0 = 2\sin 2\vartheta_v \cdot \delta\vartheta_0, \quad \delta U_0 = \sqrt{\mu}\exp(-\chi_0)/\cos\vartheta_v. \tag{67}$$

These expressions are inserted into (60) and (62) to give

$$A_- = A_{-0} + \delta A_{-0} = (A_{-00} + k_{-0}\delta\vartheta_0)\exp[2(\alpha_- - 1)\chi_0]$$
$$\cong A_{-00}\left(\frac{\cos\vartheta_v \cdot \delta U_0}{\sqrt{\mu}}\right)^{2(1-\alpha_-)}, \tag{68}$$

where we have neglected $k_{-0}\delta\vartheta_0$ against A_{-00}. Also, we have

$$A_+ = \delta A_{+0} = -\frac{\delta X_0}{2\sin 2\vartheta_v} \cdot k_{+0}\left(\frac{\sqrt{\mu}}{\delta U_0 \cos\vartheta_v}\right)^{2(\alpha_+ - 1)}. \tag{69}$$

Notice that both $1 - \alpha_-$ and $\alpha_+ - 1$ are positive.

The formulas (66) can equally well be expressed in terms of δX_1 and δU_1, i.e., the deviations from the coordinates X_1 and U_1 which belong to the collision trajectory with collision paramter A_{+1}. Therefore, we write instead of (66) the linear relations

$$\delta A_{+1} = l_{+1}\delta X_1 + l_{+2}\delta U_1, \quad \delta A_{-1} = l_{-1}\delta X_1 + l_{-2}\delta U_1, \tag{70}$$

where l_{+1}, l_{+2}, l_{-1}, and l_{-2} depend on A_{+1} and the mass ratio μ/ν.

All these trajectories which are near a collision have U_0 near $\sqrt{\mu}\pi/2$, but U_1 has some value such that $|U_1|$ is bounded away from $\sqrt{\mu}\pi/2$. In the square (25) the points (U_0, U_1) are located along the right-hand side boundary. Our main interest lies in the variation of R in a direction normal to this boundary. Therefore, we can assume that U_1 is fixed and $\delta U_1 = 0$. In the limit of $U_0 \to \sqrt{\mu}\pi/2$, the value of X_0 is given by $2\cos^2\vartheta_v$ independently of U_1, while the value of X_1 depends on U_1, or, equivalently, both X_1 and U_1 depend on the collision parameter A_{+1}.

The second equation (70) can be combined with (68) to give

$$\delta X_1 = \frac{A_{-00}}{l_{-1}} \cdot \left(\frac{\cos\vartheta_v \cdot \delta U_0}{\sqrt{\mu}}\right)^{2(1-\alpha_-)} \tag{71}$$

because $A_- = \delta A_{-1}$. On the other hand, we have $A_+ = A_{+1} + \delta A_{+1}$ where δA_{+1} can be neglected compared to A_{+1}. Therefore, we get from (69) the relation

$$\delta X_0 = -2\sin 2\vartheta_v \frac{A_{+1}}{k_{+0}} \cdot \left(\frac{\cos\vartheta_v \cdot \delta U_0}{\sqrt{\mu}}\right)^{2(\alpha_+ - 1)}. \tag{72}$$

Notice that $2(1-\alpha_-) > 1$, while $2(\alpha_+ - 1) < 1$.

The detailed discussion of the collision trajectories is completed for the case of ϑ_0 near ϑ_v. The sign of A_{+1} determines the sign of x according to (55) and, ultimately, the sign of X_1. Therefore, if $A_{+1} > 0$, we have both $X_0 > 0$ and $X_1 > 0$; and if $A_{+1} < 0$, we have $X_0 > 0$ and $X_1 < 0$. The case $A_{+1} = 0$ is of some interest, because it gives the collision trajectory $x = 0$, $u = 0$ with y going from 0 to 2 and back to 0, and v going from $+\infty$ down to $-\infty$. Therefore, the virial is simply calculated from (29), namely

$$R = \int_{-\infty}^{+\infty} \frac{2\,dv}{1 + v^2/\nu} = 2\pi\sqrt{\nu}. \tag{73}$$

This turns out to be the largest value which R ever attains. The corresponding momentum coordinates are $U_0 = \sqrt{\mu}\pi/2$ and $X_0 = 2\cos^2\vartheta_v$ with either $U_1 = \sqrt{\mu}\pi/2$ and $X_1 = -2\cos^2\vartheta_v$ or $U_1 = -\sqrt{\mu}\pi/2$ and $X_1 = 2\cos^2\vartheta_v$.

C. Trajectories near the x axis

The effect of the hyperbolic points in the vector field (43) has been discussed extensively in the previous subsection. The present subsection will examine the elliptic points. Their neighborhood is described by (44), or, in terms of the variable $z = 1/(1 + e^{2\chi})$, by the equations

$$2z(1-z)\frac{d\vartheta}{dz} = \vartheta - \mu\psi, \quad z\frac{d\psi}{dz} = \mu\vartheta - \psi, \tag{74}$$

which are similar to (52) and could serve as a basis for our further discussion.

There is a difficulty, however, because the parameter z runs from 0 to 1 and back to 0, as the trajectory runs up along the x axis and back down again into the origin. In contrast to z, the momentum u changes monotonically, going from $+\infty$ through 0, when the trajectory reaches the end of the x axis, and down to $-\infty$ when the trajectory goes back into the origin. If one uses u as the parameter, the procedure is equivalent to investigating the geodesic flow (32) in the neighborhood of the geodesic $v = \psi = 0$. It turns out that

$$\zeta = -\frac{u}{(\mu + u^2)^{1/2}} = \pm\sqrt{1-z} \tag{75}$$

is a convenient parameter to use along the geodesic $v = \psi = 0$, and brings the equations into a well-known standard form.

The most straightforward method to obtain the relevant equations is to go back to (2). The first two equations are solved under the assumption that $y = 0$, so that $x = 2/(1 + u^2/\mu) = 2(1 - \zeta^2)$. The second two equations are linearized with respect to y and v, and ζ is used as a parameter instead of t, with $d\zeta/dt = 1/4\sqrt{\mu}(1-\zeta^2)^{1/2}$,

$$\frac{dy}{d\zeta} = 4\mu^{3/2}(1-\zeta^2)^{1/2}v, \quad \frac{dv}{d\zeta} = -\frac{\sqrt{\mu}}{2(1-\zeta^2)^{5/2}}y. \tag{76}$$

The parameter ζ goes from -1 to $+1$. In particular, ζ has a positive rate of increase with time near $\zeta = 0$, when both $u = v = 0$ and the trajectory reaches its point of return at the boundary $x^2 + y^2 = 4$.

The two first-order equations can be written as a single second-order equation of the Legendre type in terms of a function $p(\zeta)$ which is related to y by $y = (1 - \zeta^2)^{3/4}p(\zeta)$,

$$\frac{d}{d\zeta}\left((1-\zeta^2)\frac{dp}{d\zeta}\right) + \left(\frac{3}{4} + \frac{2\mu^2 - 9/4}{1-\zeta^2}\right)p = 0. \tag{77}$$

The solutions are known as Legendre functions $P^{im}_{1/2}(\zeta)$ and $Q^{im}_{1/2}(\zeta)$ of spin-$\frac{1}{2}$ and imaginary magnetic quantum number im where $m = 1/2(8\mu^2 - 9)^{1/2}$. They can be written as

$$\begin{aligned} p(\zeta) &= (1-\zeta^2)^{im/2}F(im/2 + \tfrac{3}{4},\ im/2 - \tfrac{1}{4};\ \tfrac{1}{2};\ \zeta^2), \\ q(\zeta) &= \zeta(1-\zeta^2)^{im/2}F(im/2 + \tfrac{5}{4},\ im/2 + \tfrac{1}{4};\ \tfrac{3}{2};\ \zeta^2), \end{aligned} \tag{78}$$

in terms of hypergeometric functions.

In spite of their appearance, these two functions have real values, as can be ascertained most easily from the transformation formulas for hypergeometric functions (Whittaker–Watson, p. 201)

$$\begin{aligned} p(\zeta) &= \frac{\Gamma(\frac{1}{2})\Gamma(-im)}{\Gamma(\frac{5}{4} - im/2)\Gamma(\frac{1}{4} - im/2)} z^{im/2}F(\tfrac{3}{4} + im/2, \\ &\quad -\tfrac{1}{4} + im/2;\ 1 + im;\ z) + \text{c.c.}, \\ q(\zeta) &= \frac{\Gamma(\frac{3}{2})\Gamma(-im)}{\Gamma(\frac{3}{4} - im/2)\Gamma(-\frac{1}{4} - im/2)} z^{im/2}F(\tfrac{5}{4} + im/2, \\ &\quad +\tfrac{1}{4} + im/2;\ 1 + im;\ z) + \text{c.c.} \end{aligned} \tag{79}$$

For small values of z, i.e., near the origin, only the terms $z^{im/2}$ and $z^{-im/2}$ are important, and we find the spiral of Fig. 1.

The explicit formulas for y in terms of ζ and z show that these trajectories stay near the x axis. However, v does not remain bounded when ζ goes to ± 1. This is consistent with the vector field in the autonomous region, because when $|\zeta|$ goes to 1, i.e., $z \to 0$ and $\chi \to \infty$, the trajectory comes out of the elliptic point.

Let us consider in detail the spiral and the trajectories in its neighborhood. The leading terms in (79) yield

$$y \cong Az^{3/4 + im/2} + \text{c.c.} \cong A\exp[-(3/2 + im)\chi] + \text{c.c.}, \tag{80}$$

where A is a complex constant. Two consecutive zeros, corresponding to χ_0 and χ_1, have to satisfy the condition $m(\chi_0 - \chi_1) = \pi$ provided both χ_0 and χ_1 are large. Since we are looking at a trajectory which leaves the origin, we assume that $\chi_0 > \chi_1$. Moreover, $\vartheta \approx 0$ and both U_0 and U_1 are near $\sqrt{\mu}\pi/2$. In view of (47), we can write

$$(\pi/2\sqrt{\mu} - U_1)/(\pi/2\sqrt{\mu} - U_0) \cong \exp(\chi_0 - \chi_1) = \exp(\pi/m), \tag{81}$$

which is the lower limit of the ratio (50) in Sec. III A.

As χ decreases, the consecutive zeros in y have to be found by inserting the full expansions (79) or (78) into (80). Again, since we are studying a trajectory which

leaves the origin, goes up along the x axis to $x=2$, and comes back down again to the origin, the value of ξ_1 is always larger than ξ_0. Equivalently, we have $U_1 < U_0$ because of (75) and (79). Given U_0 in the interval $(-\sqrt{\mu}\pi/2, +\sqrt{\mu}\pi/2)$, there is exactly one value of U_1 corresponding to the first consecutive zero ξ_1 of a function y like (80) with a zero at ξ_0 corresponding to U_0. Therefore, we have a function $U_1 = g(U_0) < U_0$ whose final slope near $\sqrt{\mu}\pi/2$ is given by (81).

From our construction both X_0 and X_1 are positive. But, more specifically, since $x = 2/(1+u^2/\mu)$ along any of these trajectories (e.g., simply check the conservation of energy when $y = v = 0$), we have

$$X_0 = X_1 = 2. \tag{82}$$

The consecutive points along the x axis where $y=0$ can also be interpreted as conjugate points along the trajectory $v=0$ in momentum space.

IV. THE EXISTENCE OF TRAJECTORIES WITH A GIVEN BINARY SEQUENCE

A. Arcs in momentum space

The virial was introduced in Sec. II A as the principal quantity to be used to establish the existence of trajectories with a given binary sequence such as described in Sec. I C. The viral R has the momenta U_0 and U_1 as its natural, independent variables, and according to (23) we have the relations

$$\frac{\partial R}{\partial U_0} = X_0, \quad \frac{\partial R}{\partial U_1} = -X_1. \tag{83}$$

The pairs (X_0, U_0) and (X_1, U_1) correspond to consecutive intersections of the trajectory with the x axis. These are four different functions R according as the signs of X_0 and X_1, and we shall refer to these four combinations with the names a through b as defined in (24).

The detailed shape of the function $R(U_0, U_1)$ was found by working out the three limiting cases of Sec. III and by making numerical computations. It should be emphasized that, apart from these limits which are the boundaries for the domains of U_0 and U_1, the corresponding trajectories have perfectly smooth behavior. Away from the boundaries, the function R and its derivatives are smooth in terms of U_0 and U_1 because they are obtained from integrating a well behaved system of ordinary differential equations, i.e., (2), between two points which are a finite time interval away from each other. The inequality (26) is verified by taking a sufficiently dense grid of values U_0 and U_1, in addition to the results of Sec. III.

It may be helpful to visualize the trajectory from (X_0, U_0) to (X_1, U_1) as a curve in the three-dimensional space of u, v, and ψ which was used in Sec. II B to describe the geodesic flow. All these curves go from ψ_0 = some multiple of π to ψ_1 = some other (or same) multiple of π, because y vanishes at both endpoints. If these curves are projected into the (u, v) plane, we get the hodograph. Therefore, we shall refer to them as arcs in momentum space.

Such an arc can also be projected into the (u, ψ) plane. Its starting momentum u_0 and its final momentum u_1 are then directly related to U_0 and U_1 through (7). The starting and final values, ψ_0 and ψ_1, are multiples of π (modulo 2π). The arc goes through ψ = multiple of π only at its endpoints. If we rewrite the first and last equations (32) for $\sin\psi = 0$, as

$$\frac{d\psi}{du} = -\frac{2v}{\nu(1+u^2/\mu+v^2/\nu)}, \tag{84}$$

we see that the direction of the arc in the (u, ψ) projection at its endpoints gives the values of the momentum v. Therefore, we can reconstruct the arc in (u, v, ψ) space from its projection into the (u, ψ) plane. Since ψ matters only modulo 2π, one can just as well think of a (u, ψ) cylinder, rather than a plane.

The domains of variation for U_0 and U_1 can now be described in detail. For the four combinations of signs of X_0 and X_1, we have the following domains:

$$x_0 > 0, x_1 > 0: -\sqrt{\mu}\pi/2 < U_0 < \sqrt{\mu}\pi/2, \; -\sqrt{\mu}\pi/2 < U_1 < g(U_0), \tag{85a}$$

$$x_0 < 0, \; x_1 > 0: -\sqrt{\mu}\pi/2 < U_0, \; U_1 < \sqrt{\mu}\pi/2, \tag{85b}$$

$$x_0 < 0, \; x_1 < 0: -\sqrt{\mu}\pi/2 < U_1 < \sqrt{\mu}\pi/2, -\sqrt{\mu}\pi/2 < U_0 < g(U_1), \tag{85c}$$

$$x_0 > 0, \; x_1 < 0: -\sqrt{\mu}\pi/2 < U_0, \; U_1 < \sqrt{\mu}\pi/2, \tag{85d}$$

where $g(U_0)$ has been defined at the end of the last section.

These four domains can be represented in a single diagram in the manner shown in Fig. 2. There are four quadrants which correspond to the four combinations of signs for x_0 and x_1. In each quadrant there is a square for the domain $-\sqrt{\mu}\pi/2 < U_0, \; U_1 < \sqrt{\mu}\pi/2$. In the case of (i) and (iii) the squares are further subdivided, because the two momenta U_0 and U_1 have a corresponding arc only when U_0 and U_1 satisfy certain inequalities involving the function $g(U)$.

The axes in each of the (U_0, U_1) squares of Fig. 2 have been chosen so as to bring out the relations with the three limiting cases of Sec. III. The trajectories near the origin of the xy plane (Sec. III A) are first in the corner $U_0 = U_1 = \sqrt{\mu}\pi/2$ of (i), then along the boundary

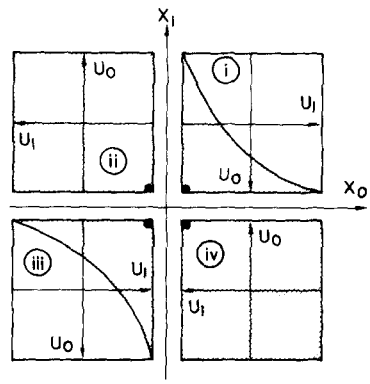


FIG. 2. The four domains of the function $R(U_0, U_1)$ which correspond to the four combinations of sign X_1 and sign X_1. The axes in each domain have been chosen such that the maxima of R are in the center of the figure, as indicated by the heavy dots.

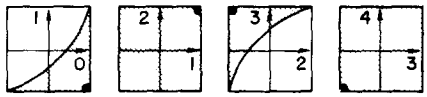


FIG. 3. Sequence of domains for R corresponding to the sequence of sign X given by $++--+$.

$U_1 = -\sqrt{\mu}\pi/2$ followed by the boundary $U_0 = -\sqrt{\mu}\pi/2$ of (iv), and finally in the corner $U_0 = U_1 = -\sqrt{\mu}\pi/2$ of (i) again. Similar trajectories near the origin are located in the corners $U_0 = U_1 = \sqrt{\mu}\pi/2$ and $U_0 = U_1 = -\sqrt{\mu}\pi/2$ or (iii), and along the boundaries $U_1 = \sqrt{\mu}\pi/2$ and $U_0 = +\sqrt{\mu}\pi/2$ or (ii).

The collision trajectories are all to be found along the "inner" boundaries. The special cases which were discussed in Sec. III B, are all found along the boundary $U_0 = \sqrt{\mu}\pi/2$ of (i) and (iv). The collision parameter A_{+1} is positive in (i) and negative in (iv), arcs with the same $|A_+|$ are facing each other across the x_0 axis in Fig. 2. The arc with the largest virial, $A_{+1} = 0$, is near the origin of the x_0 and x_1 axes. Instead of the collision parameter A_{+1}, one can just as well use the final momentum U_1 to distinguish one collision arc from another, provided the sign of x_1 is also given. The remaining arcs with a collision at either end are found along the other "inner" boundaries, like $U_1 = -\sqrt{\mu}\pi/2$ for (i) and (ii), etc. The maximum value of the virial is always near the origin of the x_0 and x_1 axes, while the virial vanishes at the outer end.

The arcs near the x axis (Sec. III C) are found along the curved portions of the boundary (i) and (iii). For (i) these arcs lie along the positive x axis, and for (iii) along the negative x axis. The value of the virial reaches a maximum in the middle of the curved boundaries and vanishes at the ends. Also, the virial vanishes along the outer boundaries in (ii) and (iv) where the corresponding arcs are near the origin of the (x,y) plane.

Each point in one of the shaded domains in Fig. 2 gives rise to exactly one arc if we further specify that $v_0 > 0$. This is an essential assertion and the whole argument for the existence of more complex trajectories depends on this proposition. As pointed out before, the proof for its correctness is given through the discussion of the various boundaries in Fig. 2, and through numerical computation of the "interior." Although the latter is not always accepted as conclusive by all mathematicians, it is believed to be valid in this case for two reasons. First, the ordinary differential equations are well behaved for U_0 and U_1 in the "interior." One can easily establish criteria for the density of the grid which is appropriate if a particular numerical procedure is used to integrate the differential equations. Second, the integration extends over a finite range of the independent variable. If Eqs. (32) are used, this range has the upper limit given by (73).

B. Composite arcs and their domains

The simple arcs in momentum space of the preceding section can be put together to form composite arcs in the following manner. First, pick a finite sequence of binaries $a_0, a_1, \ldots, a_n$ (where $a_i = \pm 1$). Second, pick a finite sequence of momenta $U_0, U_1, \ldots, U_n$. Finally,

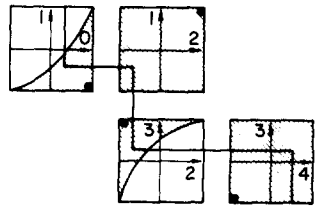


FIG. 4. The same sequence of domains for R as in Fig. 3 is now arranged so that corresponding axes in different domains are parallel. A particular sequence of momenta, U_0 through U_4, is represented by a sequence of intersecting, mutually perpendicular lines.

consider the sequence of simple arcs (U_0, U_1), $(U_1, U_2), \ldots, (U_{n-1}, U_n)$, where, in addition, the signs of the position coordinates are restricted by the condition sign $(X_i) = a_i$.

The last restriction imposes limitations on the choice of the momenta $U_0, \ldots, U_n$. If $a_i \neq a_{i+1}$, any combination of momenta U_i and U_{i+1} within the interval from $-\sqrt{\mu}\pi/2$ to $+\sqrt{\mu}\pi/2$ is acceptable according to the preceding subsection. However, if $a_i = a_{i+1}$, the values of U_i and U_{i+1} have to be subjected to the inequalities (85a) and (85c). Therefore, if the sequence of binaries $a_0, a_1, \ldots, a_n$ is given, the momenta $U_0, U_1, \ldots, U_n$ have to be restricted to certain nontrivial domains in the $(n+1)$-dimensional Euclidean space. These domains result from (85) in a straightforward manner.

It may be of some help to consider first an example. Let $n = 4$, and take $a_0 = +1$, $a_1 = +1$, $a_2 = -1, a_3 = -1$, $a_4 = +1$. The relevant four domains are shown in Fig. 3. The heavy dot in one of the corners of each square gives the location of the maximum value for the virial R in the particular domain.

In order to obtain a better picture of the restrictions which are inherent in the domains of Fig. 3, it is advantageous to rearrange the four squares and the directions of the various U axes in them. Coordinate axes in different diagrams are made parallel if they belong to the same variable. The sequence of momenta $U_0, U_1, \ldots, U_4$ is indicated by a sequence of alternating vertical and horizontal segments, as in Fig. 4.

The generalization of this scheme is quite obvious. Also, one finds immediately the following rule. Any restrictions among the momenta U_i occur only if they belong to the same string of identical binaries. To be more precise, consider a sequence $a_{i-1}, a_i, a_{i+1}, \ldots, a_{i+j-1}, a_{i+j}$ which is part of a longer sequence $a_0, a_1, \ldots, a_n$, and where $a_{i-1} \neq a_i = a_{i+1} = \cdots = a_{i+j-1} \neq a_{i+j}$. To be definite, let us assume that $a_{i-1} = a_{i+j} = -1$, where $a_i = a_{i+1} = \cdots = a_{i+j-1} = +1$. If $j = 1$, the arc (U_{i-1}, U_i) belongs to the case (85d). There is no restriction on the values of U_i whatever the values of U_{i-1} and U_{i+1} may

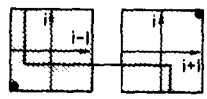


FIG. 5. Two consecutive domains, corresponding to $R(U_{i-1}, U_i)$ and $R(U_i, U_{i+1})$, show that the intermediate momentum U_i can be freely varied from $-\mu\pi/2$ to $+\mu\pi/2$ while U_{i-1} and U_{i+1} stay fixed.

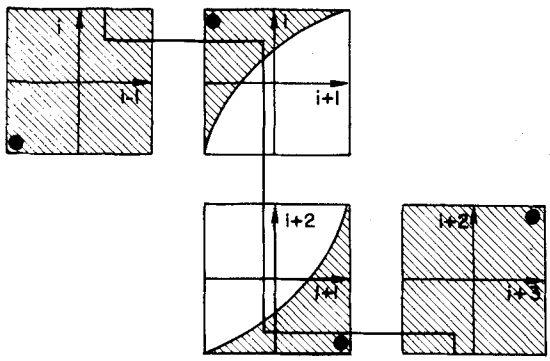


FIG. 6. Four consecutive domains show that the intermediate momenta U_i, U_{i+1}, and U_{i+2} have their common domain of variation independently of the values of U_{i-1} and U_{i+3} provided that the signs of X_{i-1} and X_{i+3} are both different from the common signs of X_i, X_{i+1}, and X_{i+2}.

be. Therefore, if $j=1$, the common domain for the momenta becomes a simple Cartesian product. Whatever applies to the sequences $U_0, \ldots, U_{i-1}$ and $U_{i+1}, \ldots, U_n$ the momentum U_i provides a factor $-\sqrt{\mu}\pi/2 < U_i < \sqrt{\mu}\pi/2$ to the product. This situation is shown in Fig. 5.

Consider now the case where $j>1$. Again, let us look at a particular example, e.g., $j=3$ which is shown in Fig. 6. The domain of variation for $U_i, U_{i+1}, \ldots, U_{i+j-1}$ is not restricted by the values of either U_{i-1} or U_{i+j} whatever they may be. However, there are the restrictions

$$\begin{aligned}
&-\sqrt{\mu}\pi/2 < U_i < \sqrt{\mu}\pi/2,\\
&-\sqrt{\mu}\pi/2 < U_{i+1} < g(U_i) < U_i,\\
&-\sqrt{\mu}\pi/2 < U_{i+2} < g(U_{i+1}) < U_{i+1},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&-\sqrt{\mu}\pi/2 < U_{i+j-1} < g(U_{i+j-2}) < U_{i+j-2},
\end{aligned} \tag{86}$$

which follow directly from (85a). The domain (86) has a somewhat contorted shape inside a j-dimensional cube. This domain will be called W_j henceforth.

W_1 is just the interval $-\sqrt{\mu}\pi/2 < U < \sqrt{\mu}\pi/2$. The shape of W_2 is the triangular region in the first square of Fig. 2. The domain W_3 can be visualized from Fig. 6 if the third square is imagined to sit perpendicularly to the second with the $i+1$ axis in common. In general, a string of j identical binaries contributes a factor W_j to the domain of the momenta $U_0, \ldots, U_n$.

The domain of the sequence $(U_0, \ldots, U_n)$ is a Cartesian product of domains W_j, one for each string of consecutive identical binaries in the sequence $(a_0, \ldots, a_n)$.

C. A trajectory for each Bernoulli sequence

The composite arcs of the preceding subsection can be interpreted as continuous curves in the (u, ψ) projection which was discussed in Sec. IV A. In order to make this interpretation unambiguous, it will be further required that the momentum v be positive at the even-numbered endpoints of the simple arcs, i.e., U_0, U_2, etc., and that v be negative at the odd-numbered endpoints, i.e., U_1, U_3, etc. Such a condition is natural if one thinks of a composite arc in the (x, y) plane where each simple arc is a trajectory between two consecutive intersections with the x axis, and the composite arc alternates between the upper and the lower half-plane.

The composite arc describes now a well-defined sequence of trajectories in the (u, v, ψ) space of the geodesic flow. The projection of this sequence into the (u, ψ) plane is continuous. Its tangent turns continuously except, possibly, at the endpoints of the simple arcs. A virial R can be defined for this composite arc by adding up the virials of all its component simple arcs. The total virial R depends on both the assumed sequence of binaries $(a_0, a_1, \ldots, a_n)$ and the sequence of momenta $(U_0, U_1, \ldots, U_n)$. In terms of the functions $R(U_0, U_1)$ in Sec. II A we have

$$R = R(U_0, U_1) + R(U_1, U_2) + \cdots + R(U_{n-1}, U_n), \tag{87}$$

where the indices $(a_0, a_1), (a_1, a_2), \ldots, (a_{n-1}, a_n)$ on the various R in the sum have been omitted for simplicity's sake.

If the composite arc is a real trajectory rather than an artificial sequence of simple arcs, the position coordinates X at the endpoints of consecutive simple arcs have to agree with each other. In view of the basic relation (23) for the virial R, one can write the condition for a continuous trajectory rather than a composite arc as

$$\frac{\partial R}{\partial U_1} = 0, \quad \frac{\partial R}{\partial U_2} = 0, \quad \cdots, \quad \frac{\partial R}{\partial U_{n-1}} = 0. \tag{88}$$

Thus, if R is considered as a function of the momenta $U_0, \ldots, U_n$ in one of the appropriate domains of the preceding subsection, then the continuous trajectories are realized at an extremum of R with respect to $U_1, \ldots, U_{n-1}$ at fixed values of U_0 and U_n.

If the binary sequence $(a_0, \ldots, a_n)$ has at least one sign change, the domain for the momenta $U_0, \ldots, U_n$ becomes a Cartesian product with U_0 and U_n belonging to different factors. Keeping U_0 and U_n fixed modifies the corresponding factor domains W of (86) by setting a fixed upper (lower) bound to the inequalities, but this does not restrict the possible values of U_0 and U_n. If the binary sequence $(a_0, \ldots, a_n)$ has all identical signs, say $a_0 = \cdots = a_n = +1$, there is both an upper bound U_0 and a lower bound U_n in the sequence of inequalities (86). These two bounds have to be compatible with the inequalities, i.e., with the repeated occurrence of the function $g(U)$.

A particularly interesting case arises when the composite arc is closed, i.e., if we require that $a_0 = a_n$ and $U_0 = U_n$. Because of our interpretation in terms of a trajectory which alternates between the upper and the lower half of the (xy) plane, it is now necessary to make n an even number. From the discussion in the previous paragraph it appears that there have to be sign changes in the binary sequence. It is not possible to have a closed composite arc with all its intersections only in the positive x axis. *A fortiori*, there cannot be any closed continuous trajectory of this kind. However, if there are sign changes and $U_0 = U_n$ is considered as one independent variable, the domain of variation for $U_0, U_1, \ldots, U_{n-1}$ is again a Cartesian product of domains

W. The condition for the occurrence of a continuous closed trajectory, usually called a closed orbit, now becomes

$$\frac{\partial \mathcal{R}}{\partial U_1} = \frac{\partial \mathcal{R}}{\partial U_2} = \cdots = \frac{\partial \mathcal{R}}{\partial U_n} = 0, \tag{89}$$

where $\mathcal{R}$ is the function (87) with $U_0 = U_n$. In other words, a closed orbit is realized at an extremum of $\mathcal{R}$, and the function $\mathcal{R}$ can be defined for any arbitrary binary sequence of even length which has at least one sign change.

In order to show the existence of an extremum for the function $\mathcal{R}$ of a composite (closed) arc, we show that $\mathcal{R}$ has at least one maximum in its domain, i. e., in the appropriate Cartesian product of domains W_j. This claim is based in turn on two simple facts. First, $\mathcal{R}$ is continuous and continuously differentiable in the interior of its domain. Second, the derivatives of $\mathcal{R}$ along the boundaries of its domain have always a positive component pointing toward the interior. We conclude that $\mathcal{R}$ has a maximum in the interior, although it has not been possible to find an explicit statement of this theorem, nor will an attempt be made to provide a proof here.

The main task is to show that there is a direction with a positive derivative into the interior at each boundary point. This is accomplished by taking some arbitrary momentum U_i in $\mathcal{R}$, and letting it vary freely while keeping the other momenta U_l with $l \neq i$ fixed. The momentum U_i belongs to one of the domains W_j of the preceding subsection, and we shall investigate the various possible cases for j which can occur.

If U_i belongs to a domain W_1, then sign $x_{i-1} \neq$ sign $x_i \neq$ sign x_{i+1}, and U_i can vary freely in the interval from $-\sqrt{\mu}\pi/2$ to $+\sqrt{\mu}\pi/2$. The situation is represented in Fig. 5, where the vertical lines are fixed by the assumed values of U_{i-1} and U_{i+1}, but the horizontal line corresponds to U_i and its location varies between the two extremes of $-\sqrt{\mu}\pi/2$ and $+\sqrt{\mu}\pi/2$. Let us choose sign $x_i = +1$ for definiteness. If U_i approaches $-\sqrt{\mu}\pi/2$, the arc (U_{i-1}, U_i) has a collision at U_i and the arc (U_i, U_{i+1}) lies entirely in the autonomous region. (A collision always occurs at a boundary which contains the heavy dot.) We can now use formula (23) together with the results of Secs. III A and B, and we find that at $U_i = -\sqrt{\mu}\pi/2$ we have

$$\frac{\partial R(U_{i-1}, U_i)}{\partial U_i} = -2\cos^2\vartheta_v, \quad \frac{\partial R(U_i, U_{i+1})}{\partial U_i} = +2\cos^2\vartheta_c. \tag{90}$$

Therefore, we find that

$$\frac{\partial \mathcal{R}}{\partial U_i} = \frac{\partial R(U_{i-1}, U_i)}{\partial U_i} + \frac{\partial R(U_i, U_{i+1})}{\partial U_i} > 0 \tag{91}$$

at $U_i = -\sqrt{\mu}\pi/2$ because $0 < \vartheta_c < \vartheta_v < \pi/2$. An exactly similar argument applies near $U_i = +\sqrt{\mu}\pi/2$ to show that $\partial R/\partial U_i < 0$.

Let us now look at a case where U_i belongs to a domain W_j with $j > 1$. The corresponding situation is shown in Fig. 6, and part of the relevant discussion has already taken place there. If U_i is varied freely with U_{i-1} and U_{i+1} fixed, its lower limit comes from the requirement in (86) that $U_{i+1} < g(U_i)$. This corresponds to an arc along the x axis as discussed in Sec. III C. On the other hand, the arc (U_{i-1}, U_i) is essentially arbitrary; therefore, we now find that

$$-2 < \frac{\partial R(U_{i-1}, U_i)}{\partial U_i} < 0, \quad \frac{\partial R(U_i, U_{i+1})}{\partial U_i} = 2. \tag{92}$$

Again $\partial \mathcal{R}/\partial U_i > 0$ at the lower boundary for U_i at constant U_{i-1} and U_{i+1}. On the other hand, the upper boundary of U_i is $+\sqrt{\mu}\pi/2$, where the arc (U_{i-1}, U_i) is entirely in the autonomous region and the arc (U_i, U_{i+1}) has a collision at U_i. The argument now is exactly similar to the one in the preceding paragraph, with the result that $\partial R/\partial U_i < 0$ at the upper end of the permissible interval for U_i.

A slightly different situation arises if we look at U_{i+1} in Fig. 6. Now the horizontal lines are fixed at the value U_i and U_{i+2}, and the location of the vertical line at U_{i+1} is allowed to vary. Its interval is now limited by two of the inequalities (86), namely $U_{i+1} < g(U_i)$ and $U_{i+2} < g(U_{i+1})$. In both cases one of the limiting arcs lies along the x axis, while the other is essentially arbitrary. Thus, we have again relations similar to (92) at the lower end of the interval for U_{i+1} and corresponding results at the upper end.

This type of reasoning always applies when a boundary point of the domain for a composite arc is approached by fixing all the momenta except one, say U_i. The derivative $\partial \mathcal{R}/\partial U_i$ is then positive at the lower end and negative at the upper end of the permissible interval for U_i. All the generic boundary points can be reached in this manner. However, there are boundary points which can only be approached by allowing two or more momenta simultaneously to converge to some limiting values in a judiciously chosen manner. In that case, more than two of the arcs become exceptional and the inequality (91) has now more than two terms on the left. There does not seem to be anything new in these exceptional cases, only a proliferation of different combinations of circumstances.

In summary, the existence of both a trajectory with fixed end momenta U_0 and U_n as well as a closed orbit, corresponding to any given binary sequence, has been demonstrated.

D. Relations between initial position and final momentum

The argument for the existence of trajectories with given end momenta and of closed orbits was based on the existence of a maximum for the virial R of a composite arc. Let us now discuss the consequences for the trajectories if they are due to a maximum (rather than some other extremum). Before doing so a number of definitions have to be given.

The virial for a simple arc depends on an initial momentum U' and a final momentum U'' through the function $R(U', U'')$ which was defined in Sec. II A. We associate an initial position X' and a final position X'' with a simple arc. These positions are functions of U' and U'',

$$X'(U', U'') = \frac{\partial R}{\partial U'}, \quad X''(U', U'') = -\frac{\partial R}{\partial U''}. \tag{93}$$

If we consider a composite arc given by the momenta $U_0, U_1, \ldots, U_m$, we shall assume that the momenta $U_1, \ldots, U_{m-1}$ have already been adjusted so as to give a continuous trajectory. Therefore, we write its virial $\mathcal{R}_m$ as a function of the two momenta U_0 and U_m, and we have the relations

$$\mathcal{R}_m(U_0, U_m) = R(U_0, U_1) + U(U_1, U_2) + \cdots + R(U_{m-1}, U_m), \tag{94}$$

where the momenta $U_1, \ldots, U_m$ satisfy the equations

$$\frac{\partial R(U_{i-1}, U_i)}{\partial U_i} + \frac{\partial R(U_i, U_{i+1})}{\partial U_i} = 0 \quad \text{for } 0 < i < m. \tag{95}$$

Further, we use the notation $R_m(U_{m-1}, U_m) = R(U_{m-1}, U_m)$ so that $\mathcal{R}_m = \mathcal{R}_{m-1} + R_m$ provided $\partial \mathcal{R}_{m-1}/\partial U_{m-1} = -\partial R_m/\partial U_{m-1}$. Of course, $\mathcal{R}_1 = R_1$. In order to simplify the writing, we shall indicate derivatives by attaching indices to $\mathcal{R}$ and R. It is understood that the binary sequence $a_0, a_1, \ldots, a_m, \cdots$ is fixed in all these arguments.

The initial position X_0 and the final position X_m are given by the formula

$$X_0 = \frac{\partial \mathcal{R}_m}{\partial U_0} = \mathcal{R}_{m;0}, \quad X_m = -\frac{\partial \mathcal{R}_m}{\partial U_m} = -\mathcal{R}_{m;m}. \tag{96}$$

Because of $\mathcal{R}_m = R_m + \mathcal{R}_{m-1}$, one can also write $X_m = -R_{m;m}$. If U_0 is held fixed and U_m is allowed to vary, all the intermediate momenta $U_1, \ldots, U_{m-1}$ vary also because they are coupled through (95). This shows up explicitly if we take the derivatives of X_m with respect to U_m and U_0. In particular, since $R_{m;m-1} + \mathcal{R}_{m-1;m-1} = 0$ as a consequence of (95), we have

$$\begin{aligned} &\delta X_m = -R_{m;m,m}\delta U_m - R_{m;m,m-1}\delta U_{m-1}, \\ &R_{m;m-1,m}\delta U_m + (R_{m;m-1,m-1} + \mathcal{R}_{m-1;m-1,m-1})\delta U_{m-1} \\ &\quad + \mathcal{R}_{m-1;m-1,0}\delta U_0 = 0. \end{aligned} \tag{97}$$

If we eliminate δU_{m-1} between these two equations, and use the notation

$$r_{m-1} = R_{m;m-1,m-1} + \mathcal{R}_{m-1;m-1,m-1}, \tag{98}$$

then we can write the derivatives of X_m in the form

$$\begin{aligned} X_{m;m} &= -\mathcal{R}_{m;m,m} = -R_{m;m,m} + (R_{m;m,m-1})^2/r_{m-1}, \\ X_{m;0} &= -\mathcal{R}_{m;m,0} = R_{m;m,m-1} \cdot \mathcal{R}_{m-1;m-1,0}/r_{m-1}. \end{aligned} \tag{99}$$

The quantity r_m obviously satisfies the recurrence relation

$$r_m = R_{m+1;m,m} + R_{m;m,m} - (R_{m;m,m-1})^2/r_{m-1}, \tag{100}$$

with the starting value $r_1 = R_{2;1,1} + R_{1;1,1}$ because $\mathcal{R}_1 = R_1$.

Equation (99) expresses $\mathcal{R}_{m;m,0}$ in terms of $\mathcal{R}_{m-1;m-1,0}$. One can immediately replace the latter by $\mathcal{R}_{m-2;m-2,0}$, etc. In this way one finds the formula

$$X_{m;0} = (-1)^m \cdot \prod_{i=1}^{m} R_{i;i,i-1} \Big/ \prod_{j=1}^{m-1} r_j, \tag{101}$$

which involves only the virials R_i for the simple arcs from U_{i-1} to U_i.

The last expression can be interpreted in terms of the second variations of the virial for the trajectory from U_0 to U_m. The virial for the composite arc (87) or (94) has the second variation in terms of $U_1, U_2, \ldots, U_{m-1}$ given by

$$\begin{aligned} \delta^2 \mathcal{R}_m = {} & R_{1;1,1}\delta U_1^2 + R_{2;2,2}\delta U_2^2 + 2R_{2;2,1}\delta U_2 \delta U_1 + R_{2;1,1}\delta U_1^2 \\ & + \cdots + R_{m-1;m-1,m-1}\delta U_{m-1}^2 + 2R_{m-1;m-1,m-2}\delta U_{m-1}\delta U_{m-2} \\ & + R_{m-1;m-2,m-2}\delta U_{m-2}^2 + R_{m;m-1,m-1}\delta U_{m-1}^2. \end{aligned} \tag{102}$$

This quadratic form in $\delta U_1, \ldots, \delta U_{m-1}$ has the diagonal elements $R_{i;i,i} + R_{i+1;i,i}$ for $i = 1, \ldots, m-1$ and the off-diagonal elements $R_{i+1;i+1,i}$ in the positions $(i, i+1)$ and $(i+1, i)$. The matrix elements in the position (i,j) where $j < i-1$ or $j > i+1$ all vanish, i.e., the matrix is symmetric and tridiagonal. The latter is a consequence of our definitions (87) and (94) for the virial.

Now let p_i be the determinant of the first i rows and columns. Then we have the well-known recursion relation

$$p_i = (R_{i;i,i} + R_{i-1;i,i})p_{i-1} - (R_{i;i,i-1})^2 p_{i-2}, \tag{103}$$

which starts with $p_1 = R_{2;1,1} + R_{1;1,1}$. Moreover, one gets the correct expression for p_2 if the recursion is applied already for $i = 2$ provided one defines $p_0 = 1$. In terms of the quotient $q_i = p_i/p_{i-1}$ one can write (101) equally well as

$$q_i = R_{i+1;i,i} + R_{i;i,i} - (R_{i;i,i-1})^2/q_{i-1}, \tag{104}$$

where the initial $q_1 = R_{2;1,1} + R_{1;1,1}$. A comparison with (100) shows that $r_i = q_i$, and the formula (101) becomes, therefore,

$$X_{m;0} = \frac{(-1)^m}{p_{m-1}} \cdot \prod_{i}^{m} R_{i;i,i-1} = -X_{0,m}. \tag{105}$$

The last part of this equality arises directly from (96) which shows that $X_{0;m} = \mathcal{R}_{m;0,m} = \mathcal{R}_{m;m,0} = -X_{m,0}$.

If the trajectory from U_0 to U_m is connected with a maximum of the virial, the second variation (102) is negative definite. A well-known theorem of matrix theory states that the determinants p_i alternate in sign as a function of i, starting with $p_0 = 1$. The quantity $(-1)^{m-1}p_{m-1}$ is, therefore, positive. Moreover, it was shown in Sec. IIA on the basis of extensive computer experiments that the mixed derivative $\partial^2 R/\partial U' \partial U''$ is always positive, as stated in Eq. (26). Thus, we find that $\partial X_0/\partial U_m$ is also positive provided the trajectory arises from a maximum in the virial (87).

This is a generalization of the inequality (27) to a trajectory of any length. It is also a generalization of a statement which was made in the preceding paper where $\partial U_m/\partial X_0$ at constant U_0 was claimed to be positive under the additional assumption $U_0 = 0$. The importance of the inequality (27) was emphasized in Sec. IIA because it allowed the determination of an arc if its initial and final momenta, rather than its initial momentum and position, are known. The same is true on the basis of (105) for any trajectory provided its binary sequence is given in addition to the initial and final momenta, and provided the trajectory maximizes the virial (87).

APPENDIX

Formulas (63) and (64) will be derived in this appendix. For this purpose, we start with the equations

of motion (2) and we insert the expressions $x = x_0 + \delta x$, $y = y_0 + \delta y$, $u = u_0 + \delta u$, $v = v_0 + \delta v$. Immediately, we specialize to $x_0 = u_0 = 0$ with $y_0 > 0$, and we integrate Eqs. (2) with $\delta x = \delta y = \delta u = \delta v = 0$. From the conservation of energy, it follows that

$$t = \int_0^{y_0} d\eta \left(\frac{\nu \cdot \eta}{2 - \eta}\right)^{1/2}, \tag{A1}$$

provided $y_0 = 0$ for $t = 0$.

It turns out to be convenient to use $z = y_0/2$ rather than t as the independent variable. If the time derivative is indicated by a dot, and the derivative w.r.t. z is indicated by a prime, we can write

$$\begin{aligned}
\delta\dot{x} &= \frac{dz}{dt}\delta x' = \frac{1}{2}\left(\frac{1-z}{\nu z}\right)^{1/2}\delta x' = \frac{\delta u}{\mu}, \\
\delta\dot{u} &= \frac{dz}{dt}\delta u' = \frac{1}{2}\left(\frac{1-z}{\nu z}\right)^{1/2}\delta u' = -\frac{\delta x}{[\delta x^2 + (2z + \delta y)^2]^{3/2}}, \\
&\frac{1}{2}\left(\frac{1-z}{\nu z}\right)^{1/2}\delta y' = \frac{\delta v}{\nu}, \\
&\frac{1}{2}\left(\frac{1-z}{\nu z}\right)^{1/2}\delta v' = -\frac{2z + \delta y}{[\delta x^2 + (2z + \delta y)^2]^{3/2}} + \frac{1}{4z^2},
\end{aligned} \tag{A2}$$

where z is the independent variable. The further development is easiest if δu and δv are eliminated so that

$$\left(\frac{1-z}{z}\right)^{1/2}\left[\left(\frac{1-z}{z}\right)^{1/2}\delta x'\right]' = 4\nu^2 \frac{\partial}{\partial \delta x}\left(\frac{1}{\partial x^2 + (2z + \delta y)^2}\right), \tag{A3}$$

$$\left(\frac{1-z}{z}\right)^{1/2}\left[\left(\frac{1-z}{z}\right)^{1/2}\delta y'\right]' = 4\frac{\partial}{\partial \delta y}\left(\frac{1}{\delta x^2 + (2z + \delta)^2}\right) + \frac{1}{z^2}.$$

The right-hand sides can be expanded in powers of δx and δy with the help of Legendre polynomials.

After some straightforward transformation one finds that

$$\begin{aligned}
z(1-z)\delta x'' - \tfrac{1}{2}\delta x' + \frac{\nu^2}{2z} &= -\nu^2 \sum_{n+2m>0} \frac{(-1)^{n+m}(n+2m+2)!}{2^{n+4m+2}(m+1)!\,m!\,n!} \\
&\quad \times \frac{\delta y^n \delta x^{2m+1}}{z^{n+2m+1}}, \\
z(1-z)\delta y'' - \tfrac{1}{2}\delta y' - \frac{1}{z}\delta y &= -\sum_{n+2m>1} \frac{(-1)^{n+m}(n+2m+1)!}{2^{n+4m} m!\,m!\,n!} \\
&\quad \times \frac{\delta y^n \delta x^{2m}}{z^{n+2m}}.
\end{aligned} \tag{A4}$$

The convergence of these expansions is assured as long as $\delta x^2 + \delta y^2 < 4z^2$. All the linear terms in (A4) are on the left-hand side.

The first step in solving (A4) is to ignore the right-hand sides because they are at least quadratic in δx and δy. The solutions of the resulting homogeneous equation in δy are ignored altogether because they amount to shifting the time of collision away from $t = 0$, i.e., $z = 0$, and to changing the total energy away from its standard value $-\frac{1}{2}$. Among the two solutions of the homogeneous equation for δx we retain only the following:

$$\delta x = A z^{\alpha} F(\alpha - 1, \alpha; 2\alpha - 1/2; z), \tag{A5}$$

where $\alpha = 1/4[3 + (9 - 8\nu^2)^{1/2}]$. The other, linearly independent solution starts with the $1/4[3 - (9 - 8\nu^2)^{1/2}]$ power of z, so that νx does not vanish as z goes to zero, i.e., the trajectory does not go into a collision with the origin.

The approximation (A5) can now be improved if (A5) is inserted into the lowest terms on the right-hand side of (A4) and if we retain always the lowest significant powers of z. This procedure is quite systematic and yields the expansion (63). It is worth inserting these formulas into the total energy

$$\begin{aligned}
&\frac{\mu^2}{4}\frac{1-z}{z}(\delta x')^2 + \frac{1-z}{4z}(2 + \delta y')^2 \\
&\quad - \sum_{n+2m\geq 0} \frac{(-1)^{n+2m}(n+2m)!}{2^{n+4m} m!\,m!\,n!} \frac{\delta y^n \delta x^{2m}}{z^{n+2m+1}} = -1,
\end{aligned} \tag{A6}$$

in order to check the powers of z which are compensated. The A and B terms in (63) are needed to eliminate all negative powers.

[1] A. Einstein, Verh. Dtsch. Phys. Ges. 19, 82—92 (1917).
[2] J.B. Keller, Ann. Phys. 4, 180—8 (1958); J.B. Keller and S.I. Rubinow, Ann. Phys. 9, 24—75 (1960); V.P. Maslov, *Theorie des perturbations et methodes asymptotiques* (Gauthier-Villars, Paris, 1972); M.V. Berry and K.E. Mount, Rep. Prog. Phys. 35, 315—97 (1972); A. Voros, Ann. Inst. Henri Poincaré 24, 31—90 (1976).
[3] M.C. Gutzwiller, J. Math. Phys. 8, 1979—2000 (1967); 10, 1004—20 (1969); 11, 1791—1806 (1970); 12, 343—58 (1971), which are referred to by the roman numerals I, II, III, IV. Also M.C. Gutzwiller, J. Math. Phys. 14, 139—52 (1973) to be referred to as AK1.
[4] V.I. Arnold and A. Avez, *Problems ergodignes de la Mecanique Classique* (Gauthier-Villars, Paris, 1967); J. Moser, *Stable and Random Motions in Dynamical Systems* (Princeton U.P., Princeton, N.J., 1973).
[5] J. Hadamard, J. Math. Pure Appl. 4, 27—87 (1898); E. Artin, Abh. Math. Sem. Hamburg 3, 170—5 (1924); G.A. Hedlund, Bull. Am. Math. Soc. 241—60 (1939); D.V. Anosov, Proc. Steklov Inst. Math. 97 (translated by Am. Math. Soc., Providence, 1969); E. Hopf, Bull. Am. Math. Soc. 77, 863—77 (1971).
[6] H. Huber, Math. Annalen 138, 1—26 (1959); H.P. McKean, Comm. Pure Appl. Math. 25, 225—46 (1972); A.B. Venkov, Proc. Steklov Inst. Math. 125, 1—48 (1973).
[7] J. Hadamard, "Sur le billiard non-Euclidien," Soc. Sci. Bordeaux, Proces Verbaux 1898, 147 (1898).
[8] Y. Colin de Verdiere, Compos. Math. 27, 83—106, 159—84 (1973); J. Chazarain, Invent. Math. 24, 65—82 (1974); J.J. Duistermaat and V.W. Guillemin, Invent. Math. 29, 39—79 (1975); A. Weinstein, *Lecture Notes in Mathematics* 459 (Springer, Berlin, 1975), pp. 341—372.
[9] R.B. Balian and C. Bloch, Ann. Phys. (N.Y.) 60, 401—47 (1970); 63, 592—606 (1971); 64, 271—307 (1971); 69, 76—160 (1972).
[10] M. Kac, Am. Math. Monthly 73, 1—23 (1966); M. Berger, C.R. Acad. Sci. Paris 263A, 13—7 (1966); H.P. McKean and I.M. Singer, J. Diff. Geom. 1, 43—69 (1967).
[11] I.C. Percival, J. Phys. A. Math. Nucl. Gen. 7, 794—802 (1974); W. Eastes and R.A. Marcus, J. Chem. Phys. 61, 4301—6 (1974); D.W. Noid and R.A. Marcus, J. Chem. Phys. 62, 2119—24 (1974).
[12] W.H. Miller, J. Chem. Phys. 63, 996—9 (1975).
[13] S. Smale, Bull. Am. Math. Soc. 73, 747—817 (1967), cf. p. 801.

Further applications of harmonic mappings of Riemannian manifolds to gravitational fields*

Ahmet Eriş

Department of Physics, Middle East Technical University, Ankara, Turkey
(Received 4 October 1976)

We consider static, spherically-symmetric and stationary, axially-symmetric gravitational fields in the framework of harmonic mappings of Riemannian manifolds. In this approach the emphasis is on a correspondence between the solution of the Einstein field equations and the geodesics in an appropriate Riemannian configuration space. Using Hamilton–Jacobi techniques, we obtain the geodesics and construct the resulting space–time geometries. For static space–times with spherical symmetry we obtain the well-known solutions of Schwarzschild, Reissner–Nordström, and Janis–Newman–Winicour, where in the latter two cases the source for the geometry is the static, source-free Maxwell field and the massless scalar field respectively. When we consider the source to be a triad of massless scalar fields with the internal symmetry SU(2) × SU(2), we find that the space–time geometry is once again that of Janis–Newman–Winicour. Finally we formulate the stationary, axisymmetric gravitational field problem in terms of composite mappings and obtain the Weyl, Papapetrou, and the generalized Lewis and van Stockum classes of axisymmetric solutions.

1. INTRODUCTION

Variational principles which can be cast into the form of Eells and Sampson's theory of harmonic mappings of Riemannian manifolds[1] occupy a special place in physics. In fact it is possible to formulate all compensating field theories in the framework of generalized harmonic mappings from a Riemannian manifold M into a principal fiber bundle P over M.[2] The main advantage of this formulation lies in the direct geometrical insight it affords into the structure of the Lagrangian density. In an earlier paper[3] we had considered stationary, cylindrically-symmetric vacuum gravitational fields where the Hilbert Lagrangian density is in the form of the "energy" functional for harmonic mappings of Riemannian manifolds. Consequently we were able to read off the geometry of the configuration space for such gravitational fields and we had demonstrated the correspondence between the solutions of the Einstein field equations and the geodesics of the configuration space M'. The geometry of M' is a two-dimensional space of constant curvature with Lorentz signature and we had shown that depending on the range of values taken on by the constants of motion of M', the space–time solutions with whole cylinder symmetry fall into three distinct classes. In this case the crucial constant is the "energy" first integral and when it is positive, zero, or negative, that is for timelike, null, or spacelike geodesics on the hyperboloid we have the vacuum solutions of Lewis,[4] van Stockum,[5] and Lewis, respectively. We had also discussed the specialization to static, spherically-symmetric gravitational fields where the Einstein field equations were formulated as geodesics of a flat configuration space. The null-cone structure of the configuration space once again was observed to play an important role in delineating the regions available to the Schwarzschild and Kantowski–Sachs solutions. However in this case geodesics cross the null cone and we had shown that this corresponds to passage through the Schwarzschild horizon.

In the present work we consider the formulation of static, spherically-symmetric gravitational fields with sources in terms of the geodesics of a configuration space M'. Including an electromagnetic field as the source for the geometry we obtain the Reissner–Nordström solution. The null cone defined on the configuration space separates the Reissner–Nordström solution into three classes depending on the sign of e-m where e is the electric charge and m is the gravitational mass. When we consider the source to be a massless scalar field we obtain the solution of Janis, Newman, and Winicour.[6] In this case M' is a three-dimensional, conformally-flat space with Lorentz signature. Finally, we consider a massless triad of scalar fields with the internal symmetry group SU(2)×SU(2) coupled to the gravitational field. The resulting space–time geometry turns out to be identical to that of Janis, Newman, and Winicour. We find that in the flat-space limit total energy of the static "pion" field necessarily vanishes,[7] however in general for curved space–time such a condition is not required.

Once the geodesics of M' are obtained it is possible to generate new classes of solutions by using composite mappings. Starting from the fact that the composition of a totally geodesic map and a harmonic map is harmonic,[1] we introduce a third Riemannian manifold M'' which is one-dimensional and consider the map $f=k\circ h$ where $h:M\to M''$ and $k:M''\to M'$. The requirement that f be harmonic means that we are interested in the geodesics of M' parametrized by a function satisfying Laplace's equation on M. Expressing this function in terms of a multipole expansion we obtain three classes of axially-symmetric solutions of the Einstein field equations of gravitation. Because of their cylindrically-symmetric subclasses, which correspond to the monopole term in the expansion, these solutions are called stationary, axisymmetric solutions of Lewis and van Stockum. When a variational principle first used by Ernst[8] is formulated in terms of composite mappings we obtain the well known Weyl[9] and Papapetrou[10] classes of axisymmetric solutions. For problems with spherical symmetry, since M is already one-dimensional, the composite map formulation amounts to a coordinate transformation in the space–time metric.

 Copyright © 1977 American Institute of Physics

2. STATIC, SPHERICALLY-SYMMETRIC SPACE-TIMES

We had earlier considered the geometry of the configuration space for static, spherically-symmetric vacuum gravitational fields.[3] For this purpose we had started with the well-known general line element describing such a geometry

$$dS^2 = -B^{-2}dt^2 + A^{-2}[dr^2 + r^2(d\Theta^2 + \sin^2\Theta\, d\Phi^2)], \tag{1}$$

where A and B are functions of r only. Looking at the Einstein Lagrangian which is obtained in this case,[11] we had observed that the solutions of the Einstein field equations for vacuum can be formulated as harmonic maps $f: M \to M'$ where

$$ds^2 = dr^2 + r^2(d\Theta^2 + \sin^2\Theta\, d\Phi^2) \tag{2}$$

and

$$ds'^2 = A^{-3}B^{-1}dA^2 + 2A^{-2}B^{-2}dAdB, \tag{3}$$

are the metrics for M, M', respectively. The metric for M given by Eq. (2) does not change if we consider a static, spherically-symmetric gravitational field coupled to the source-free Maxwell field, the zero-rest-mass scalar field or to a triad of massless scalar fields with the internal symmetry group SU(2)×SU(2). Constructing the appropriate source Lagrangians[12] we see that the coupled field equations for these cases can be formulated as the harmonic maps $f: M \to M'$ where

$$ds'^2 = A^{-3}B^{-1}dA^2 + 2A^{-2}B^{-2}dAdB + BA^{-1}dA_0^2, \tag{4}$$

$$ds'^2 = A^{-3}B^{-1}dA^2 + 2A^{-2}B^{-2}dAdB - A^{-1}B^{-1}d\phi^2, \tag{5}$$

and

$$ds'^2 = A^{-3}B^{-1}dA^2 + 2A^{-2}B^{-2}dAdB - A^{-1}B^{-1}(1 + \boldsymbol{\pi}\cdot\boldsymbol{\pi}/4f^2)^{-2}d\boldsymbol{\pi}\cdot d\boldsymbol{\pi}, \tag{6}$$

are the metrics for M', respectively. In Eq. (4) A_0 is the only nonzero component of the electromagnetic vector potential, in Eq. (5) ϕ stands for the scalar field, and in Eq. (6) $\boldsymbol{\pi}$ is a three-component Cartesian vector denoting the triad of scalar fields. Since the metric coefficients in Eq. (1) depend only on r, M is effectively one-dimensional and there is a one-to-one correspondence between the solutions of the field equations and the geodesics of the configuration space.

For a unified treatment of static problems with spherical symmetry we consider the transformations

$$A = R^{-2}e^{-2\theta}, \quad B = e^{2\theta}, \tag{7}$$

which enable us to write the metrics in Eqs. (3)–(6) for the configuration space M' in the form

$$ds'^2 = dR^2 - R^2 d\Sigma^2, \tag{8}$$

where $d\Sigma^2$ is the metric of a Riemannian manifold. For vacuum solutions this metric is given by

$$d\Sigma^2 = d\theta^2, \tag{9}$$

and as we had discussed previously, results in a flat configuration space M'. In this system of coordinates the Hamiltonian is given by

$$H = P_R^2 - R^{-2}P_\theta^2, \tag{10}$$

and the momentum P_θ which is a constant of motion is identified with the Schwarzschild mass m. For problems with sources the metrics

$$d\Sigma^2 = d\theta^2 - e^{4\theta}dA_0^2, \tag{11}$$

$$d\Sigma^2 = d\theta^2 + d\phi^2, \tag{12}$$

and

$$d\Sigma^2 = d\theta^2 + (1 + \boldsymbol{\pi}\cdot\boldsymbol{\pi}/4f^2)^{-2}d\boldsymbol{\pi}\cdot d\boldsymbol{\pi} \tag{13}$$

lead, respectively, to solutions where the source of the gravitational field is the free Maxwell field, the zero-rest-mass scalar field and the triad of massless scalar fields.

We shall now investigate the geodesics of these configuration spaces using the Hamilton–Jacobi theory. We start by considering the static, spherically-symmetric free Maxwell field to be the source for the geometry. In this case the Hamilton–Jacobi equation is written as

$$\frac{\partial S}{\partial \lambda} + \left(\frac{\partial S}{\partial R}\right)^2 - R^{-2}\left[\left(\frac{\partial S}{\partial \theta}\right)^2 - e^{-4\theta}\left(\frac{\partial S}{\partial A_0}\right)^2\right] = 0, \tag{14}$$

where

$$\lambda = r^{-1}, \tag{15}$$

is the affine parameter for the geodesics. The complete solution of Eq. (14) is readily found by separation of variables to be

$$S = -\varepsilon\lambda + \alpha A_0 + (\varepsilon R^2 + \epsilon)^{1/2} - \tfrac{1}{2}(\alpha^2 e^{-4\theta} + \epsilon)^{1/2}$$
$$\begin{cases} -\epsilon^{1/2}\coth^{-1}\left(\dfrac{\varepsilon}{\epsilon}R^2 + 1\right)^{1/2} + \dfrac{\epsilon^{1/2}}{2}\coth^{-1}\left(\dfrac{\alpha^2}{\epsilon}e^{-4\theta} + 1\right)^{1/2}, \\ \qquad \epsilon \geq 0 \\ -(-\epsilon)^{1/2}\tan^{-1}\left(-\dfrac{\varepsilon}{\epsilon}R^2 - 1\right)^{1/2} \\ \quad + \dfrac{(-\epsilon)^{1/2}}{2}\tan^{-1}\left(-\dfrac{\alpha^2}{\epsilon}e^{-4\theta} - 1\right)^{1/2}, \quad \epsilon < 0. \end{cases} \tag{16}$$

Various constants appearing in S are expressed in terms of the momenta as

$$\alpha = P_{A_0}, \quad \epsilon = P_\theta^2 - e^{-4\theta}P_{A_0}^2, \quad \varepsilon = P_R^2 - R^{-2}\epsilon. \tag{17}$$

Solving for the metric coefficients, by differentiating S with respect to three constants in Eq. (17), and constructing the space–time metric (1) we obtain the well-known Reissner–Nordström solution

$$dS^2 = -\left(1 - \frac{m^2 - e^2}{4r^2}\right)^2\left[\left(1 + \frac{m}{2r}\right)^2 - \frac{e^2}{4r^2}\right]^{-2}dt^2 + \left[\left(1 + \frac{m}{2r}\right)^2 - \frac{e^2}{4r^2}\right]^2(dx^2 + dy^2 + dz^2), \tag{18}$$

$$A_0 = \frac{e}{r}\left[\left(1 + \frac{m}{2r}\right)^2 - \frac{e^2}{4r^2}\right]^{-1}, \tag{19}$$

where $r^2 = x^2 + y^2 + z^2$, m is the gravitational mass and e is the electric charge. In terms of the constants of motion on M' we have

$$m/e = \pm\cosh\epsilon^{1/2}, \quad \epsilon \geq 0,$$
$$m/e = \pm\cos(-\epsilon)^{1/2}, \quad \epsilon < 0, \tag{20}$$
$$e = \alpha.$$

We next consider the massless scalar field as the

source for the geometry. Using Eqs. (8) and (12), the Hamilton–Jacobi equation is written as

$$\frac{\partial S}{\partial \lambda}+\left(\frac{\partial S}{\partial R}\right)^2-R^{-2}\left[\left(\frac{\partial S}{\partial \theta}\right)^2+\left(\frac{\partial S}{\partial \phi}\right)^2\right]=0, \tag{21}$$

where λ is again given by Eq. (15). The complete solution of Eq. (21) is

$$\begin{aligned} S=&-\varepsilon/r+\beta\theta+\gamma\phi+(\varepsilon R^2+\beta^2+\gamma^2)^{1/2}\\ &-(\beta^2+\gamma^2)^{1/2}\coth^{-1}\left(\frac{\varepsilon R^2}{\beta^2+\gamma^2}+1\right)^{1/2}, \end{aligned} \tag{22}$$

where the constants appearing in the solution are given by

$$\beta=P_\theta,\quad \gamma=P_\phi,\quad \varepsilon=P_R^2-R^{-2}(P_\theta^2+P_\phi^2). \tag{23}$$

The space–time metric (1) constructed from Eq. (22) is

$$\begin{aligned} dS^2=&-\left|\frac{1-\mu/2r}{1+\mu/2r}\right|^{2\beta/\mu}dt^2+\left|\frac{1+\mu/2r}{1-\mu/2r}\right|^{2\beta/\mu}\\ &\times\left(1-\frac{\mu^2}{4r^2}\right)^2(dx^2+dy^2+dz^2), \end{aligned} \tag{24}$$

where $r^2=x^2+y^2+z^2$ and $\mu\equiv(\beta^2+\gamma^2)^{1/2}>0$. The expression for the scalar field

$$\phi=\frac{\gamma}{\mu}\ln\left|\frac{1-\mu/2r}{1+\mu/2r}\right| \tag{25}$$

vanishes for $\gamma=0$ and in this case Eq. (24) reduces to the usual Schwarzschild metric. This solution was first obtained by Janis, Newman, and Winicour.[6]

We finally consider the source for the geometry to be a triad of massless scalar fields with the internal symmetry group SU(2)×SU(2). The Hamilton–Jacobi equation constructed using Eqs. (8) and (13) is written in a separable form as

$$\begin{aligned} &\frac{\partial S}{\partial \lambda}+\left(\frac{\partial S}{\partial R}\right)^2-R^{-2}\left(\frac{\partial S}{\partial \theta}\right)^2-k^2R^{-2}\left\{\left(\frac{\partial S}{\partial \zeta_1}\right)^2\right.\\ &\left.+\sin^{-2}\zeta_1\left[\left(\frac{\partial S}{\partial \zeta_2}\right)^2+\sin^{-2}\zeta_2\left(\frac{\partial S}{\partial \zeta_3}\right)^2\right]\right\}=0, \end{aligned} \tag{26}$$

$$k\equiv f^{-1}\neq 0,$$

where we have defined three new "pion" coordinates through the transformations

$$\begin{aligned} \pi^1&=2f\tan\frac{\zeta_1}{2}\sin\zeta_2\cos\zeta_3,\\ \pi^2&=2f\tan\frac{\zeta_1}{2}\sin\zeta_2\sin\zeta_3,\\ \pi^3&=2f\tan\frac{\zeta_1}{2}\cos\zeta_2. \end{aligned} \tag{27}$$

Constructing the space–time metric (1) we find that the solution is exactly that of Janis, Newman, and Winicour, given by Eq. (24), with

$$\mu\equiv(\beta^2+\gamma^2k^2)^{1/2}>0,\quad k\neq 0. \tag{28}$$

The expression for the "pion" field Π, where

$$\Pi=(1+\pi\circ\pi/4f^2)^{-1}\pi, \tag{29}$$

is given by

$$\begin{aligned} \Pi^1&=(\nu/\gamma k)\sin\phi,\\ \Pi^2&=(\sigma/k\nu)\cos\phi,\\ \Pi^3&=[(\nu^2-\sigma^2)^{1/2}/k\nu]\cos\phi, \end{aligned} \tag{30}$$

where

$$\phi=\frac{\gamma k^2}{2\mu}\ln\left|\frac{1-\mu/2r}{1+\mu/2r}\right|, \tag{31}$$

and γ, ν, σ are constants satisfying $\sigma^2<\nu^2<\gamma^2$. In terms of the momenta we have

$$\begin{aligned} \gamma^2&=P_{\zeta_1}^2+\sin^{-2}\zeta_1\nu^2,\\ \nu^2&=P_{\zeta_2}^2+\sin^{-2}\zeta_2\sigma^2,\\ \sigma&=P_{\zeta_3}. \end{aligned} \tag{32}$$

The expression for Π given by Eq. (29) is determined up to a transformation belonging to the symmetry group SU(2)×SU(2). For the representation of the group see, for example, Meetz.[12] In the limit of no interaction between the components of the "pion" field, i.e., $f\to\infty$, we have three independent massless scalar fields

$$\Pi^i=\frac{e^i}{2\mu}\ln\left|\frac{1+\mu/2r}{1-\mu/2r}\right|, \tag{33}$$

where $\mu\equiv(\beta^2+\gamma^2)^{1/2}>0$ and $(e^1)^2+(e^2)^2+(e^3)^2=\gamma^2$. The resulting space–time geometry is again that of Janis, Newman, and Winicour given by Eq. (24). These results imply that one can devise a procedure for constructing solutions to field equations of gravity coupled to a multiplet of zero-rest-mass scalar fields with internal symmetry from solutions to the coupled Einstein-massless scalar field equations. A similar procedure, relating the solutions of the coupled Einstein–Yang–Mills equations to the solutions of the coupled Einstein–Maxwell equations was recently presented by Yasskin.[13]

In order to present the solutions in a unified scheme we shall now turn to an examination of the phase space for M'. As in the case of the metrics for M' we shall find it convenient to write the Hamiltonian in the form

$$H_{s'}=P_R^2-R^{-2}H_\Sigma, \tag{34}$$

where H_Σ denotes the "kinetic energy" associated with the metric $d\Sigma^2$ in question. For cases under investigation we have

$$H_\Sigma=P_\theta^2, \tag{35}$$

$$H_\Sigma=P_\theta^2+P_\phi^2, \tag{36}$$

$$H_\Sigma=P_\theta^2+k^2[P_{\zeta_1}^2+\sin^{-2}\zeta_1(P_{\zeta_2}^2+\sin^{-2}\zeta_2P_{\zeta_3}^2)], \tag{37}$$

$$H_\Sigma=P_\theta^2-e^{-4\theta}P_{A_0}^2. \tag{38}$$

Denoting the "energy" constant for $H_{s'}$ by ε and that for H_Σ by ϵ we have

$$\varepsilon=P_R^2-R^{-2}\epsilon. \tag{39}$$

For vacuum solutions and solutions where the zero-rest-mass scalar field (or a multiplet of scalar fields with internal symmetry) is the source for the Einstein field equations, i.e., for Eqs. (35)–(37), the "energy" constant ϵ is necessarily positive definite. Hence, defining P_Σ through $\epsilon=R^2P_\Sigma^2$ and considering a two-dimen-

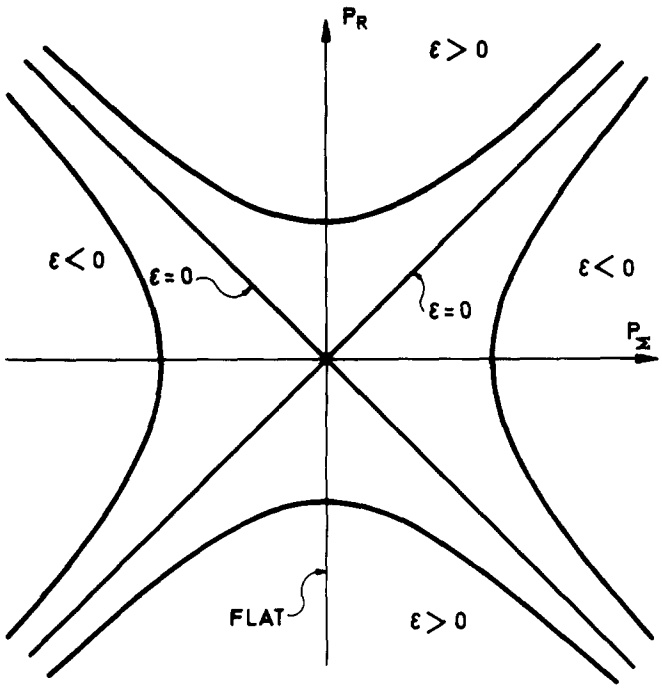


FIG. 1. A submanifold of the phase space for static, spherically-symmetric vacuum gravitational fields and those where the source for the geometry is the massless scalar field. Particular examples of solutions are the hyperbolas in this figure.

sional submanifold of the phase space spanned by the momenta P_R and P_Σ, solutions to the Einstein field equations are represented by the family of hyperbolas

$$\varepsilon = P_R^2 - P_\Sigma^2. \tag{40}$$

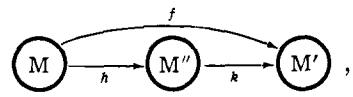
In Fig. 1 we can see an illustration of this situation. The vertical axis $P_\Sigma = 0$ corresponds to flat space–time where, depending on the solution, the Schwarzschild mass m and the scalar field strength ϕ vanish. As far as solutions to the Einstein field equations are concerned there is no distinction between the different branches of the hyperbolas but the null cone is prohibited since the space–time metric coefficients become unbounded there. For the remaining case, where we have a source-free Maxwell field as the source for the Einstein field equations, the most crucial constant is the "energy" constant ϵ and for representation of the solutions it suffices to consider the two-dimensional submanifold of the phase space spanned by the momenta P_Σ and P_{A_0}. As shown in Fig. 2 the Reissner-Nordström solution is represented by the family of hyperbolas

$$\epsilon = P_\theta^2 - e^{-4\theta} P_{A_0}^2. \tag{41}$$

For $\epsilon > 0$ we have $m > |e|$ and these solutions are represented by the family of hyperbolas which lie within the "future" and the "past" null cones of this figure. For $\epsilon < 0$ we have $m < |e|$ and solutions are represented by the family of hyperbolas which lie in the "elsewhere" region. The null cone itself represents solutions where $m = |e|$ and the origin corresponds to flat space–time where $m = |e| = 0$.

3. COMPOSITE MAPPINGS AND STATIONARY SPACE-TIMES WITH AXIAL SYMMETRY

So far in this work we have considered problems where the coefficients of the space–time metric depended on one variable only. This was essential in the reduction of the solution of the Einstein field equations to a study of the geodesics of the appropriate configuration space. The solution to the more general problem where metric coefficients are functions of more than one variable is equivalent to constructing the extremal submanifolds of the configuration space. There is, however, a subclass of solutions to the general axisymmetric problem which once again requires only a knowledge of the geodesics of the configuration space. Examples of such stationary, axially symmetric gravitational fields are the Weyl and Papetrou solutions which are obtained by assuming a functional dependence between the metric coefficients. From the viewpoint of harmonic mappings of Riemannian manifolds these solutions are constructed by introducing a third Riemannian manifold M'' which is one-dimensional. As indicated below,



the map $f: M \to M'$ can then be regarded as a composite map

$$f = k \circ h, \tag{42}$$

where $h: M \to M''$ and $k: M'' \to M'$. Constructing the associated energy functional $h: M \to M''$ we find that the condition for h to be harmonic is given simply by

$$\Delta h = 0, \tag{43}$$

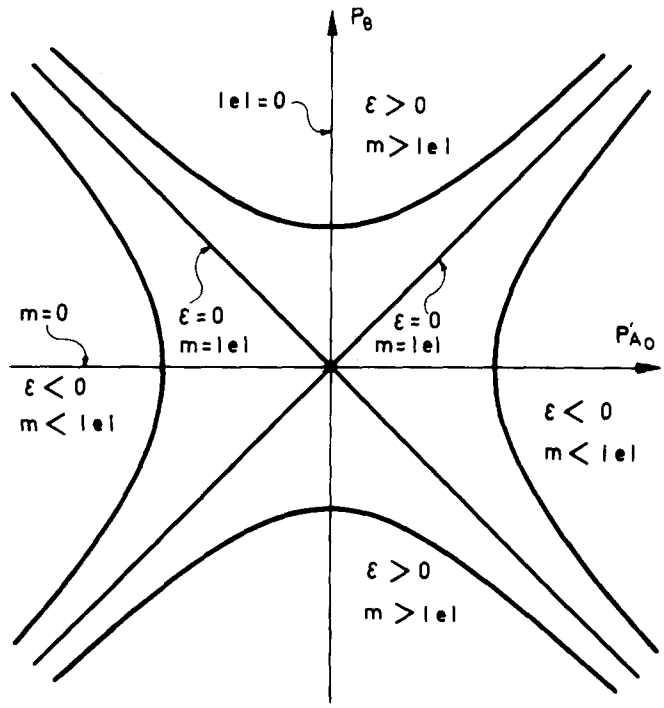


FIG. 2. A submanifold of the phase space for static, spherically-symmetric gravitational fields where the source-free Maxwell field is the source for the geometry. Note that we have $P'_{A_0} = e^{-2\theta} P_{A_0}$ for the horizontal axis here.

where Δ is the covariant Laplace operator on M. On the other hand, the requirement that f be harmonic means that k is a totally geodesic map, since M'' is one-dimensional and the composition of a harmonic and a totally geodesic map is harmonic.[1] We shall now reformulate the problem of stationary, axisymmetric vacuum gravitational fields in terms of such composite maps.

We start by considering three Riemannian manifolds, M, M'', and M' with the metrics

$$ds^2 = d\rho^2 + dz^2 + \rho^2 d\phi^2, \tag{44}$$

$$ds''^2 = d\Omega^2, \tag{45}$$

$$ds'^2 = d\psi^2 - \tfrac{1}{4}e^{4\psi}d\omega^2, \tag{46}$$

respectively, and the maps $h: M \to M''$, $k: M'' \to M'$. In terms of local coordinates the component of h is Ω and those of k are $y^A = \{\psi, \omega\}$. The condition that h be a harmonic map is given by Eq. (43) as

$$\Omega_{\rho\rho} + \frac{1}{\rho}\Omega_\rho + \Omega_{zz} = 0, \tag{47}$$

whereas the condition that k be harmonic means we are interested in the geodesics of M' given by Eq. (46) with the affine parameter Ω. Hence, by virtue of the present formulation, the expression for ψ and ω describing the geodesics of M' in terms of the affine parameter Ω which satisfies Eq. (47) provide us with solutions to the Einstein field equations for the metric

$$dS^2 = -\rho e^{2\psi}(dt - \omega d\phi)^2 + \rho^{-1}e^{2\gamma - 2\psi}[(d\rho)^2 + (dz)^2] + \rho e^{-2\psi}(d\phi)^2. \tag{48}$$

Geodesics of the configuration space (46) were studied in detail previously[3] where we had seen that the null cone structure of M' distinguishes between the three classes of solutions $L\omega_\pm$, $L\psi$, and V. Corresponding to each class we now have a larger class of axially symmetric solutions given by

$$e^{2\psi} = \beta^{-1/2}\sinh\Omega, \quad \omega = \beta^{1/2}\coth\Omega, \quad \beta > 0, \tag{49}$$

$$e^{2\psi} = |\beta|^{-1/2}\cos\Omega, \quad \omega = |\beta|^{1/2}\tan\Omega, \quad \beta < 0, \tag{50}$$

and

$$e^{2\psi} = \Omega, \quad \omega = e^{-2\psi}, \quad \beta = 0, \tag{51}$$

respectively, where Ω is an arbitrary solution of Laplace's equation (47) which can be expressed in terms of a multipole expansion. For the particular choice

$$\Omega = a + b\ln\rho, \tag{52}$$

we have the subset of solutions $L\omega_\pm$, $L\psi$, and V.

Apart from these three classes of solutions, the well-known Weyl[9] and Papapetrou[10] solutions can also be constructed using composite mappings. For this purpose we consider a different formulation of the stationary, axially symmetric gravitational field problem due to Ernst.[8] We start by defining the function Φ of ρ, z such that

$$\rho^{-1}\eta^2\nabla\omega = \hat{n}\times\nabla\Phi, \tag{53}$$

where ∇ has its usual meaning in the flat space cylindrical coordinates ρ, z, ϕ, and $\hat{n}$ is a unit vector in the azimuthal direction. In terms of Φ the vacuum Einstein field equations for the metric

$$dS^2 = -\eta(dt - \omega d\phi)^2 + \eta^{-1}[e^{2\gamma}(d\rho^2 + dz^2) + \rho^2 d\phi^2], \tag{54}$$

can be derived from the variational principle $\delta I = 0$,

$$I = 2\pi\int \eta^{-2}[(\nabla\eta)^2 + (\nabla\Phi)^2]\rho d\rho dz. \tag{55}$$

Thus we see that geodesics of M' with the metric

$$ds'^2 = \eta^{-2}(d\eta^2 + d\Phi^2), \tag{56}$$

in terms of the affine parameter Ω, satisfying (47) yields a class of solutions for the vacuum Einstein field equations for the space–time metric (54). Expressions for η and Φ as functions of Ω can be obtained by solving the Hamilton–Jacobi equation

$$\frac{\partial S}{\partial\Omega} + \eta^2\left[\left(\frac{\partial S}{\partial\eta}\right)^2 + \left(\frac{\partial S}{\partial\Phi}\right)^2\right] = 0, \tag{57}$$

to get the complete solution

$$S = -\beta\Omega + \alpha\Phi + (\beta - \alpha^2\eta^2)^{1/2} - \beta^{1/2}\coth^{-1}[\beta^{1/2}(\beta - \alpha^2\eta^2)^{-1/2}]. \tag{58}$$

In terms of the complex function

$$\xi = \frac{1 + \eta + i\Phi}{1 - \eta - i\Phi}, \tag{59}$$

the final result becomes

$$\xi = -e^{i\nu}\coth\Omega, \tag{60}$$

where ν is any real constant and Ω is a solution of the Laplace equation (47). Expressing Ω in terms of a multipole expansion one obtains the well-known solutions of Weyl[9] (for $\nu = 0$) and Papapetrou[10] (for $\nu = \pi/2$).

Theorems on the solution of the partial differential equation of Hamilton and Jacobi enable us to conclude that the solutions to the totally geodesic problems are exhaustive. That is to say, once M' is specified by imposing symmetries on space–time and prescribing the sources of the geometry, we get the complete solution for the space–time metric using the Hamilton–Jacobi theory for the geodesics of M'. Hence, for example, the set of solutions $L\omega_-$, $L\omega_+$, $L\psi$, and V exhaust the class of stationary, cylindrically-symmetric exterior solutions of the Einstein field equations. Similarily, the Schwarzschild, the Reissner–Nordström and the Janis–Newman–Winicour solutions exhaust the class of static, spherically-symmetric solutions of the Einstein field equations with appropriate sources. A similar argument is not true for the stationary axially symmetric fields obtained by using composite mappings. For example, the set of all axisymmetric solutions obtained in this section do not exhaust the class of stationary, axially-symmetric vacuum solutions of the Einstein field equations in the sense that there are other known exact solutions which cannot be expressed in terms of composite mappings. This is because of the fact that the nature of composite mappings implies a functional dependence between the coefficients for the space–time metric which in general need not be the case. The axisymmetric solutions of Lewis and van Stockum are characterized on the phase space for M' in the same way as their cylindrically-symmetric subclasses $L\omega_-$, $L\omega_+$, $L\psi$, and V. However, in the Ernst formulation the

change in the signature results in M' becoming the Poincaré upper half plane and the solutions of Weyl and Papapetrou are then represented by circles in the submanifold of the phase space spanned by the momenta P_η and P_Φ.

For problems with spherical symmetry the manifold M, described by the metric in Eq. (2), is already one-dimensional. In this case the composite map formulation does not lead to a new situation since the introduction of a third Riemannian manifold which is again one-dimensional merely replaces the space–time coordinate r by a harmonic function of r, that is $r' = ar + b$ with a, b constants. This is then equivalent to a coordinate transformation in the space–time metric.

ACKNOWLEDGMENTS

I would like to thank Professor Y. Nutku, Dr. M. Gürses, and Dr. R. Güven for helpful suggestions and many stimulating discussions.

*Research partially supported by the Turkish Scientific and Technical Research Council (TBTAK).

[1]J. Eells, Jr. and J.H. Sampson, Am. J. Math. 86, 109 (1964); Ann. Inst. Fourier 14, 61 (1964).
[2]Y. Nutku, Ann. Inst. Henri Poincaré A 21, 175 (1974).
[3]A. Eriş and Y. Nutku, J. Math. Phys. 16, 1431 (1974).
[4]T. Lewis, Proc. Roy. Soc. London A 136, 176 (1932).
[5]W.J. van Stockum, Proc. Roy. Soc. Edinburgh 57, 135 (1937).
[6]A.I. Janis, E.T. Newman, and J. Winicour, Phys. Rev. Lett. 20 878 (1968).
[7]S. Deser, M.J. Duff, and C.J. Isham, preprint (1976).
[8]F.J. Ernst, Phys. Rev. 167, 1175 (1968).
[9]H. Weyl, Ann. Phys. 54 117 (1917).
[10]A. Papapetrou, Ann. Phys. 12, 309 (1953).
[11]F.A. Kaempffer, Phys. Rev. 165, 1420 (1968).
[12]K. Meetz, J. Math. Phys. 10, 589 (1969).
[13]P.B. Yasskin, Phys. Rev. D 12, 2212 (1975).

Linearization stability of the Einstein–Maxwell system*

Judith M. Arms

Department of Mathematics, University of California, Berkeley, California 94720

We obtain conditions on a compact Cauchy surface sufficient to insure linearization stability of the coupled Einstein–Maxwell equations, and identify these conditions with the absence of globally defined infinitesimal symmetries of the fields. The appropriate domain for these symmetries is a circle bundle over spacetime. The vector potential of the electromagnetic field is identified with the connection on the bundle.

I. INTRODUCTION

This paper extends the work of Fischer and Marsden[1] and Moncrief[2] on the linearization stability of vacuum spacetimes to the case of gravity coupled with a sourceless electromagnetic field coming from a single-valued vector potential on a spacetime V with compact Cauchy surface M. Linearization stability concerns the validity of linear perturbation theory. A solution to the exact equations is said to be linearization stable if any solution to the linearized equations (relative to the exact solution) approximates to first order a curve of exact solutions. This property is not trivial, as is demonstrated by the counterexample of Brill and Deser.[3]

Our main results are two theorems giving sufficient conditions for linearization stabiltiy. Our definitions and conventions follow Misner, Thorne, and Wheeler[4] up to units. Let ${}^4g$ be the Lorentz metric (signature $-+++$) on the spacetime V, and g the restriction of ${}^4g$ to the Cauchy surface M. Greek indices range from 0 to 3 inclusive and are raised and lowered by ${}^4g$; Latin indices range from 1 to 3 and are manipulated by g. A semicolon (vertical bar) represents covariant differentiation with respect to ${}^4g$ (g). Let k be the second fundamental form or extrinsic curvature of M (so $k_{ij} = -Z_{i;j}$, where Z is the future pointing unit normal vector field on M); and let π be the conjugate momentum of g, $\pi^{ij} = (k^m_m g^{ij} - k^{ij})g^{1/2}$, where $g^{1/2} = \sqrt{\det g}$, so that π is a tensor density. Suppose ${}^4A$ is the vector potential for the electromagnetic field ${}^4F$, and E and B are the electric and magnetic field densities observed by a Gaussian normal set of observers on M. In terms of these quantities, the Einstein–Maxwell equations may be written as constraint equations and evolution equations. The constraints are (cf. Refs. 5 and 4, p. 525)

$$\begin{aligned} \mathcal{H} &= [-\tfrac{1}{2}(\pi^k_k)^2 + \pi^{jk}\pi_{jk} + \tfrac{1}{2}(E^jE_j + B^jB_j)]g^{-1/2} - Rg^{1/2} = 0, \\ \mathcal{J}_i &= -2\pi_{ij}{}^{|j} + E^j_{\ |j}A_i - [ijk]E^jB^k = 0, \\ \mathcal{K} &= -E^k_{\ |k} = 0 \end{aligned} \tag{1.1}$$

(where R is the scalar curvature of g and $[ijk]$ is the completely antisymmetric tensor density such that $[123] = +1$). Linearization stability of the four-dimensional equations reduces to that of the constraint equations.

In what follows, all tensors either are C^∞ or lie in suitable Sobolev spaces as discussed below.

Theorem 1: A solution $({}^4g, {}^4F)$ to the coupled Einstein–Maxwell equations is linearization stable if the following three conditions are satisfied:

(i) $\mathrm{tr}\pi = \pi^k_k$ is a constant multiple of the volume element on M.

(ii) One of g, π, E, or B is nontrivial; that is, at least one of π, E, or B is not identically zero, or g is not flat.

(iii) There are no simultaneous symmetries of all the fields on M; that is, if $\mathcal{L}_Y g = 0$, $\mathcal{L}_Y \pi = 0$, $\mathcal{L}_Y E = 0$, and $\mathcal{L}_Y B = 0$, then $Y = 0$.

($\mathcal{L}_Y T$ is the Lie derivative of the tensor T with respect to Y; e.g., $\mathcal{L}_Y g_{ij} = Y_{i|j} + Y_{j|i}$.)

Theorem 1 parallels the work of Fischer and Marsden. Theorem 2 below establishes global conditions for linearization stability (in contrast to the Cauchy data conditions of Theorem 1), as Moncrief[2] did for the vacuum case. The most geometric and concise description of these global conditions requires that the vector potential ${}^4A$ be identified with the connection 1-form ω of a circle bundle P over V, following Wu and Yang.[6] The bundle has a projection map $\tau: P \to V$ such that $\tau^{-1}(x)$ is isomorphic to the circle S^1 for each $x \in V$ and $\omega = \tau^* {}^4A + d\theta$, where θ is the angular coordinate on S^1. The circle action on the bundle corresponds to the trivial gauge transformation ${}^4A \to {}^4A + d\lambda$, where λ is a constant.

Theorem 2: A solution $({}^4g, {}^4F)$ of the coupled Einstein–Maxwell equations is linearization stable if there are no simultaneous symmetries of the connection one form ω and the pullback via τ of ${}^4g$ to the bundle except the circle action; that is, if $\mathcal{L}_Y \omega = 0$ and $\mathcal{L}_Y \tau^* {}^4g = 0$ together imply that Y is an infinitesimal generator of the canonical action of the circle on the bundle.

We conjecture that these conditions are necessary as well as sufficient; see Sec. V.

The proof of these theorems uses the implicit function theorem and a splitting lemma. The implicit function theorem gives the following: Suppose Φ is a C^r map, $r \geq 1$, between two Banach spaces X and Y such that $\Phi(x_0) = 0$, the kernel of $D\Phi$ at x_0 splits (i.e., $X = X_1 \otimes \ker D\Phi(x_0)$, where each summand is closed), and $D\Phi(x_0)$ is onto Y. Then the set $\{x \mid \Phi(x) = 0\}$ is a C^r manifold near x_0 with tangent space at x_0 equal to $\ker D\Phi(x_0)$. The conclusion of this theorem implies that a solution h to the linearized equation $D\Phi(x_0) \cdot h = 0$ is tangent to a curve of solutions to the exact equation $\Phi = 0$. (In order to apply this theorem, it is necessary to work with Banach or Hilbert spaces, such as Sobolev spaces of functions appropriate to insure sufficient smoothness of

 Copyright © 1977 American Institute of Physics

the map Φ. Regularity theorems can then be used to generalize the results to spaces of C^∞ functions.) This theorem reduces the proof to showing $D\Phi$ is onto and its kernel splits. The proof is further reduced by the following splitting lemma: If $D\Phi^*$, the adjoint operator to $D\Phi$, is elliptic (in the sense of Douglis and Nirenberg as extended by Hormander[7]), then the spaces X and Y in the discussion above may be split; $X = \ker D\Phi \oplus$ range $D\Phi^*$ and $Y = \text{range} D\Phi \oplus \ker D\Phi^*$. From the second splitting we see that $D\Phi$ is onto when $D\Phi^*$ is injective. Thus, to prove linearization stability it suffices to prove $D\Phi^*$ is elliptic and injective. For further discussion of these points, see Refs. 1 and 8.

Section II gives the computations that prove Theorem 1. In Sec. III we lay the groundwork for Sec. IV, where we identify elements of $\ker D\Phi^*$ with symmetries of the metric and the connection (the gravitational and electromagnetic fields) on P.

II. CAUCHY DATA CONDITIONS

Let A be the restriction of ${}^4A$ to M; then B^i defined above is $[ijk]A_{k|j}$. Let $\eta^i = -E^i$ be the conjugate momentum for A. Let $\Phi(g, A, \pi, \eta) = (\mathcal{H}, \mathcal{J}, \mathcal{K})$; then (1.1) can be written as $\Phi = 0$. We differentiate with respect to (g, A, π, η) and compute

$$
\begin{aligned}
&D\Phi(h,b,p,f)\\
&= (\{[-\pi^k{}_k\pi^{ik} + 2\pi^{ik}\pi^j{}_k + \tfrac{1}{2}(E^iE^j + B^iB^j)]h_{ij}\\
&\quad + [\tfrac{1}{4}(\pi^k{}_k)^2 - \tfrac{1}{2}\pi^{ki}\pi_{ki} - \tfrac{1}{4}(E^kE_k + B^kB_k)]h^i{}_j\\
&\quad - \pi^k{}_kb^i{}_i + 2\pi^{ij}P_{ij} + [ijk]B_ib_{k|j} - E^kf_k\}g^{-1/2}\\
&\quad + \{-\tfrac{1}{2}Rh^k{}_k + (h^k{}_k)^{|j}{}_{|j} - h^{kj}{}_{|kj} + R^{ij}h_{ij}\}g^{1/2}\\
&\quad - 2\pi^{kj}{}_{|j}h_{ik} + 2\pi^{jk}(\tfrac{1}{2}h_{jk|i} - h_{ij|k}) - 2p^{|j}{}_{ij}\\
&\quad + E^j_{|j}b_i + E^j(b_{i|j} - b_{j|i}) - f^j{}_jA_i\\
&\quad + f^j(A_{j|i} - A_{i|j}),\ f^k{}_{|k}
\end{aligned}
$$

and

$$
\begin{aligned}
&D\Phi^*(N,X,\lambda)\\
&= (N[-\pi^k{}_k\pi^{ij} + 2\pi^{ik}\pi^j{}_k + \tfrac{1}{2}(E^iE^j + B^iB^j)\\
&\quad + \{\tfrac{1}{4}(\pi^k{}_k)^2 - \tfrac{1}{2}\pi^{kl}\pi_{kl} - \tfrac{1}{4}(E^kE_k + B^kB_k)\}\\
&\quad \times g^{ij}]g^{-1/2} + [NR^{ij} - \tfrac{1}{2}NRg^{ij} + N^{|k}{}_{|k}g^{ij}\\
&\quad - N^{|ij}]g^{1/2} - \mathcal{L}_X\pi^{ij},\ [ijk](NB')_{k|j} + \mathcal{L}_XE^i,\\
&\quad N(2\pi_{ij} - \pi^k{}_kg_{ij})g^{-1/2} + \mathcal{L}_Xg_{ij},\\
&\quad - NE_ig^{-1/2} + \mathcal{L}_XA_i - \lambda_{,i}),
\end{aligned}
$$

where $B'^i = B^ig^{-1/2}$ and R^{ij} is the Ricci curvature of g. The arguments for $D\Phi^*$ are a function N, a vector field X^i, and a function λ defined only up to addition of a constant (i.e., a function modulo the constants). These are dual to the denstiy $\mathcal{H}$, the one form density $\mathcal{J}_i$, and the divergence density $\mathcal{K}$, respectively.

We show $D\Phi^*$ is elliptic. Consider the symbol of $D\Phi^*$, namely σ_ξ: $(s, V^k) \to ([\xi^k\xi_kg^{ij} - \xi^i\xi^j]g^{1/2}s + [\xi_l\pi^{li}\delta^j_k + \xi_l\pi^{lj}\delta^i_k - \xi_k\pi^{ij}]V^k,\ [\xi^kE^i - \delta^i_k\xi_lE^l]V^k,\ [g_{ik}\xi_j + g_{jk}\xi_i]V^k,\ \xi_iA_kV^k)$. If $\xi \neq 0$ and $\sigma_\xi(s, V^k) = 0$, from the third factor we get $V_i\xi_j + V_j\xi_i = 0$, so $V = 0$. Then the trace of the first factor reduces to $(n-1)\xi^k\xi_kg^{1/2}s = 0$ (where n is the dimension of $M = 3$) so $s = 0$. Therefore, the symbol is injective; i.e., $D\Phi^*$ is elliptic.

We now prove Theorem 1. We begin by showing $D\Phi^*$ is injective, i.e., the equation

$$D\Phi^*(N,X,\lambda) = 0 \tag{2.1}$$

implies $N = 0$, $X = 0$, and $\lambda = \text{const}$. Taking the trace of the first factor in (2.1) yields

$$
\begin{aligned}
&N[\tfrac{1}{4}(\pi^k{}_k)^2 - \tfrac{3}{2}\pi^{jk}\pi_{jk} - \tfrac{1}{4}(E^kE_k + B^kB_k)]g^{-1/2}\\
&\quad + [2N^{|k}{}_{|k} + -\tfrac{1}{2}NR]g^{1/2} - X^k(\pi^j{}_j)_{|k} = 0.
\end{aligned}
\tag{2.2}
$$

Using $\mathcal{H} = 0$ to eliminate the R term in (2.2), we get

$$
\begin{aligned}
&N[\tfrac{1}{2}(\pi^k{}_k)^2 - 2\pi^{jk}\pi_{jk} - \tfrac{1}{2}(E^kE_k + B^kB_k)]g^{-1/2}\\
&\quad + 2N^{|k}{}_{|k}g^{1/2} - X^k(\pi^j{}_j)_{|k} = 0.
\end{aligned}
\tag{2.3}
$$

Condition (i) implies the last term of (2.3) is zero. Observe that

$$
\begin{aligned}
&[\tfrac{1}{2}(\pi^k{}_k)^2 - 2\pi^{jk}\pi_{jk}]g^{-1/2}\\
&\quad = [-\tfrac{1}{2}(\pi^k{}_kg_{ij} - 2\pi_{ij})(\pi^l{}_sg^{ij} - 2\pi^{ij})]g^{-1/2}\\
&\quad = -2k_{ij}k^{ij}g^{1/2}
\end{aligned}
\tag{2.4}
$$

and that $k = 0$ iff $\pi = 0$. If we use (2.4), multiply (2.3) through by $-N$, and integrate, we find

$$
\begin{aligned}
&2\int N^2k_{ij}k^{ij}g^{1/2}d^3x + \tfrac{1}{2}\int N^2E^kE_kg^{-1/2}d^3x\\
&\quad + \tfrac{1}{2}\int N^2B^kB_kg^{-1/2}d^3x + 2\int N^{|k}N_{|k}g^{1/2}d^3x = 0.
\end{aligned}
$$

N must be constant, and if π, E, or B is nonzero, N must be zero. If π, E, and B are all identically zero, $\mathcal{H} = 0$ becomes $R = 0$ and the first equation of (2.1) becomes $NR^{ij} - \frac{1}{2}NRg^{ij} = 0$, which together make $NR^{ij} = 0$. In dimension three, $R^{ij} = 0$ iff g is flat; so if g is not flat, N again must be zero. Thus condition (ii) forces N to be zero. (2.1) is now reduced to

$$
\begin{aligned}
&\mathcal{L}_X\pi = 0,\quad \mathcal{L}_X\eta = -\mathcal{L}_XE = 0,\\
&\mathcal{L}_Xg = 0,\quad \mathcal{L}_XA - d\lambda = 0.
\end{aligned}
\tag{2.5}
$$

The exterior derivative at the last equation is

$$d\mathcal{L}_XA = \mathcal{L}_XdA = \mathcal{L}_XB'^b = 0, \tag{2.5'}$$

where B'^b is the covariant zero weight vector corresponding to the vector density B. By condition (iii), (2.5) and (2.5') imply X is zero. This further reduces (2.5) to $d\lambda = 0$, i.e., $\lambda = 0$ (mod constants). Thus (i), (ii), and (iii) suffice to insure that the kernel of $D\Phi^*$ is trivial, and Theorem 1 is proved.

III. $D\Phi^*$ AND THE CIRCLE BUNDLE

Global conditions for linearization stability will be established in Sec. IV by identifying the arguments of $D\Phi^*$ with vector fields on a certain five-dimensional space. The key to this identification is that the dual nature of the map Φ: $\Phi = 0$ represents the constraint equations, and in addition, $D\Phi^*$ can be used to state the evolution equations concisely. This is seen most easily when the constraint and evolution equations are derived from the variational principle

$$
\begin{aligned}
I &= \int(\pi^{ij}g_{ij,0} + \eta^iA_{i,0} - N\mathcal{H} - X^i\mathcal{J}_i - \lambda\mathcal{K})dt\\
&= \int(\pi^{ij}g_{ij,0} + \eta^iA_{i,0} - \Phi\cdot(N, X^i, \lambda)dt.
\end{aligned}
$$

(cf. Ref. 4, Chap. 21), where ${}^4g_{0i} = X_i$, ${}^4g_{00} = -(N^2 - X^iX_i)$, and ${}^4A_0 = -\lambda$. Observe that the integrand is a density on M, so with the addition of dt it becomes a density on V. Varying with respect to the fields g, A, π, and η, integrating by parts, and dropping divergences yields

$$\int[h_{ij}(-\pi^{ij}_{,t}) + b_i(-\eta^i_{,t}) + p^{ij}g_{ij,t} + f^iA_{i,t} - D\Phi(h_{ij}, b_i, p^{ij}, f^i)\circ(N, X^i, \lambda)]dt$$
$$= \int[\{h_{ij}, b_i, p^{ij}, f^i\} \circ \{((-\pi^{ij}_{,t}), (-\eta^i_{,t}), (g_{ij,t}), (A_{i,t})) - D\Phi^*(N, X^i, \lambda)\}]dt$$

where h_{ij}, b_i, p^{ij}, and f^i represent the variations in g_{ij}, A_i, π^{ij}, and η^i, respectively. From this we derive a concise statement of the evolution equations:

$$\frac{\partial}{\partial t}(g, A, \pi, \eta) = J \circ D\Phi^*(N, X, \lambda),$$

where J is the antisymmetric matrix

$$\begin{pmatrix} 0 & Id \\ -Id & 0 \end{pmatrix}.$$

A choice of arguments (N, X, λ) for $D\Phi^*$ represents a choice of coordinates (the gauge for gravity) and potential (the gauge for electromagnetism). We refer to the choice $(1,0,0)$ as Gaussian normal coordinates and (special) Coulomb gauge. (For a sourceless field, $A^i_{|i,t} = 0$ is equivalent to $\lambda^{|i}_{|i} = 0$. Classical Coulomb gauge, $A^i_{|i} = 0$, and our special Coulomb gauge, $\lambda = 0$, are particular cases of the equivalent conditions.) A nontrivial element (N, X, λ) of $\ker D\Phi^*$ corresponds to a choice of coordinate system and gauge in which the vector field $\partial/\partial t$ is an infinitesimal symmetry for the gravitational and electromagnetic fields.

Let $\tau: P \to V$ be a (trivial) principal circle bundle over V with connection 1-form ω, as in Sec. I. (The bundle would be nontrivial in the case that the vector potential ${}^4A$ is multivalued.) (N, X, λ) can be identified with a vector field on P just as Moncrief[2] identified the arguments (N, X) of his $D\Phi^*$ with the vector field $\partial/\partial t$ on V. The connection ω is a 1-form invariant under the circle action and equal to 1 on the generators of the action. Given a local section $i: U \subset V \to P$ and a coordinate system x^μ on U, there is a coordinate system (x^μ, θ) on $\tau^{-1}(U)$, where θ is the angular coordinate on S^1. In such a coordinate system ω has the coefficients $(A_\mu, 1)$. (For more details, see Refs. 6, 9–11.) Given a section over M, a choice of (N, X, λ) determines the coordinates x^μ and the section over V, thus determining the coordinates on P. If $\partial/\partial t$ represents a tangent to the section above the timelike curves of the coordinate system x^μ, then Coulomb gauge $(\lambda = 0)$ means $\omega(\partial/\partial t) = 0$.

From this point on we work in Gaussian normal Coulomb gauge with its coordinates (t, x^i, θ) on P. To avoid confusion, we relabel elements of $\ker D\Phi^*$ as (C, Y, μ). We have interpreted (C, Y, μ) geometrically; now we interpret $J \circ D\Phi^*(C, Y, \mu)$. Let ${}^5Y = (C, Y^i, -\mu)$ in Gaussian normal Coulomb gauge be a vector field on P invariant on each fiber. Suppose there is another coordinate system $(\bar{t}, \bar{x}^i, \bar{\theta})$ such that $\partial/\partial\bar{t} = {}^5Y$. Then $J \circ D\Phi^*(C, Y, \mu) = (h, b, p, f)$ is the timelike change of the fields (g, A, π, η) in the barred coordinate system; i.e., the change along the integral curves of ${}^5Y$. (h, b, p, f) is the three-plus-one breakdown of $\mathcal{L}_{{}^5Y}\tau^*{}^4g$ and $\mathcal{L}_{{}^5Y}\omega$, just as (g, A, π, η) is the three-plus-one breakdown of $\tau^*{}^4g$ and ω. The reasoning above applies only to vector fields with $C^2 - Y_iY^i > 0$, but the results extend to all cases because all the operators are linear.

IV. GLOBAL CONDITIONS

We begin the proof of Theorem 2 by showing that certain vector fields on P give rise to elements of $\ker D\Phi^*$. Suppose $\mathcal{L}_{{}^5Y}\tau^*{}^4g = 0$ and $\mathcal{L}_{{}^5Y}\omega = 0$. A calculation shows that ${}^5Y$ must be θ-invariant. The arguments of the preceding section show that $J \circ D\Phi^*(C, Y, \mu)$ is zero; thus $(C, Y, \mu) \in \ker D\Phi^*$. If ${}^5Y$ is a generator of the circle action, C and Y are zero and μ is constant. Call this the trivial Killing field for $\tau^*{}^4g$ and ω; it generates the trivial element of $\ker D\Phi^*$. For each nontrivial Killing field for both $\tau^*{}^4g$ and ω, there is a nontrivial element of $\ker D\Phi^*$.

Furthermore, distinct Killing fields give rise to distinct elements of $\ker D\Phi^*$. This follows immediately from:

Lemma: If ${}^5Y$ is Killing for $\tau^*{}^4g$ and ω, and ${}^5Y$ vanishes on the Cauchy surface M, then ${}^5Y$ vanishes everywhere.

Proof: Because ${}^5Y$ is θ-invariant, ${}^4Y = \tau^*{}^5Y$ is well defined. A calculation shows that $\mathcal{L}_{{}^4Y}{}^4g = 0$ and $0 = \mathcal{L}_{{}^5Y}\omega_\nu = \mathcal{L}_{{}^4Y}A_\nu - \mu_{,\nu}$. By the first equation and Lemma 4.2 in Ref. 2, ${}^4Y$ vanishes everywhere; then the second equation reduces to $0 = -d\mu$, so μ is constant and therefore identically zero. □

From the proof we note that for all simultaneous Killing fields ${}^5Y$, $\mathcal{L}_{{}^5Y}{}^4F = \mathcal{L}_{{}^4Y}d^4A = -d\mu^2 = 0$. In fact, ${}^5Y$ can be characterized by the requirements $\mathcal{L}_{{}^4Y}{}^4g = 0$ and $\mathcal{L}_{{}^4Y}{}^4A$ is exact; ${}^5Y^\theta$ is then defined by $d({}^5Y^\theta) = \mathcal{L}_{{}^4Y}{}^4A$, where θ indicates the fifth coordinate, along the fiber.

P can be regarded as a special case of Kaluza–Klein–Jordan–Thiry theory; see, e.g., Refs. 12, 13. The Kaluza–Klein ${}^5g$ is $\tau^*{}^4g + \omega \otimes \omega$; the Killing fields for ${}^5g$ which are θ-invariant are exactly the simultaneous Killing fields for $\tau^*{}^4g$ and ω.

Conversely, for each element of $\ker D\Phi^*$ we can construct a θ-invariant simultaneous Killing field. We use five of the twenty equations $\mathcal{L}_{{}^5Y}\tau^*{}^4g = 0$ and $\mathcal{L}_{{}^5Y}\omega = 0$ to extend ${}^5Y = (C, Y^i, -\mu)$ off M, and show that the other fifteen are then satisfied:

$$\mathcal{L}_{{}^5Y}\tau^*{}^4g_{00} = \mathcal{L}_{{}^4Y}{}^4g_{00} = -2C_{,0} = 0,$$
$$\mathcal{L}_{{}^5Y}\tau^*{}^4g_{0i} = \mathcal{L}_{{}^4Y}{}^4g_{0i} = -C_{,i} + Y^j_{,0}g_{ij} = 0,$$
$$\mathcal{L}_{{}^5Y}\omega_0 = \mathcal{L}_{{}^4Y}A_0 - \mu_{,0} = 0$$

when integrated succesively are ordinary differential equations having unique solutions that give ${}^5Y = (C, Y^i, -\mu)$ off M. $\mathcal{L}_{{}^5Y}\tau^*{}^4g_{\theta\theta} = 0$ always; and $\mathcal{L}_{{}^5Y}\tau^*{}^4g_{\theta\mu} = 0$ and $\mathcal{L}_{{}^5Y}\omega_\theta = 0$ because ${}^5Y$ is θ-invariant. This leaves $\mathcal{L}_{{}^5Y}\tau^*{}^4g_{ij}$ and $\mathcal{L}_{{}^5Y}\omega_i$. We will show that these components satisfy evolution equations which, for vanishing initial data, have only the trivial solution; thus these components, which vanish on M, must vanish everywhere.

Let $F({}^4g, {}^4A)=0$ represent the coupled Einstein–Maxwell equations. These may be considered as equations on P because everything is θ-invariant. For any vector field ${}^5Y$ on P, $\mathcal{L}_{5_Y}(F({}^4g,{}^4A))=DF({}^4g,{}^4A)\circ\mathcal{L}_{5_Y}(\tau^{*\,4}g,\omega)=0$. That is, $\mathcal{L}_{5_Y}\tau^{*\,4}g$ and $\mathcal{L}_{5_Y}\omega$ satisfy the linearized field equations. Consider these equations as constraint and evolution equations. The evolving quantities are

$$h_{ij}=Dg_{ij}\cdot\mathcal{L}_{5_Y}(\tau^{*\,4}g,\omega)=\mathcal{L}_{5_Y}\tau^{*\,4}g_{ij}=\mathcal{L}_{4_Y}{}^4g_{ij},$$
$$b_i=DA_i\cdot\mathcal{L}_{5_Y}(\tau^{*\,4}g,\omega)=\mathcal{L}_{5_Y}\omega_i=\mathcal{L}_{4_Y}{}^4A_i-\mu_{,i},$$
$$p^{ij}=D\pi^{ij}\circ\mathcal{L}_{5_Y}(\tau^{*\,4}g,\omega),$$
$$f^i=D\eta^i\circ\mathcal{L}_{5_Y}(\tau^{*\,4}g,\omega).$$

Note that h and b are exactly those quantities which we wish to prove equal to zero. The perturbed lapse, shift, and gauge, i.e., the perturbations of N, X, and λ, are zero because these are exactly $\mathcal{L}_{5_Y}\tau^*g_{00}$, $\mathcal{L}_{5_Y}\tau^{*\,4}g_{0i}$, and $\mathcal{L}_{5_Y}\omega_0$, which were set equal to zero to define ${}^5Y$ off M. (h,b,p,f) on M is exactly the derivative along ${}^5Y$ of (g,A,π,η), that is, $J\circ D\Phi^*(C,Y,\mu)=0$, so these quantities are initially zero. They satisfy the linearized evolution equations with Gauss–Coulomb gauge conditions. (h,b,p,f) identically zero is a solution because the equations are linear, and uniqueness has been proved by Choquet-Bruhat.[14] Thus ${}^5Y$ is a simultaneous Killing field for $\tau^{*\,4}g$ and ω, and Theorem 2 is proved.

From the above we can also conclude

Proposition: The dimension of $\ker D\Phi^*$ is equal to the number of linearly independent nontrivial simultaneous Killing fields of $\tau^{*\,4}g$ and ω.

V. DISCUSSION

We conjecture that the hypotheses of Theorem 2 are necessary as well as sufficient. Bourguignon, Ebin, and Marsden[15] have shown that, in dimension greater than or equal to three, linearization stability for the scalar curvature implies that the kernel of the adjoint map is trivial. An analogous proof shows that linearization stability of the Einstein–Maxwell equations implies that there are no nontrivial simultaneous Killing fields.

Coll[16] relates killing fields for ${}^4g$ and ${}^4F$ to solutions to the linearized constraint equations. His work differs from this paper in that he linearized with respect to ${}^4F$ instead of ${}^4A$; thus he obtains results on V instead of on P. He does not discuss linearization stability for the coupled system.

We expect our results to generalize in several ways. Just as Fischer, Marsden, and Moncrief have extended their results in the vacuum case to asymptotically flat spacetimes, the present results should also extend to the asymptotically flat case. The way we have handled the gauge group suggests linearization stability results for inexact electromagnetic fields (i.e., multiply-valued vector potentials) and non-Abelian gauge (Yang–Mills) fields coupled with gravity. Any field coupled with gravity for which the coupled field equations can be derived from a variational principle is susceptible to attack by these methods.

The formulation of the coupled theory in Section III provides a foundation for other results besides linearization stability. For example, the general decomposition theorem in Ref. 17 can be applied to give a simultaneous splitting of the gravitational and electromagnetic fields similar to Moncrief's splitting of gravity alone.[13] When g is flat and π, E, and B are zero, this splitting reduces to the decomposition of Barbance, Deser, and Berger-Ebin (see, e.g., Ref. 3) for gravity and the classical splitting of the electric and magnetic fields into transverse fields, gauge perturbations, and a subspace orthogonal to the constraint $\mathrm{div}E=0$.

ACKNOWLEDGMENTS

It is a pleasure to thank J.L. Kazdan, J.E. Marsden, V. Moncrief, R.K. Sachs, and A.H. Taub for many helpful discussions. Special thanks are due to Professor Marsden for a critical reading of the manuscript.

*Partially supported by NSF Grants MPS-74-21029 and PHY-76-08961.

[1]A.E. Fischer and J.E. Marsden, Bull. Amer. Math. Soc. 79, 997-1003 (1973); Proc. Symp. Pure Math. AMS 27, Part 2, 219–63 (1975).
[2]V. Moncrief, J. Math. Phys. 16, 493–98 (1975).
[3]D. Brill and S. Deser, Comm. Math. Phys. 32, 291–304 (1973).
[4]C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[5]R. Arnowitt, S. Deser, and C.W. Misner, *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962), pp. 227–65.
[6]T.T. Wu and C.N. Yang, Phys. Rev. D 12, 3845–57 (1975).
[7]L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin (1963), pp. 267–68; Ann. Math. 83, 129–209 (1966).
[8]A.E. Fischer and J.E. Marsden, Duke Math. J. 42, 519–47 (1975).
[9]S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry, Vol. I* (Interscience, New York, 1963), pp. 63–112.
[10]V.A. Lugo, preprint (1976).
[11]A. Trautman, Rep. Math. Phys. 1, 29–62 (1970).
[12]P.G. Bergmann, *Introduction to the Theory of Relativity* (Prentice-Hall, New York, 1946), pp. 254–80.
[13]A. Lichnerowicz, *Théories relativistes de la gravitation et de l'ectromagnétisme* (Masson, Paris, 1955), pp. 151–234, esp. pp. 151, 174–98.
[14]Y. Choquet-Bruhat, *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962), pp. 130–68.
[15]J.-P. Bourguignon, D.G. Ebin, and J.E. Marsden, C.R. Acad. Sci. Paris Sér. A 282, 867–70 (1976).
[16]B. Coll, C.R. Acad. Sci. Paris Sér. A 281, 1109–12 (1975).
[17]J.M. Arms, A.E. Fischer, and J.E. Marsden, C.R. Acad. Sci. Paris Sér. A 281, 517–20 (1975).
[18]V. Moncrief, J. Math. Phys. 16, 1556–60 (1975).

The theory of one-dimensional Schrödinger operators with a periodic potential*

N. N. Meiman

Formerly member of the Institute for Theoretical and Experimental Physics †
(Received 16 September 1976)

We carry out the spectral analysis of the n-band Hill equation by examining the external problem for a class of entire functions with least derivation from zero. The spectrum is characterized in terms of hyperellyctic integrals and the geometry of admissible points (band edges).

I. INTRODUCTION

It is well known that there is a connection between the solution of the Cauchy problem for the Korteweg–de Vries equation (KdV) [1] and the spectral properties of the Schrödinger operator

$$H[u] = -\partial^2/\partial x^2 + u(x;t). \tag{0.1}$$

In particular, if $u(x;t=0)$ falls off sufficiently rapidly as $|x| \to \infty$, so that

$$\int_{-\infty}^{+\infty} |xu(x;t=0)|\, dx < \infty, \tag{0.2}$$

then as $t \to \pm\infty$ the solution of the Cauchy problem for KdV with initial data $u(x;t=0)$ goes over into an n-soliton regime, where n is the number of points in the discrete spectrum of $H[u(x;t=0)]$. It is also known that the spectrum of $H[u(x;t)]$ is independent of t. The pure soliton solution serves as a reflectionless potential for the Schrödinger operator, i.e., it is highly degenerate and unstable, but it is and this is one of its most important characteristics—asymptotically stable in the sense that neighboring initial data, in the course of evolution according to the KdV become neighboring n-soliton regimes.

Recently a number of interesting mathematical results have been obtained for the Cauchy problem of the KdV equation with periodic initial data. The corresponding Schrödinger operator $H[u(x;t)]$ is an operator with a periodic potential, and its Bloch spectrum is independent of t. The investigation of the periodic Cauchy problem for the KdV equation aroused a new wave of interest in the study of the Schrödinger operator with a periodic potential. Novikov[2] first gave an argument in support of the analogy between the n-soliton solution for the KdV equation with a rapidly decreasing potential and the solution of the periodic problem for KdV, for which the Bloch spectrum of the corresponding operator has n forbidden bands. Such potentials and operators are now said to be *finite-band* or, more precisely, *n-band*. Its and Matveev[3] deepened the analogy between the n-soliton and n-band solutions. Still, it is important to note that this analogy refers entirely to the fact that both the n-soliton solution for a rapidly dropping potential and the n-band solution in the periodic case and cases of degeneracy, even very high degeneracy, i.e., solutions of similar form where the mechanisms of degeneracy have much in common (e.g., the property of being reflectionless).

The real significance of the n-soliton solutions is that they serve as asymptotic regimes for the KdV equation with nondegenerate initial data, and these regimes differ markedly from both the background and the tail. There is nothing similar for the n-band periodic solutions. Such solutions are obtained only when the initial condition $u(x;t=0)$ is itself n-band. The n-band solutions have not dynamical significance, and in this sense there is not real analogy between n-soliton and n-band solutions.

The mathematical expression of this fact is that, in the space of functions vanishing at infinity, potentials with a finite number of negative levels are nondegenerate, while, in the space of periodic functions, potentials with a finite number of gaps are infinitely degenerate.

Let $u_a(x)$ be a potential with period a and E_{2j-1}, E_{2j}, $j=1,2,\cdots$, be the forbidden bands or gaps in the Bloch spectrum of the Schrödinger operator $H[u]$. The half-line $E < E_0$ is also forbidden. The boundary points of the bands E_ν, $\nu=0,1,2,\cdots$, are points of the spectra of the periodic or antiperiodic problems for the operator $H[u]$, and completely determine the Bloch spectrum. We also naturally assume that the quantities a^2E_ν are dimensionless. The potential $u_a(x) = u_a(x/a)$.

In studying the properties of the Schrödinger operator and the KdV problem, the solution of the inverse scattering problem (ISP) plays a special role. Of central importance in solving the ISP is the forming of the trace of the shift matrix (monodromy matrix), for which we shall use the standard notation $F(E)$. The function $F(E)$ is a real entire function in the plane of the spectral variable E, with asymptotic $\cos(a\sqrt{E})$ at infinity,[4] and the Bloch spectrum is just the set of interior points of the linear manifold

$$\mu_u\{E : \mathrm{Im}F(E) = 0,\ |F(E)| \le 1\}. \tag{0.3}$$

For a nondegenerate potential the set μ_u consists of an infinite number of intervals, while for the n-band potential it consists of n finite intervals

$$[E_{2(j-1)}, E_{2j-1}],\quad j=1,2,\ldots,n, \tag{0.4}$$

and one semi-infinite interval $[E_{2n}, \infty)$. It follows that in the case of an n band potential $F(E)$ is the solution of the following extremum problem, which we call the n-band e. p.):

Among the class of entire real functions $f(E)$ which increase no faster than $\exp(a|E|^{1/2})$ and are normalized by $(n+1)$ ordered conditions

$$f(E_{2(j-1)}) = (-1)^{j-1},$$
$$j=1,2,\ldots,n+1,\ E_0 < E_2 < \cdots < E_{2n}, \tag{0.5}$$

 Copyright © 1977 American Institute of Physics

we find the function for which the deviation from zero along $[E_{2n}, \infty)$

$$L_f = \sup |f(E)|, \quad E_{2n} < E < +\infty, \tag{0.6}$$

has its minimum value $L_f = 1$.

In fact, the point E_{2n} is among the nodes—the points at which the norming conditions are assigned—so that $L \geq 1$, but according to(0.4) , $(E_{2n}, \infty) \subset \mu_u$. One sees also that it is invalid to assume that the points $\{E_0, E_2, \ldots, E_{2n}\}$ can be chosen arbitrarily in the region (0.5), since it then does not follow that for arbitrarily selected points we get the minimum deviation $L_f = 1$.

The existence of one or more extremal functions follows from the compactness of the comparison class in the space with norm

$$\|f\| = \sup |f(E) \exp(ia\sqrt{E})|, \quad E \in C_1 \setminus [E_{2n}, \infty), \tag{0.7}$$

where C_1 is the complex plane.

The present work is based on the solution of the ISP—this has become standard—and on the solution of extremal problems of the theory of analytic functions in Refs. 5—7. This is the main new point.

The general method developed in Refs. 5—7 for solving extremal problems for analytic functions is based on a detailed study of the structure of the sets $\mu\{E: \mathrm{Im} f(E) = 0, \ |f(E)| \leq L_f\}$.

One introduces the following auxiliary functions:

$$\rho_f(E) = \overline{L}_f^{-1}[f(E) - i\{L_f^2 - f^2(E)\}^{1/2}\}, \ ap_f(E) = i \ln \rho_f(E). \tag{0.8}$$

The functions ρ_f and p_f are single-valued in $C_1 \setminus \mu$; the sign of the root is chosen so that for $E \to -\infty$, $\rho(E) \sim 2 L^{-1} f(E)$. (In Ref. 7 these functions are denoted by w and ψ).

In Ref. 7 it is shown that the derivative $f'(E)$ has so many zeros that, if we draw through these zeros lines of constant argument of the function $\rho_f(E)$

$$\arg \rho_f(E) = \arg \rho_f(\xi), \tag{0.9}$$

along which $|\rho_f(E)|$ falls off as we move away from the points ξ, until their intersection with μ, then the set μ supplemented by these lines is the boundary of a simply connected region $\mathcal{G}_0 = C_1 \setminus \mu$. This region is called a fundamental region (denoted by F in Ref. 7).

The function $t = ap(E)$, as shown in Ref. 7, maps the fundamental region $\mathcal{G}_0$ conformally onto the upper half-plane $\mathrm{Im} t > 0$, with a finite number of vertical cuts with their feet on the real axis. We call this domain T. The coordinates of the feet of the cuts and their lengths are parameters for the domain T and simultaneously for the domain $\mathcal{G}_0$, which they determine to within a parallel real translation.

For the function $F(E)$, which is half the trace of the monodromy matrix, the general picture simplifies considerably. The set μ consists of the $n+1$ intervals (0.4). The supplementary system of cuts drawn through the zeros of the derivative $F'(E)$ consists of the n gaps (E_{2j-1}, E_{2j}) (on each segment there is just one zero ξ_j). The fundamental domain $\mathcal{G}_0$ is a plane with cut $[E_0, \infty)$. In this case the domain T is symmetric with respect to the imaginary axis. The functions $\rho(E)$ and $\rho^{-1}(E)$ are eigenvalues of the monodromy matrix, while $p(E)$ is the so-called quasimomentum. We shall give more details below.

In this paper we give a detailed description of the spatial region $(a; E_0, E_2, E_4, \ldots, E_{2n})$ whose points are the boundary points of admissible bands for the spectrum of the n-band operator $H[u_a]$.

In particular, we prove the natural assumption, made in the work of Dubrovin and Novkov,[8] that an arbitrary periodic potential is approximated by finite-band potentials.

For the non-self-adjoint Schrödinger operator with a complex periodic potential, we give the localization of the Bloch spectrum and the theory of expansion in eigenfunctions analogous to the usual theory of the Fourier integral. Such a treatment is also useful in the self-adjoint case.

We should like to call attention to the principal of increase, of flux which is presented in the Appendix. This principal and the results which follow from it, whose discussion we have omitted, enable one by an extremely simple and unified method to obtain very precise estimates of different perturbations.

Novikov, Dubrovin, and Krichever have obtained many profound results concerning periodic and almost-periodic problems for the KdV equation. These results are presented in a recent survey by Dubrovin, Matveei, and Novikov.[9] The periodic problem has also been treated by V. A. Marchenko.[10]

1. BLOCH SPECTRUM OF THE SCHRÖDINGER OPERATOR AND THE TRACE OF THE MONODROMY MATRIX

The Bloch spectrum of the Schrödinger operator $H[u]$ with periodic potential $u(x+a) = u(x)$, $u(x) = u(x/a)$ is the set of values E for which all solutions of the equation $(H - E)\psi = 0$ are bounded on the whole real axis. The potential $u(x)$ is assumed to be sufficiently smooth and, except in the last section, to be real. We know that the Bloch spectrum consists of the half-line $E_0 \leq E$, from which are excluded an infinite or finite number of forbidden bands (gaps) (E_{2j-1}, E_{2j}), $j = 1, 2, \cdots$. If the gaps are finite in number, we have a finite-band potential. This is a case of extremely high degeneracy. In the general case there are an infinite number of gaps, and the spectrum of the periodic problem consists of the ends of the gaps with even index j, while the spectrum of the antiperiodic problem consists of the ends of gaps with odd index:

$$\begin{aligned} &\psi(x+a, E_{2j-1}) = (-1)^j \psi(x; E_{2j-1}), \\ &\psi(x+a; E_{2j}) = (-1)^j \psi(x; E_{2j}). \end{aligned} \tag{1.1}$$

Unless otherwise stated, we assume that all the gaps are nondegenerate.

Let E be a point of the Bloch spectrum and $\varphi_{1,2}(x; E)$ some fundamental system of solutions of the equation $(H - E)\psi = 0$.

For a displacement through one period

$$\varphi_i(x+a;E)=a_{i1}\varphi_1(x;E)+a_{i2}\varphi_2(x;E). \tag{1.2}$$

The matrix

$$T(E)=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{1.3}$$

is called the shift matrix or monodromy matrix.

Since the Wronskian is constant, while, on the other hand,

$$W(\varphi_1(x+a),\varphi_2(x+a))=\det T(E)\cdot W(\varphi_1(x),\varphi_2(x)), \tag{1.4}$$

we have

$$\det T(E)=1. \tag{1.5}$$

When we transform to another fundamental system of solutions, the matrix $T(E)$ goes over into a conjugate matrix, so the trace $\mathrm{tr}T(E)$ and the eigenvalues $\rho_{1,2}(E)$ of the matrix $T(E)$ are independent of the choice of basis. We note that the dimensions of E are cm^{-2}, so $T(E)$ depends, not on E, but on the dimensionless quantity $a^2(E-E_0)$. (This assertion is equivalent to the statement that when we substitute $x=\nu\hat{x}$, $\nu>0$, the new period and spectrum are related to the old ones by the relations $\hat{a}=a\nu^{-1}$, $\hat{E}=E\nu^2$. The monodromy matrix is obviously not changed.) It follows from the general theory that $\mathrm{tr}T[a^2(E-E_0)]$ is an entire function of $a^2(E-E_0)$. We introduce the special notation

$$F[a^2(E-E_0)]=\tfrac{1}{2}\mathrm{tr}T[a^2(E-E_0)],\quad F(0)=1. \tag{1.6}$$

According to the crude asymptotic estimate (cf., for example, Ref. 4), for $|a^2E|\to\infty$, over the entire plane,

$$F[a^2(E-E_0)]\sim\cos a\sqrt{E-E_0}. \tag{1.7}$$

Since $\det T(E)=1$, $\rho_1\cdot\rho_2=1$, and we can introduce the quasimomentum $p(E-E_0)$ according to the formula [cf. below, Eq. (1.22)]

$$\rho_{1,2}[a^2(E-E_0)]=\exp[\mp iap(E-E_0)]. \tag{1.8}$$

The boundedness of all solutions of the equation $(H-E)\psi=0$ is equivalent to the equality $|\rho_{1,2}(E)|=1$, so that the points of the spectrum are characterized by the condition $|\rho_{1,2}(E)|=1$, or, what is the same thing, the reality of the quasimomentum $p(E)$. From the equation $\det T[a^2(E-E_0)]=1$ it follows that

$$\begin{aligned}&\rho_{1,2}[a^2(E-E_0)]=F[a^2(E-E_0)]\mp i\{1-F^2[a^2(E-E_0)]\}^{1/2},\\ &ap(E-E_0)=i\ln\rho_1[a^2(E-E_0)],\\ &F[a^2(E-E_0)]=\cos ap(E-E_0),\quad p(0)=0.\end{aligned} \tag{1.9}$$

The sign of the root in (1.9) is chosen so that $|\rho_1[a^2(E-E_0)]|\to\infty$ for $E\to-\infty$.

The allowed bands and gaps of the Bloch spectrum are characterized respectively by the relations

$$\begin{aligned}&\mathrm{Im}F[a^2(E-E_0)]=0,\quad |F[a^2(E-E_0)]|<1,\\ &\mathrm{Im}F=0,\quad |F|>1,\end{aligned} \tag{1.10}$$

while the spectra of the periodic and antiperiodic problems are determined by the equations $\rho_{1,2}[a^2(E-E_0)]=\pm1$ or, what is the same thing, the equations

$$F[a^2(E-E_0)]=\pm1. \tag{1.11}$$

The points of the spectrum of the periodic or antiperiodic problems are interior points of the Bloch spectrum only in the case of degeneracy, when the corresponding gap shrinks to a multiple point of the spectrum. In such points the shift matrix $T[a^2(E-E_0)]$ reduces to triangular Jordan form. Both functions $\rho_{1,2}[a^2(E-E_0)]$ are analytic and single-valued in the domain $\mathcal{G}$, which is obtained if we make cuts along all the allowed bands. It is convenient also to consider $\rho_{1,2}[a^2(E-E_0)]$ in the large, as an analytic function on the two-sheeted Riemann surface $\mathcal{R}$ of the function $\{1-F^2[a^2(E-E_0)]\}^{1/2}$. We can, for example, assume that $\mathcal{R}$ consists of two exemplars $\mathcal{G}_\pm$ of the region $\mathcal{G}$, pasted in standard fashion along the cuts.

The functions $\rho_{1,2}[a^2(E-E_0)]$ play a special role, and the equalities

$$|\rho_{1,2}[a^2(E-E_0)]|=1 \tag{1.12}$$

hold only on the closures of the allowed bands. Since $\rho_1[a^2(E-E_0)]\to\infty$ when $E\to-\infty$, and $\rho_1\rho_2=1$,

$$|\rho_2[a^2(E-E_0)]/\rho_1[a^2(E-E_0)]|<1,\quad E\in\mathcal{G}. \tag{1.13}$$

Along the boundary $\partial\mathcal{G}$ of the region $\mathcal{G}$

$$\rho_2[a^2(E-E_0)]=\overline{\rho_1[a^2(E-E_0)]}. \tag{1.14}$$

This relation together together with (1.13) means that $\rho_1[a^2(E-E_0)]$ belongs to the well-investigated HB class of functions relative to the domain $\mathcal{G}$.[11] From (1.13) it follows that

$$|\rho_1[a^2(E-E_0)]|>1,\quad \#\rho_2[a^2(E-E_0)]|<1,\quad E\in\mathcal{G}. \tag{1.15}$$

If we go over to the surface $\mathcal{R}$, these properties hold on the sheet $\mathcal{G}_+$, while on $\mathcal{G}_-$ the functions ρ_1 and ρ_2 interchange roles.

The function $\rho_1[a^2(E-E_0)]$ maps the domain $\mathcal{G}$ (or the sheet $\mathcal{G}_+$ of the surface $\mathcal{R}$) onto the infinite-sheeted Riemann surface consisting of complete sheets placed over the region $|\rho|>1$. The cut going through the allowed band $[E_{2(j-1)},E_{2j}]$ is then mapped onto the unit circle $|\rho|=1$, traversed in the negative direction.

Under such a mapping $\arg\rho_1[a^2(E-E_0)]$ decreases monotonically along the cut, and along the upper boundary of the cut the rigorous inequality

$$\frac{d}{dE}\arg\rho_1[a^2(E-E_0)]=-\frac{adp(E-E_0)}{dE}=\frac{a^2F'[a^2(E-E_0)]}{\{1-F^2[a^2(E-E_0)]\}^{1/2}}<0 \tag{1.16}$$

holds. The sign of the root is uniquely determined by the condition $(1-F^2)^{1/2}\sim iF$ for $E\to-\infty$. To a positive circling of the cut along the lower edge there corresponds differentiation with respect to $(-E)$, but the sign of the root also changes. Integrating (1.16) along the upper edge of the cut, we obtain

$$ap(E-E_0)\Big|_{E_{2(j-1)}}^{E_{2j-1}}=-\int_{a^2E_{2(j-1)}}^{a^2E_{2j-1}}\frac{F'[a^2(E-E_0)]}{\{1-F^2[a^2(E-E_0)]\}^{1/2}}\,da^2E=\pi, \tag{1.17}$$

since $-\pi$ is half the jump $f - ap$ $(E - E_0) = \arg\rho_1[a^2(E - E_0)]$ along the corresponding cut.

The jump in $ap(E - E_0)$ along the forbidden band is zero:

$$ap(E - E_0)\Big|_{E_{2j-1}}^{E_{2j}} = -\int_{a^2E_{2j-1}}^{a^2(E_{2j})} \frac{F'[a^2(E - E_0)]}{\{1 - F^2[a^2(E - E_0)]\}^{1/2}}\, da^2E = 0. \quad (1.18)$$

In fact, the ends of the bands are points of the spectrum of the periodic or antiperiodic problems, and at such points

$$\rho_{1,2}[a^2(E - E_0)] = \exp[\mp iap(E - E_0)] = \pm 1,$$

i.e., $ap(E - E_0)$ is equal to a multiple of π, so that the jump in $ap(E - E_0)$ along the forbidden band is also a multiple of π. On the other hand, along a forbidden band the integral in (1.18) is pure imaginary, so that the common value of the first two terms in (1.18) is zero. A similar argument applied to the allowed band leads to a new derivation of the equality (1.17).

From (1.16) and the normalization $p(0) = 0$ it follows that

$$ap(E - E_0) = -\int_0^{a^2(E-E_0)} \frac{F'[a^2(E - E_0)]}{\{1 - F^2[a^2(E - E_0)]\}^{1/2}}\, da^2(E - E_0). \quad (1.19)$$

The derivative $p'(E - E_0)$ has precisely one real zero ξ_j in each gap $[E_{2j-1}, E_{2j}]$ and these are its only zeros. The derivative $F'[a^2(E - E_0)]$ in addition to these zeros also has zeros at multiple points of the spectrum.

According to the crude asymptotic,[4] in the case of a continuous potential,

$$\begin{aligned} E_{2j} &= (2j)^2(\pi/a)^2 + a_0 + o(1) \\ E_{2j+1} &= (2j+1)^2(\pi/a)^2 + a_0 + o(1), \end{aligned} \quad (1.20)$$

where $o(1)$ go to zero for $j \to \infty$.

From the more precise estimates obtained in Refs. 10, 12, it follows that, if the potential $u_a(x)$ has k continuous derivatives, then

$$(E_{2k} - E_{2k-1}) \sim k^{-n}. \quad (1.21)$$

For an analytic potential the lengths of the gaps fall off exponentially.

We rewrite the expression (1.19) for $ap(E - E_0)$ in the form

$$ap(E) = \frac{a}{2}\int_{E_0}^{E} \frac{dE'}{(E' - E_0)^{1/2}} \times \prod_j \frac{(1 - E'\xi_j^{-1})\xi_j}{[E_{2j-1}E_{2j}(1 - E'E_{2j-1}^{-1})(1 - E'E_{2j}^{-1})]^{1/2}}. \quad (1.22)$$

The coefficient $a/2$ in front of the integral corresponds to the asymptotic $ap(E - E_0) \sim a\sqrt{E - E_0}$. The product in (1.22) runs over the gaps, and converges by virtue of (1.20)–(1.21). In the case of an n-band potential the product contains n factors.

The fact that $F[a^2(E - E_0)] = ap(E - E_0)$ is an entire function is equivalent to the fact that the periods of the integral (1.22) are multiples of $2\pi/a$, so that we get still another derivation of (1.17), which we rewrite in the form

$$ap(E)\Big|_{E_{2(j-1)}}^{E_{2j-1}} = \frac{a}{2}\int_{E_{2(j-1)}}^{E_{2j-1}} \frac{dE'}{(E' - E_0)^{1/2}} \prod_j = \pi. \quad (1.23)$$

From the fact that $ap(E - E_0)$ is real in the allowed bands and that the integrand (1.22) is imaginary along the gaps, it again follows that

$$ap(E - E_0)\Big|_{E_{2j-1}}^{E_{2j}} = \frac{a}{2}\int_{E_{2j-1}}^{E_{2j}} \frac{dE'}{(E' - E_0)^{1/2}} \prod_j = 0. \quad (1.24)$$

We introduce the dimensionless variable

$$z = a(E - E_0)^{1/2} \quad (1.25)$$

so that the domain $\mathcal{G}_0$ is mapped conformally on the upper half-plane $\mathrm{Im}z > 0$. In the dimensionless variable, $F[a^2(E - E_0)] = F(z^2)$, $ap(E - E_0) = \hat{p}(z)$, where

$$\hat{p}(z) = \int_0^z dz \prod_j \frac{z^2\zeta_j^2}{[(z^2 - z_{2j-1}^2)(z^2 - z_{2j}^2)]^{1/2}}, \quad (1.26)$$

$$z_k^2 = a^2(E_k - E_0), \quad \zeta_k^2 = a^2(\xi_k - E_0). \quad (1.27)$$

The relations (1.23) and (1.24) are now written in the form

$$\hat{p}(z)\Big|_{z_{2(j-1)}}^{z_{2j-1}} = \pi, \quad \hat{p}(z)\Big|_{z_{2j-1}}^{z_{2j}} = 0, \quad \hat{p}(z)\Big|_{z_{2j-1}}^{\zeta_j} = -\hat{p}(z)\Big|_{\zeta_j}^{z_{2j}}. \quad (1.28)$$

2. CANONICAL DOMAINS

The function $t = ap(E - E_0)$ is analytic in the region $\mathcal{G}$ and on the Riemann surface $\mathcal{R}$, but, as was stated in the Introduction, of particular interest is the behavior of $ap(E - E_0)$ in the fundamental domain $\mathcal{G}_0$, which in our case is simply the plane with a cut $[E_0, \infty)$.

In fact, in this case the forbidden bands $[E_{2j-1}, E_{2j}]$ play the role of the cuts supplementing the set μ. Each such segment goes through a zero of the derivative $F'[a^2(E - E_0)]$, and $\rho_1[a^2(E - E_0)]$ decreases in both directions away from ξ_j.

From (1.22) it follows that $t = ap(E - E_0)$ maps $\mathcal{G}_0$ conformally onto the upper half-plane $\mathrm{Im}t > 0$ with vertical cuts. From (1.23) it follows that the bases of the cuts lie at points that are multiples of π. These cuts are the images of the upper and lower edges of the gaps $[E_{2j-1}, E_{2j}]$. Since on the cut $[E_0, \infty)$, $ap(E + i0) = -ap(E - i0)$, the image of $\mathcal{G}_0$ is symmetric about the imaginary axis, and the images of the upper and lower edges of the gap $[E_{2j-1}, E_{2j}]$ are symmetric cuts with indices j and $-j$ and length

$$l_j = \frac{a}{2}\int_{E_{2j-1}}^{t_j} \frac{dE'}{(E' - E_0)^{1/2}} \left|\prod_j\right| = \frac{a}{2}\int_{t_j}^{E_{2j}} \frac{dE'}{(E' - E_0)^{1/2}} \left|\prod_j\right|. \quad (2.1)$$

Obviously

$$l_j = \ln\left|\rho_1(\zeta_j^2)\right| = \ln\left|F(\zeta_j^2) + [F^2(\zeta_j^2) - 1]^{1/2}\right|, \quad (2.2)$$

and $l_j \to 0$ as $j \to \infty$. More precisely,

$$2l_j \sim \left| \frac{d^2}{dz^2} F(z^2) \right|^{1/2}_{z=\zeta_j} (z_{2j} - z_{2j-1}), \tag{2.3}$$

and the coefficient of $(z_{2j} - z_{2j-1})$ in (2.3) tends to 1.

In the dimensionless variable $z = a\sqrt{E - E_0}$ there corresponds to the region $\mathcal{G}_0$ its conformal image, the upper half-plane $\mathrm{Im} z > 0$. Therefore, the region T is obviously also mapped conformally on this half-plane. The two mutually reciprocal functions $t = t(z)$ and $z = z(t)$ are symmetric with respect to the imaginary axis and $t \sim z$ at ∞. Obviously, $t(z) = \hat{p}(z)$ of (1.26). In terms of the dimensionless variable,

$$l_j = \int_{z_{2j-1}}^{\zeta_j} dz \left| \prod_j \frac{z^2 - \zeta_j^2}{[(z^2 - z_{2j-1}^2)(z^2 - z_{2j}^2)]^{1/2}} \right| = \int_{\zeta_j}^{z_{2j}} dz \left| \prod_j \right|. \tag{2.1'}$$

We shall call the image of $\mathcal{G}_0$ a canonical domain and denote it by T_F, T_u, T_{F_u}, or simply $T = T(t : \mathrm{Im} t > 0 \setminus \cup S_j)$ where $\cup S_j$ is the union of the cuts.

Turning to the Schrödinger operator $H[u(x)]$, we see that to each periodic potential $u_a(x)$ there corresponds an entire function $F_u(z^2)$ and a domain T_u, completely determined by the parameters l_j (with $l_j = 0$ at points $t = j\pi$, where there is no cut). The rate of falloff of l_j for $j \to \infty$ depends on the degree of smoothness of the potential.

In view of the simplicity of the parametrization of the canonical domains T_u, its is important to determine the nature of the correspondence between operators $H[u]$ and domains T_u.

Theorem 1: A family of Schrödinger operators $H[u]$ with one and the same canonical domain T are characterized by the fact that their spectra are obtained from one another by a shift and a dilation; more precisely, the quantities

$$z_k^2 = a^2(E_k - E_0), \quad \zeta_k^2 = a^2(\xi_k - E_0)$$

depend only on the domain T_k of index k, and not on the particular operator of the family.

Proof: The function $\hat{p}(z)$ that maps the half-plane $\mathrm{Im} z > 0$ conformally on the domain T, and is normalized by $\hat{p}(0) = 0$, $\hat{p}(z) \sim z$ at ∞, is uniquely given by Eq. (1.26). The points z_{2k-1} and z_{2k} are pre-images of the points $t = k(\pi \mp 0)$, while the points ζ_k are pre-images of the tops of the cuts with base at the points $k\pi$. All the parameters z_k and ζ_k appearing in (1.26) are uniquely determined by the domain T, i.e., follow from its parameters $\{l_k\}$.

The converse of the theorem is obvious: If the quantities z_k^2 and ζ_k^2 coincide for two potentials, there correspond to them the same function $\hat{p}(z)$ and the same domain T.

Another form of the theorem consists in the assertion that the trace of the monodromy matrix of any operator of the family is given in terms of the same function $F_T(z^2) = \cos \hat{p}(z)$

$$F_u[a^2(E - E_0)] = F_T(z^2), \quad F_T(0) = 1. \tag{2.4}$$

From (2.4) it follows that the traces of the monodromy matrices of two operators of the family take the same values at points $E^{(2)}$ and $E^{(1)}$ that are related by

$$a_2^2(E^{(2)} - E_{0,2}) = a_1^2(E^{(1)} - E_{0,1}). \tag{2.5}$$

(In this formulation the theorem is a special case of the general theorem 112 of Ref. 7.)

From this theorem it follows that if $u_2(x)$ and $u_1(x)$ are related by

$$u_2(x) - E_{0,2} = (a_1/a_2)^2[u_1(a_{1x}/a_2) - E_{0,1}], \tag{2.6}$$

then the domains T_2 and T_1 coincide.

To the operator $H[u(x)]$ there corresponds a canonical domain $T(t : \mathrm{Im} t > 0 \setminus \cup S_j)$, and conversely to each canonical domain there corresponds an infinite set of operators. In fact, for any domain T there exists a function $t = \hat{p}(z)$ that maps the half-plane $\mathrm{Im} z > 0$ conformally on T. The function $\hat{p}(z)$ is normalized uniquely by the conditions $\hat{p}(0) = 0$ and $\hat{p}(z) \sim z$ at ∞. Given arbitrary values $a > 0$ and E_0, we determine $p(E - E_0)$ and $F[a^2(E - E_0)]$ from the equations

$$ap(E - E_0) = \hat{p}[a^2(E - E_0)], \quad F[a^2(E - E_0)] = \cos ap(E - E_0). \tag{2.7}$$

According to the solution of the ISP, for each function $F[a^2(E - E_0)]$ there exists an infinite set of periodic potentials $u_a(x)$ with period a, for which $2F[a^2(E - E_0)]$ is the trace of the shift matrix. The degree of smoothness of the potential $u_a(x)$ is determined by the rate of decrease of the cuts in the domain T, in accordance with formula (2.3). Thus the following theorem holds:

Theorem 2: The set of Schrödinger operators $H[u(x)]$ with periodic potentials having a definite degree of smoothness is mapped surjectively on the set of canonical domains $T(t : \mathrm{Im} t > 0 \setminus \cup S_j)$. The rate of decrease of the lengths of the cuts S_j is related to the degree of smoothness of the potential by (2.3).

The fact that the function $t = ap(E - E_0)$ conformally maps the E plane with cut $[E_0, \infty)$ onto the canonical domain T, with the points $E_{2k-1} \pm i0$ and $E_{2k} \pm i0$ as pre-images of the points $t = \pm k(\pi \mp 0)$, and with the points $\xi_k \pm i0$ as pre-images of the vertices of the cuts $S_{\pm k}$, can, if we wish, be regarded as a necessary and sufficient condition for the sequence $\{E_\nu\}$ to serve as the spectrum of the periodic and antiperiodic problems for some Schrödinger operator $H[u_a]$.

Theorem 3: The necessary and sufficient condition for the existence of a Schrödinger operator $H[u]$, for which a given sequence $\{E_\nu\}$ serves as the spectrum of the periodic and antiperiodic problems, consists of satisfying Eqs. (1.23) and (1.24) and having asymptotic form

$$ap(E - E_0) \sim a\sqrt{E - E_0}.$$

If $l_k = 0$ for some k, i.e., if the kth gap is contracted to a point $E_{2k-1} = \xi_k = E_{2k}$, or $z_{2k-1} = \zeta_k = z_{2k}$, then

$$\frac{E - \xi_k}{[(E - E_{2k-1})(E - E_{2k})]^{1/2}} = 1, \quad \frac{z - \zeta_k}{[(z - z_{2k})(z - z_{2k-1})]^{1/2}} = 1,$$

and all the formulas containing these relations simplify. For the case of n-band potentials, the products in

formulas (1.22)–(1.24), (1.26)–(1.28), (2.1–2.2), and (2.5) contain no more than n factors.

Let us make more precise the concept of an n-band potential. To a potential $u(x)$ with no degeneracy there corresponds a canonical domain T for which all $l_k > 0$. We say that a potential is an n band potential if $l_k = 0$ for $k > n$. A nondegenerate n-band potential or an n-band potential of general form is one for which all $l_k > 0$ for $k \leq n$. All degenerate n-band potentials are then limiting cases of nondegenerate ones, and we can essentially restrict our consideration to the nondegenerate n-band potentials.

Let $u_a(x)$ be an n-band potential with period a, $T^{(n)}$ its canonical domain and $z = z(t)$ the normed function, discussed repeatedly above, that maps $T^{(n)}$ on $\mathrm{Im} z > 0$. The portion of the boundary of the domain $T^{(n)}$ between the points $t = -n\pi - 0$ and $t = n\pi + 0$ contains all $2n$ cuts of the domain $T^{(n)}$, and is mapped on the segment $[-z_{2n}, z_{2n}]$. If we interpret the function $z = z(t)$ as a complex heat potential, the length of the image of an arc of the isotherm $\mathrm{Im} z(t) = \mathrm{const}$ is equal to the thermal flux through the arc. In particular, $2z_n$ is equal to the thermal flux across the arc $[-n\pi - 0, n\pi + 0]$ of the boundary of the domain $T^{(n)}$. As the lengths of the cuts decrease, the thermal flux through this portion of the boundary of the domain T decreases strictly monotonically and reaches a minimum when all the cuts are contracted to a point. For such degeneracy $z(t) = t$ and $z_n = n\pi$, so that

$$a\sqrt{E_{2n} - E_0} = z_n \geq n\pi. \tag{2.8}$$

(In the Appendix we formulate and prove a general principle of increasing flux.)

The equality sign applies only when all gaps $[E_{2k-1}, E_{2k}]$ contract to a point. Then $u(x) = E_0$ and $F[a^2(E - E_0)] = \cos(a\sqrt{E - E_0})$.

Consideration of canonical domains makes it possible to prove the natural assumption of Dubrovin and Novikov[8] about the approximation of an arbitrary periodic potential by finite-band potentials:

Approximation Theorem: A continuous periodic potential $u_a(x)$ with period a can be approximated by finite-band analytic potentials $u_a^{(n)}(x)$ with the same period.

Proof: It is sufficient to prove the theorem for several times differentiable potentials $u_a(x)$, since arbitrarily smooth potentials are dense in the space of continuous periodic potentials.

From the solution of the inverse problem given in the paper of Stankevich[13] it follows that the potential $u_a(x)$ is uniquely determined by the function $F^{(n)}[a^2(E - E_0)]$ and the sequence (c_k, ϵ_k), where c_k is a point in the gap $[E_{2k-1}, E_{2k}]$, while the index $\epsilon_k = \pm 1$ can be considered as indicating the side of the cut on which the point (c_k, ϵ_k) lies. In the mapping onto the canonical domain T, the point (c_k, ϵ_k) is mapped into the point $\gamma_{\pm k}$, lying on the kth or $(-k)$th cut.

We denote by $T^{(n)}$ and $\{\gamma_k\}$ the domain and the n-point sequence that are obtained from T_F and $\{\gamma_k\}$ by removing all cuts with indices of modulus greater than n. We denote by $z^{(n)}(t)$, $\hat{p}^{(n)}$ and $F^{(n)}(z^2)$ the functions which, according to Theorem 1, correspond to the domain $T^{(n)}$. The quantity

$$\sum_{|k|>n} l_k \tag{2.9}$$

decreases as n increases at a rate corresponding to the degree of smoothness of the potential $u_a(x)$. From this in turn it follows that the difference

$$|z^{(n)}(t) - z(t)|, \tag{2.10}$$

as n increases, goes uniformly and sufficiently rapidly to zero over the whole of the canonical domain T.

According to Ref. 13, $F^{(n)}(z^2)$ and $\{\gamma_k\}^{(n)}$ uniquely determine the n-band potentials $u_a^{(n)}(x)$. From (2.10) it follows that $F^{(n)}(z^2)$ and $\hat{p}^{(n)}(z)$ tend uniformly over all of the half-plane $\mathrm{Im} z > 0$ to $F(z^2)$ and $\hat{p}(z)$, while it follows from Ref. 13 that then $u_a^{(n)}(x) \rightrightarrows u_a(x)$.

3. FINITE-BAND POTENTIALS AND FUNCTIONS WITH MINIMUM DEVIATION FROM ZERO

The Bloch spectrum of a Schrödinger operator $H[u]$ with a periodic potential with period a consists of the interior points of the set

$$\mu_u\{E : \mathrm{Im} F_u[a^2(E - E_0)] = 0;\ |F_u| \leq 1\}. \tag{3.1}$$

The double roots of the equation $F^2[a^2(E - E)] - 1 = 0$ belong to the spectrum, the simple roots are the boundaries of the gaps, and each gap contains precisely one root ξ of the derivative F'. For an n-band potential of general form the set μ_u consists of $n + 1$ components.

In the Introduction it was shown that for an n-band potential half the trace of the shift matrix $F[a^2(E - E_0)]$ is a solution of the following extremum problem:

In the comparison class consisting of entire real functions with asymptotic $f(E) \sim \cos a\sqrt{E}$ and normed by the $n + 1$ conditions (0.5),

$$f(E_{2(j-1)}) = (-1)^j, \quad E_0 < E_2 < \cdots < E_{2n},$$

find the function for which the deviation

$$L_f = \sup|f(E)|, \quad E \in [E_{2n}, \infty)$$

has its minimum value $L_f = 1$.

From the Phragmen–Lindelof maximum principle for a plane with cut $[E_{2n}, \infty)$, in such a domain analytic functions that increase toward infinity no faster than $\exp\{a|E|^{1/2}\}$ are compact in the space with norm

$$\|f\| = L_f = \sup|f(E)\exp(ia\sqrt{E - E_{2n}})|, \quad E \in C\setminus[E_{2n}, \infty). \tag{3.2}$$

In particular the comparison consisting of entire real functions that increase toward infinity no faster than an $|E|^{1/2}$ and are normed by (0.5) is compact in this metric. From the compactness of the comparison family it follows that there exists a function f belonging to the family for which the deviation L_f has the minimum value L. From (0.5), $L \geq 1$. On the other hand, $F_u[a^2(E - E_0)]$ belongs to the comparison family, so that if the point $E_0 < E_2 < \cdots < E_{2n}$ is taken from the spectrum of the n-band operator $H[u_a]$, half the trace of the shift matrix is an extremal function with $L_F = 1$. It then follows immediately that the points $E_0 < E_2 < \cdots < E_{2n}$ for which the

minimum deviation $L > 1$ cannot belong to the spectrum of the n-band operator $H[u_a]$.

We introduce the special notations $\mathcal{M}(a;E_{0,2n})$ and $L(a;E_{0,2n})$ for the comparison class and minimum deviation described above.

We define an n-band extremal problem (*n-band EP*) as an extremal problem with auxiliary condition $L(a;E_{0,2n}) = 1$. This additional condition separates out that part of the space $(a;E_0, E_2, \ldots, E_{2n})$ whose points are edges of gaps—with one from each gap—for the periodic operator $H[u_a]$. The purpose of the next few paragraphs is to determine this region. According to (2.8) we know, for example, that for the points of this region

$$a\sqrt{E_{2n} - E_0} \geq n\pi. \tag{3.3}$$

In the n-band EP the nodes E_0 and E_{2n} are distinguished, since E_{2j} for $j \neq 0, n$ can be either the left or right end of the jth gap, $[E_{2j-1}, E_{2j}]$ or $[E_{2j}, E_{2j-1}]$, whereas the points E_0 and E_{2n} must necessarily be the right ends of the gaps$(-\infty, E_0]$ and $[E_{2n-1}, E_{2n}]$.

In Ref. 5–7 and 14, 15 a general method was devel oped for solving the problem of minimum deviation for analytic functions, but the specific proof was given functions that were entire functions of exponential type (EET) deviating least from zero along the real axis, so we shall sometimes go over from the variable E to the variable

$$q = \sqrt{E - E_{2n}} \quad (z = a\sqrt{E - E_0}), \tag{3.4}$$

i.e., from the energy to the momentum.

We need to recall certain concepts, terms, and results of the papers cited. Let $f(q)$ be EET, L_f its deviation on the real axis, with $L_f < \infty$.

We denote by $\mathcal{M}_f$ the set of roots of the equation

$$f^2(q) - L_f^2 = 0, \tag{3.5}$$

and by $\mathcal{R}_f$ the set of real roots of (3.5), i.e., the set of points with maximum deviation of $f(q)$ on the real axis, axis, and by $T\sinh f$ we denote any maximal subset of the set $\mathcal{R}_f$, at whose points $f(q)$ assumes values with a alternating signs. Let $\{\alpha\}_f$ and $\{\alpha'\}_f$ by systems of intervals into which $\mathcal{R}_f$ and $T\sinh f$ divide the real axis, N_α and $N_{\alpha'}$ be the numbers of zeros of the function $f(q)$ in the α and α' intervals, and r the number of zeros of the function $f(q)$ in the half-plane $\mathrm{Im} q > 0$.

Of fundamental importance is the quantity introduced in[7]:

$$I_f = \sum_{\alpha} [N_\alpha/2]\,\mathrm{tr} = \sum_{\alpha'} [N_{\alpha'}/2]\,\mathrm{tr}, \tag{3.6}$$

which is called the defect index of the function $f(q)$.

All the investigations in Ref. 7 are based on a detailed study of the set

$$\mu_f\{q : \mathrm{Im} q \geq 0,\ \mathrm{Im} f = 0,\ |f| \leq L_f\}, \tag{3.7}$$

the parts of which that are in the upper half-plane $\mathrm{Im} q > 0$ being regarded as cuts. The set μ_f in turn determines the fundamental domain $\mathcal{G}_f$. We must draw through each zero of the derivative $f'(q_\nu) = 0$, $\mathrm{Im} q_\nu > 0$ a cut up to its intersection with μ_f, drawn along the line of constant argument of the function

$$L_f^{-1}[f(q) - i[L_f^2 - f^2(q)]^{1/2}] \tag{3.8}$$

along which the modulus of the function decreases as we move away from the point q_ν. The set μ_f supplemented in this way does not spoil the connectedness of the upper half-plane and is, as shown in Ref. 7, the boundary of a simply-connected region $\mathcal{G}_f$.

The set $\mathcal{M}_f$ belongs to μ_f and breaks it up into arcs along which $f(q)$ alternately changes monotonically from $-L_f$ to L_f and from L_f to $-L_f$, i.e., it plays a role for μ_f analogous to that of the set $T\sinh f$ for the real axis. The number of points of $\mathcal{M}_f$ that do not belong to $T\sinh f$ is exactly equal to $2I_f$. [This property can be taken as the definition of the defect index, and one can then obtain Eq. (3.13) as a consequence.]

Of particular interest are functions with finite defect index (FDI) with finite deviation $L_f < \infty$. These functions are completely characterized by the irrepresentation

$$f(q) = L_f \cos\varphi(q), \tag{3.9}$$

where $\varphi(q)$ is the hyperelliptic integral

$$\varphi(q) = a\int_0^q dq\ \frac{P_{2I}(q)}{[Q_{2I}(q)\,\overline{Q}_{2I}(q)]^{1/2}}. \tag{3.10}$$

Here $I = I$, $P_{2I}(q)$ and $Q_{2I}(q)$ are real polynomials of degree $2I_f$ with leading coefficients 1, and a is the type of the function $f(q)$. The integral is over a path in the fundamental region $\mathcal{G}_f$.

Since $f(q)$ is an entire function, the periods of the integral (3.10) are multiples of $2a^{-1}\pi$, where a multiplicity different from 1 denotes some degeneracy.

In Refs. 7, 15 it was shown that the solutions of various extremal problems are functions with FDI, where the defect index is less than the number of normalizing conditions. The unknown coefficients in the polynomials $P(q)$ and $Q(q)$ and the deviation L are found from the conditions of the problem.

We state Theorems 4 and 5 of Ref. 7, which are of fundamental importance for us, and which we shall cite in the following as Theorems A and B. In both theorems the comparison class in taken to be the family of functions entire in functions of type no higher than a, normed by their values at n points of the upper half-plane and at $2k$ real points.

Theorem A: Let $u(q)$ be a function of type a, belonging to the comparison family, and $L_u < \infty$. If the defect of the function $I(u) < n + k$, and if between any two neighboring points of the set $T\sinh f$ there are an even number of real nodes, then $u(q)$ is the unique extremal function of the comparison family.

Theorem B: If the defect index of the function $u(q)$ is $I(u) \geq n + k$, and if between any two neighboring points of the set $\mathcal{R}_u$ there are an even number of real nodes, then $u(q)$ is not an extremal function of the family.

In the variable q, the comparison family $\mathcal{M}(a; E_{0,2n})$ goes over into a family that is entire in functions of type no greater than a, normed by conditions at the points $i\kappa_{2(j-1)}$, where $\kappa_{2(j-1)} = \sqrt{E_{2n} - E_{2(j-1)}}$:

$$f(i\kappa_{2(j-1)}) = (-1)^{j-1},\ j = 1, 2, \ldots, n,$$
$$f(0) = (-1)^n,\ f'(0) = 0. \tag{3.11}$$

Note: From the symmetry of the normalizing conditions (3.11) about the imaginary axis it follows that together with the solution $u(q)$ of the extremal problem, $u(-q)$ and $\frac{1}{2}[u(q)+u(-q)]$ will also be solutions, which is possible only if $u(q)$ is an even function. Therefore, although under the substitution $E=q^2+E_{2n}$ the family $\mathcal{M}$ goes over into the family of even functions of q, the extremal problem can be treated on the family of all functions.

One could also not transform to the q plane, and treat the extremal problems in the plane with cut $[E_{2n}, \infty)$. Theorems A and B remain valid, but the definitions are modified as follows. Let $f(E)$ be a function that is real on $[E_{2n}, \infty)$ and analytic in the plane with the cut.

By $\mathcal{M}_f$ we again mean the set of all roots of the equation

$$f^2(E)-L_f^2=0, \quad L_f=\sup|f(E)|. \tag{3.12}$$

By $\mathcal{R}_f$ we mean the set of roots of Eq. (3.12) that lie on the cut $[E_{2n}, \infty)$, i.e., each root is considered on both edges of the cut. $T\sinh f$ is any maximal subset of the $\mathcal{R}_f$, in which $f(E)$ alternately takes the values $\pm L_f$. $\{\alpha\}$ and $\{\alpha'\}$ are systems of intervals into which the set $\mathcal{R}_f$ and correspondingly $T\sinh f$ subdivide the cut $[E_{2n}, \infty)$. The defect index I_f of the function $f(E)$ is given by the equation

$$I_f=\sum_{\alpha}[N_\alpha/2]+r=\sum_{\alpha'}[N_{\alpha'}/2]+r, \tag{3.13}$$

where r is the number of zeros of the function $f(E)$ outside the cut $[E_{2n}, \infty)$. The set

$$\mu_f\{E:\mathrm{Im}f=0, \ |F|\leq L_f\} \tag{3.14}$$

contains the cut $[E_{2n}, \infty)$ and is considered to be a cut in the plane. The set μ_f supplemented by the cuts drawn through the zeros of the derivative of the function (0.8), as described in the Introduction, is the boundary of a simply connected region $\mathcal{G}_f$, called the fundamental region.

The entire functions with finite defect index and finite deviation $L_f<\infty$ coincide with the class of functions that are representable in the form

$$f(E)=L_f\cos\varphi(E), \tag{3.15}$$

where $\varphi(E)$ is the hyperelliptic integral

$$\varphi(E)=a\int^E \frac{P_I(E)}{[Q_{2I}(E)]^{1/2}}\frac{dE}{(E-E_{2n})^{1/2}}, \quad I=I_f, \tag{3.16}$$

in the region $\mathcal{G}_f$. The periods of the integral (3.16) are multiples of $2a^{-1}\pi$.

The number of points of the set $\mathcal{M}_f$ that do not belong to $T\sinh f$ is $2I_f$.

Theorems A and B cannot be applied immediately to the n-band extremal problem since one of the nodes, namely E_{2n}, coincides with a point of maximum deviation, but the results given enable us to prove the following theorem which we need:

Theorem C: Let $F(E)$ be a solution of the extremal problem for the family $\mathcal{M}(a;E_{0,2n})$ and suppose that in some left semineighborhood of the point E_{2n}, $|F(E)|>1$. Then $L_F=1$, and $F(E)$ is a solution of the n-band extremal problem.

Proof: We use the variable $q: F(E)=F(q^2+E_{2n})$. We assume that $L_F>1$; then the point $q=0$ is an interior point of the set μ_F, and this set contains a cut along the the imaginary axis up to the point where $F(q^2+E_{2n})=L_F$. From the monotonic behavior of $F(q^2+E_{2n})$ along the arc of the set μ_F between two neighboring points of maximum deviation, it follows that at the two nearest points of maximum deviation on μ_F to the left and right of the point $q=0$

$$F(q^2+E_{2n})=-L_F F(E_{2n}). \tag{3.17}$$

From (3.17) it follows that on arcs of μ_F between the point $q=0$ and these points there is one zero of the function $F(q^2+E_{2n})$. If these zeros are complex, their contribution to I_F is 2, while if they are real it is 1. From (3.11) it follows that $F(q^2+E_{2n})$ has no fewer than n pure imaginary roots, and $I_F\geq n+1$. Since $L_F>1$, Theorem B is applicable ($k=q$), and the function $F(q^2+E_{2n})$, contrary to assumption cannot be extremal.

For a nondegenerate solution of the n-band extremal problem, E_{2n} is the right end of a forbidden band, at which $|F|>1$; therefore according to Theorem C, the necessary and sufficient condition for a solution of the extremal problem for the family $\mathcal{M}(a;E_{0,2n})$ to be a solution of the n-band extremal problem is that there be a left semineighborhood of the point E_{2n} in which $|F(E)|>1$. According to the solution of the ISP, this is simultaneously a necessary and sufficient condition for $E_0, E_2, \ldots, E_{2n}$ to be the boundaries of gaps (one for each gap) for the n-band Schrödinger operator $H[u_a]$. It is convenient to call the point in the n-dimensional space $\{z_2^2, z_4^2, \ldots, z_{2n}^2\}=\{z_{2j}^2\}$,

$$z_\nu^2=a^2(E_\nu-E_0), \quad z_\nu^2=a^2q_\nu^2+a^2(E_{2n}-E_0), \tag{3.18}$$

and admissible point. In the following section we describe the set of admissible points.

4. ASYMPTOTIC SOLUTION OF THE n-BAND EXTREMAL PROBLEM AND THE SET OF ADMISSIBLE POINTS

The parameters $\{l_k\}$ of the canonical domain T and the parameters $\{z_{2k}^2\}$, $\{z_{2k-1}^2\}$, $\{\xi_k^2\}$, and the function $\hat{p}(z)$ which maps the upper half-plane $\mathrm{Im}z>0$ on T, are related by $4n$ relations, of which $3n$ are independent:

$$\hat{p}(z)\Big|_{z_{2(j-1)}}^{z_{2j-1}}=\pi, \quad \hat{p}(z)\Big|_{z_{2j-1}}^{z_{2j}}=0, \tag{4.1}$$

$$\begin{aligned} il_j&=\int_{z_{2j-1}}^{\xi_j} dz\prod_{k=1}^{n}\frac{z^2-\xi_k^2}{[(z^2-z_{2k-1}^2)(z^2-z_{2k}^2)]^{1/2}}, \\ il_j&=\int_{\xi_j}^{z_{2j}} dz\prod_{k=1}^{n}. \end{aligned} \tag{4.2}$$

It is sometimes convenient to use the energy variable with the mapping function $t=ap(E-E_0)$. The relations then take the form

$$ap(E-E_0)\Big|_{E_{2(j-1)}}^{E_{2j-1}}=\pi, \quad ap(E-E_0)\Big|_{E_{2j-1}}^{E_{2j}}=0, \tag{4.1'}$$

$$-il_j=a\int_{E_{2j-1}}^{\xi_j}\frac{dE}{2(E-E_0)^{1/2}}\prod_{k=1}^{n}\frac{E-\xi_k}{[(E-E_{2k-1})(E-E_{2k})]^{1/2}},$$

$$il_j = a\int_{\xi_j}^{E_{2j}} \frac{dE}{2(E-E_0)^{1/2}} \prod_{k=1}^{n}. \tag{4.2'}$$

The relations are written in a form that does not assume the inequalities $E_{2j-1} < E_{2j}$, only the inequalities

$$E_0 < E_2 < \circ\cdots < E_{2n}, \quad E_{2n-1} \leq E_{2n}, \tag{4.3}$$

and the fact that the points ξ_k belong to the corresponding gaps is assumed.

In place of the dimensionless parameters z_k^2 and ζ_k^2 it is sometimes convenient to use dimensionless parameters $a^2q_k^2$ and $a^2\xi_k^2$:

$$a^2q_k^2 = a^2(E_k - E_{2n}), \quad a^2q_k^2 = (z_k^2 - z_{2n}^2),$$
$$a^2\hat{\xi}_k^2 = (\xi_k^2 - z_{2n}^2). \tag{4.4}$$

In all cases there are $4n$ essential parameters related by $3n$ independent relations, so that we can assign no more than n parameters.

The system of parameters $\{l_k\}$, i.e., the gaps in T, can be assigned arbitrarily for any $l_k \geq 0$. Because of the uniqueness of the conformal mapping [with $\hat{p}(0) = 0$ and $\hat{p}(z) \cdot z$ at infinity] the other parameters are uniquely determined. The situation is different for the systems $\{z_{2k}^2\}$ and $\{\xi_k^2\}$. The closed set O_n of admissible points $\{z_{2k}^2\}$ is, as we know, the set of points for which the n-band extremal problem has a solution.

If we are given the n parameters $\{z_{2k}^2\}$ or, what is the same thing, the quantities $\{a^2q_{2k}^2\}$, then the remaining parameters must be found from Eqs. (4.1)–(4.2) or (4.1′)–(4.2′). But does there exist a real solution of this system, is it unique, and does it satisfy the inequality $E_{2n-1} \leq E_{2n}$? One can easily answer all these questions in the asymptotic case, when

$$z_{2k}^2 - z_{2k-1}^2 \gg 1, \quad k = 1, 2, \ldots, n. \tag{4.5}$$

This is the situation, for example, for fixed values $E_{0,2n}$ and increasing a. Under the condition (4.5), independent of whether the given quantities are the $\{a^2q_{2k}^2\}$ or $\{a^2\hat{\zeta}_k^2\}$, the system has a unique solution with the following asymptotic behavior:

$$l_j \approx a|q_{2(j-1)}|, \quad a^2\left(\xi_j - \frac{E_{2(j-1)} + E_{2j-1}}{2}\right) \approx 2a|q_{2(j-1)}|,$$
$$\ln a^2(E_{2j-1} - E_{2(j-1)}) \approx -a|q_{2(j-1)}| + \ln a|q_{2(j-1)}| + \ln 2. \tag{4.6}$$

The last of the relations (4.6) can be replaced by

$$\ln a(|q_{2(j-1)}| - |q_{2j-1}|) \approx -a|q_{2(j-1)}|. \tag{4.7}$$

From this asymptotic estimate it follows that the allowed bands contract to points according to an exponential law, while the zeros of the derivative $F'[a^2(E-E_0)]$ tend toward the left ends of the forbidden bands; for any fixed $E \neq \xi_j$, $E_{2(j-1)}$, E_{2j-1},

$$\frac{E - \xi_j}{[(E - E_{2(j-1)})(E - E_{2j-1})]^{1/2}} \sim 1. \tag{4.8}$$

This last result enables us to verify the validity of the asymptotic estimate (4.6), since, according to (1.22), the integrand for $ap(E-E_0)$ consists of a product of expressions (4.8) for different j with $(E_{2n}-E)^{-1/2}$.

From (1.22), (4.8) and the first relation of (4.1′) it follows that

$$\frac{a}{2(E_{2n} - \check{E}_j)^{1/2}} \int_{E_{2(j-1)}}^{E_{2j-1}} \frac{(\xi_j - E)\,dE}{[(E - E_{2(j-1)})(E - E_{2j-1})]^{1/2}} \sim \pi,$$
$$\check{E}_j \in (E_{2(j-1)}, E_{2j-1}),$$

or

$$a\left(\xi_j - \frac{E_{2(j-1)} + E_{2j-1}}{2}\right) \sim \pi \circ 2(E_{2n} - E_{2(j-1)})^{1/2},$$

i.e., we get the second relation in (4.6).

From (1.22), (4.8), and (4.1′)–(4.2′), we find

$$l_j \sim a\int_{\xi_j}^{\xi_j + \alpha_j} \frac{dE}{2(E_{2n}-E)^{1/2}} \frac{E - \xi_j}{[(E - E_{2(j-1)})(E - E_{2j-1})]^{1/2}}$$
$$+ a\int_{\xi_j + \alpha_j}^{E_{2j} + \alpha_{j+1}} \frac{dE}{2(E_{2n}-E)^{1/2}}$$
$$+ a\int_{E_{2j} - \alpha_{j+1}}^{E_{2j}} \frac{dE}{(E_{2n}-E)^{1/2}} \frac{\xi_{j+1} - E}{[(E_{2j} - E)(E_{2j+1} - E)]^{1/2}}. \tag{4.9}$$

The quantities α_j are chosen so small that in the first and third integrals the factors $(E_{2n}-E)^{-1/2}$ can be taken out from under the integral sign, while the parameters (4.5) are so large that (4.8) is valid. For example, we can take $\alpha_j \approx \nu a^{-1}|q_{2(j-1)}|$, where $1 \ll \nu \ll a|q_{2(n-1)}|$. We then get, from (4.9),

$$l_j \sim a(|q_{2(j-1)}| - |q_{2j-1}|) - \ln a^2(E_{2j+1} - E_{2j}) + \ln 2aq_{2j}, \quad j < n. \tag{4.10}$$

For $j = n$ the integration interval (ξ_n, E_{2n}) needs to be split into just two intervals $(\xi_n, \xi_n + \alpha_n)$ and $(\xi_n + \alpha_n, E_{2n})$, and

$$\ln \sim a|q_{2(n-1)}|. \tag{4.11}$$

From (4.2) and (4.8) there follows

$$-\ln a^2(E_{2j-1} - E_{2(j-1)}) + \ln 2aq_{2j} \sim l_j$$

or

$$-\ln a(|q_{2(j-1)}| - |q_{2j-1}|) \sim l_j. \tag{4.12}$$

Substituting in (4.10), we get the recursion relation

$$l_j \sim l_{j+1} + a(|q_{2(j-1)}| - |q_{2j}|), \quad j < n. \tag{4.13}$$

Taking (4.11) into account, we get

$$l_j \sim a|q_{2(j-1)}|. \tag{4.14}$$

Substituting this value of l_j into (4.12), we obtain the relation (4.7) or the last relation of (4.6).

The solution obtained satisfies the inequalities

$$E_{2j-1} < \xi_j < E_{2j}. \tag{4.15}$$

The uniqueness of the solution follows from the fact that when the conditions (4.5) are satisfied, the Jacobian of the system is different from zero.

The asymptotic solution explains the mechanism of the transition from the operators $H[u_a(x)]$ with periodic potentials $u_a(x)$ to the operator $H[u(x)]$ with rapidly decreasing potential and discrete negative

$$q_0^2 < q_2^2 < \cdots < q_{2(n-1)}^2 < 0, \quad E_{2n} = 0. \tag{4.16}$$

In fact, suppose that the quantities (4.16) are fixed and that $a \to \infty$. For $|aq_{2(n-1)}| \gg 1$, the solution with the asymptotic form (4.6) determines the functions

$ap(q^2-E_0)^{1/2}$ and $F[a^2(q^2-E_0)]=\cos ap(q^2-E_0)^{1/2}$. From the solution of the ISP we know that to a given function $F[a^2(q^2-E_0)]$ there corresponds an infinite number of potentials $u_a(x)$ with period a. The potentials $u_a(x)$ can be chosen so that for $a\to\infty$ they converge to the function $u(x)$ which goes to zero sufficiently rapidly as $x\to\infty$. It follows from (4.6) that the operator $H[u(x)]$ has the discrete spectrum (4.16) and the continuous spectrum $[0,\infty)$.

For any ordered fixed system of $n+1$ nodes $E_{0,2n}$, starting from some sufficiently large a, then n-band extremal problem has a unique solution: the point $\{z_{2j}^2\}$ belongs to the set O_n, and the remaining $2n$ parameters $\{z_{2j-1}^2\}$, $\{\xi_j^2\}$, the function $\hat{p}(z)$ and the n parameters $\{l_k\}$ of the canonical domain T are uniquely determined. The inequalities

$$E_{2(k-1)}<E_{2k-1}<\xi_k<E_{2k} \tag{4.17}$$

are satisfied.

Starting from here, we reduce the period a. Then the parameters E_{2j-1} and ξ_j change ($E_{0,2n}$ are fixed), but, according to Theorem C, as long as $E_{2n-1}<E_{2n}$ the n-band extremal problem has a solution with $L=1$.

For sufficiently large a the system of analytic equations (4.1)–(4.2) has a unique solution $\{z_{2k-1}^2\}$, $\{\xi_k^2\}$, and since such solutions can neither appear nor disappear, for any $a>0$ there exists a (real) unique solution of the system. This system is, however, related to the extremal problem only so long as $L(a;E_{0,2n})=1$.

As a decreases, the inequalities (4.17) may be violated, namely E_{2k-1} as it increases may pass the value of E_{2k}, but the inequalities

$$E_{2(k-1)}<E_{2k-1}<E_{2(k+1)} \tag{4.18}$$

remain valid, as follows from the equations

$$\begin{aligned}&F(z_{2(k-1)}^2)=(-1)^{k-1},\\ &F(z_{2k-1}^2)=(-1)^k,\quad F(z_{2(k+1)}^2)=(-1)^{k+1}.\end{aligned} \tag{4.19}$$

Violation of the inequalities (4.17) for $k<n$ simply means that E_{2k} has become the left end, and E_{2k-1} the right end of the kth gap, while nothing is altered in the structure of the set $\mu(E:\mathrm{Im}F=0;|F|\le 1)$.

The picture is different if we reach a value of a^2 (we denote this value and the corresponding point $\{z_{2k}^2\}$ by a_1^2 and $\{z_{2k}^2\}_1$) for which

$$E_{2n-1}=\xi_n=E_{2n}. \tag{4.20}$$

Then the semi-infinite allowed band $[E_{2n},\infty)$ goes over into the semi-infinite allowed band $[E_{2(n-1)}^{(1)},\infty)$, where

$$E_{2(n-1)}^{(1)}=\max\{E_{2(n-1)},E_{2n-3}(a_1^2)\}. \tag{4.21}$$

With further decrease of a^2, the minimum deviation $L(a;E_{0,2n})$ on the comparison set $\mathcal{M}(a;E_{0,2n})$ becomes greater than 1, the points $E_0, E_2, \ldots, E_{2n}$ are no longer on the boundary of the set $\mu(E:\mathrm{Im}F=0;|F|\le L_F)$ and the set μ may move out into the complex plane and change its structure.

We want to show that the entire boundary ∂O_n can be irradiated from the origin, i.e., that any half-line, emerging from the origin and lying in the angle

$$0\le z_2^2\le z_4^2\le\cdots\le z_{2n}^2 \tag{4.22}$$

intersects ∂O_n in one point. We note that for $E_{2n-1}=E_{2n}$, $a=a_1$, the solution of the n-band extremal problem is also a solution of the extremal problem for the comparison set $\mathcal{M}'(a';E_{0,2n})$, which differs from the set $\mathcal{M}(a;E_{0,2n})$, which differs from the set $\mathcal{M}(a;E_{0,2n})$ in having an additional condition

$$f'(E_{2n})=0. \tag{4.23}$$

The assertion that ∂O_n can be irradiated from the origin follows from the theorem:

Theorem D: If $u(q^2)$ is a function of type a from the family $\mathcal{M}'(a;E_{0,2n})$ with defect index $I_a\le n$, then $u(q^2)$ is a unique extremal function.

The proof is analogous to proof 4 of Ref. 7, so without further explanation we shall use the necessary terms and notation from Ref. 7.

Let $f(q^2)$ be a function of the set $\mathcal{M}'(a;E_{0,2n})$ and $L_f\le L_a$; then, for sufficiently large $c>0$,

$$\begin{aligned}&|f(q^2)|\le|\hat{w}(q^2;c)|,\\ &\hat{w}(q^2;c)=u(q^2)+icv(q^2),\quad -\infty<q<\infty\end{aligned} \tag{4.24}$$

where

$$\hat{w}(z;1)=\hat{w}(z)$$

is the associated function. The equality sign in (4.24) holds only at the point $q=0$, at which $f(0)=\hat{w}(0)=(-1)^{n+1}$. The number of zeros of the function $\hat{w}(q^2;c)$ in the upper half-plane $\mathrm{Im}q>0$ coincides with the defect index of the function $u(q^2)$ and cannot exceed n. From the inequality (4.24) it follows that the number of roots of the function

$$\hat{w}(q^2;c)-f(q^2)=[u(q^2)-f(q^2)]+icv(q^2) \tag{4.25}$$

in the half-plane $\mathrm{Im}q>0$ does not exceed the number of roots of the function $\hat{w}(q^2;c)$ in the same half-plane, i.e., n. On the other hand, this number is not less than

$$\int_\infty v(q^2)/[u(q^2)-f(q^2)]\ge n+1. \tag{4.26}$$

The inequality (4.26) follows from the fact that $[u(q^2)-f(q^2)]/v(q^2)$ has n complex zeros in the region $\mathrm{Im}q>0$, and a third-order zero at $q=0$.

From Theorem D it follows that, for $a_1<a_1$, the segment $z_{2k}^2=a^2(E_{2k}-E_0)$, $0\le a\le a_1$ does not intersect ∂O_n. In fact, in the contrary case, for some $a'<a_1$, $E_{2n-1}(a')=E_{2n}$, and a solution will exist for the n-band extremal problem for $a=a'<a_1$, which contradicts Theorem D.

From Theorem D it also follows that the extremal problem for the comparison set $\mathcal{M}(\check{a};E_{0,2n})$ has a continuum of solutions if the point $\{\check{z}_{2k}^2\}$, $\check{z}_{2k}^2=\check{a}^2(E_{2k}-E_0)$ is an interior point of O_n, namely one solution of type a for any a in the interval $(a_1,\check{a})$. The points of the boundary ∂O_n are characterized by the fact that for them the n-band extremal problem has precisely one solution in the family $\mathcal{M}(a_1;E_{0,2n}\}$.

Since each ray emerging from the origin is determined by the ratios z_{2k}^2/z_{2n}^2, $k=1,\ldots,n-1$, and since ∂O_n intersects each ray in the angle (4.23) in one point, the set O_n is completely characterized by an inequality of the form

$$z_{2n}^2 \geqslant f_n\left(\frac{z_2^2}{z_{2n}^2}, \ldots, \frac{z_{2(n-1)}^2}{z_{2n}^2}\right). \tag{4.27}$$

Since on the boundary of the set $E_{2n-1}=E_{2n}$,

$$a^2(E_{2n-1}-E_0)=f_n\left(\frac{E_2-E_0}{E_{2n}-E_0}, \ldots, \frac{E_{2(n-1)}-E_0}{E_{2n}-E_0}\right). \tag{4.28}$$

The relation (3.28) is valid for all points of O_n. On ∂O_n, (4.28) goes over into the equation

$$a_1^2(E_{2n}-E_0)=f_n\left(\frac{E_2-E_0}{E_{2n}-E_0}, \ldots, \frac{E_{2(n-1)}-E_0}{E_{2n}-E_0}\right). \tag{4.29}$$

The boundary of the angle (4.22) is the asymptotic cone for the set O_n. This follows from the fact that in any bounded domain of the angle $z_{2n}^2 < R^2$, the differences $z_{2j}^2 - z_{2(j-1)}^2$ for all points of O_n are bounded below by a constant $d(R)$. Finally, the set O_n lies to the right of the hyperplane $z_{2n}^2 = n^2\pi^2$, and, as we have shown, has only one point in common with this hyperplane, $\{z_{2j}^2 = j^2\pi^2\}$, which it is natural to call a vertex of the set O_n. To a vertex of the set O_n there corresponds $F[a^2(E-E_0)] = \cos a\,(E-E_0)^{1/2}$, and $u(x)=E_0$.

The transition from an interior point $\{z_{2j}^2\}$ of the set O_n to a boundary point $\{z_{2j}^2\}_1$ on the same ray corresponds to the transition from the n-band problem to the $(n-1)$-band problem. We need only recall that to maintain the conditions of the extremal problem the farthest right node $E_{2(n-1)}^{(1)}$ is given by (4.21). This means that the projection of ∂O_n on the $(n-1)$-dimensional subspace $\{z_2^2, \ldots, z_{2(n-1)}^2\}$ is a set O_{n-1}, and we can use the contraction method. As the point $\{z_{2k}^2\}$ moves along its ray in the set O_{n-1}, the value of E_{2n} changes, since E_{2n} is the first point of maximum deviation of the solution of the n-band extremal problem on $[E_{2(n-1)}^{(1)}, \infty)$.

After n steps we contract to the null-dimensional set O_0, consisting of a single point to which, on the boundary of O_n, there corresponds a vertex of the set O_n.

5. EXPANSION IN EIGENFUNCTIONS OF THE SCHRÖDINGER OPERATOR WITH A PERIODIC POTENTIAL

In this section we transform the results of Titchmarsh[4] and then change from the variable E to the variable p. The final formulas are quite close to classical formulas for Fourier transformations, which eliminates the need to keep in our proof the main results of the theory of the Fourier integral.

In this section it is convenient to work with the dimensionless variables

$$\hat{x}=a^{-1}x, \quad \hat{E}=a^2E, \quad \hat{u}(\hat{x})=a^2u(\hat{x}), \quad \hat{u}(\hat{x}+1)=\hat{u}(\hat{x}), \quad ap=\hat{p}. \tag{5.1}$$

In addition, so long as there is no danger of a misunderstanding, we shall denote these dimensionless quantities by the same symbols as before, x, E, etc.

We recall some standard notation. We denote by $\theta(x;x_0;E)$ and $\varphi(x;x_0;E)$ a basis with initial conditions

$$\begin{aligned}&\theta(x_0;x_0;E)=1, \quad \theta'(x_0;x_0;E)=0,\\ &\varphi(x_0;x_0;E)=0, \quad \varphi'(x_0;x_0;E)=1.\end{aligned} \tag{5.2}$$

$\theta(x;x_0;E)$ and $\varphi(x;x_0;E)$ are entire functions of the spectral variable E. The values of these functions and their derivatives at $x=x_0+1$ are denoted by $\theta(E)$, $\theta'(E)$, $\varphi(E)$, $\varphi'(E)$ or simply $\theta, \theta', \varphi, \varphi'$. In this basis the shift matrix is

$$T(E)=\begin{pmatrix}\theta & \theta'\\ \varphi & \varphi'\end{pmatrix}. \tag{5.3}$$

In particular,

$$2F(E-E_0)=\theta+\varphi', \quad \theta\varphi'-\theta'\varphi=1. \tag{5.4}$$

As another basis we choose the Bloch functions $\psi_\pm(x;x_0;E)$ functions of the shift matrix:

$$\begin{aligned}&\varphi_\pm(x+1;x_0;E)=\rho_\pm(E)\,\psi_\pm(x;x_0;E),\\ &\rho_\pm(E)=\exp[\pm ip(E)],\end{aligned} \tag{5.5}$$

normalized by the equalities $\psi_\pm(x_0;x_0;E)=1$. As is well known,

$$\psi_\pm(x;x_0;E)=\exp[ixp(E)]\,\chi_\pm(x;x_0;E), \tag{5.6}$$

where $\chi_\pm(x;x_0;E)$ are functions with period 1. The basis $\psi_\pm(x;E)$ is related as follows to the basis $\theta(x;E)$, $\varphi(x;E)$ [4]:

$$\psi_\pm(x;x_0;E)=\theta(x;x_0;E)+\frac{\rho_\pm(E)-\theta(E)}{\varphi(E)}\,\varphi(x;x_0;E), \tag{5.7}$$

$$W(\psi_+,\psi_-)=\frac{\rho_-(E)-\rho_+(E)}{\varphi(E)}. \tag{5.8}$$

Let $f(x)\in L(-\infty,\infty)$ have bounded variation at the point x. For such a function Titchmarsh (Ref. 4, formula 21.6.3) obtained an expansion in the basis $\theta(x;E)$, $\varphi(x;E)$. This formula is easily brought to the form

$$\begin{aligned}f(x)=\frac{1}{4\pi}\int_\mu \frac{d(E-E_0)}{[1-F^2(E-E_0)]^{1/2}}\Bigg\{&\theta(x;x_0;E)\int_{-\infty}^{\infty}dy\,f(y)\\ &\times[\varphi(y+1;x_0;E)-F(E-E_0)\,\varphi(x;x_0;E)]-\varphi(x;x_0;E)\\ &\times\int_{-\infty}^{\infty}dy\,f(y)[\theta(y+1;x_0;E)-F(E-E_0)\,\theta(y;x_0;E)]\Bigg\}.\end{aligned} \tag{5.9}$$

The inner integral in (5.9) goes over the set μ as cut, the function $[1-F^2(E-E_0)]^{1/2}$ is single valued in $C_1\setminus\mu$ and is defined by the usual condition $(1-F^2)^{1/2}\sim iF$ for $E\to-\infty$.

Let

$$\begin{pmatrix}\theta(x;E)\\ \varphi(x;E)\end{pmatrix}=C\begin{pmatrix}\psi_+(x;E)\\ \psi_-(x;E)\end{pmatrix}, \quad \det C=W^{-1}(\psi_+,\psi_-). \tag{5.10}$$

We express $\theta(x;E)$ and $\varphi(x;E)$ in (5.9) in terms of $\psi_\pm(x;E)$ and use the fact that $\psi_\pm(x+1;E)=\rho_\pm(E)\,\psi_\pm(x;E)$ and $W(\psi_+,\psi_-)=[\rho_+-\rho_-]/\varphi(E)$. As a result we get the formula

$$\begin{aligned}f(x)=\frac{1}{8\pi}\int_\mu \frac{\varphi(E)\,dE}{[1-F^2(E-E_0)]^{1/2}}\\ \times\left\{\begin{aligned}&\psi_+(x;E)\int_{-\infty}^{\infty}dy\,f(y)\,\psi_-(y;E)\\ &+\psi_-(x;E)\int_{-\infty}^{\infty}dy\,f(y)\,\psi_+(y;E)\end{aligned}\right\}.\end{aligned} \tag{5.11}$$

We recall that formula 21.6.3 in Ref. 4 and, thus, also (5.9)–(5.11) are obtained as the result of integration over the cut $[E_0,\infty)$, and the cut $[E_0,\infty)$ has been replaced by a cut over the set μ because the integrals

over the upper and lower edges of the gaps cancel one another.

We see from (5.7) that $\psi_\pm(x;E)$ in (5.11) are actually the values on the upper and lower edges of the cuts for the same function $\psi(x;E)$, which is meromorphic on the Riemann surface of the function $[1-F^2(E)]^{1/2}$. From (5.6) it follows that $\chi_\pm(x;E)$ in turn are the values along the edges of the cuts for one and the same function $\chi(x;E)$, meromorphic on the Riemann surface. The poles of the functions $\psi(x;E)$ and $\chi(x;E)$ are points of the Riemann surface whose projections on the E plane coincide with the zeros of the function $\varphi(E)$. The zeros of $\varphi(E)$ are known to lie one in each gap of the Bloch spectrum.

From the derivation of formula 21.6.3 it follows that $\psi_-(x;E)$ is obtained from $\psi_+(x;E)$ by analytic continuation in the fundamental region $\mathcal{G}_0$, which the function $t=p(E-E_0)$ maps conformally onto the canonical domain T. The spectrum μ is then mapped on the whole real axis, the points $E+i0$ onto the opposite edges of μ—into the points $\pm p$, the upper and lower edges of the gap $[E_{2k-1}, E_{2k}]$ are mapped into the cuts $S_{\pm k}$, whose vertices are images of the zeros $\xi_k \pm i0$ of the derivative $F'(E-E_0)$. In transforming from E to the variable p, the inversion formula (5.11) goes over into the inversion formulas

$$\tilde f(p)=\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty} dx\,\psi(x;E(p))f(x),$$
$$f(x)=\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}\frac{-\varphi(E(p))}{2F'[E(p)-E_0]}\psi(x;E(p))\tilde f(-p). \tag{5.12}$$

These formulas are obviously equivalent to the formula

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} dp\,\frac{-\varphi(E(p))}{2F'[E(p)-E_0]}\psi(x;E(p))\psi(y;E(-p))=\delta(x-y). \tag{5.13}$$

The zeros of the denominator $F'(E(p)-E_0)$ lie at the vertices of the cuts $S_{\pm k}$, i.e., inside the region of integration in (5.12)—(5.13). The zeros of $\varphi(E(p))$ lie on the cuts $S_{\pm k}$, i.e., they are also inside the region of integration. In the region of integration,[4]

$$-\varphi(E(p))/2F'(E(p)-E_0)>0. \tag{5.14}$$

If we introduce Bloch functions normalized not by the condition $\psi_\pm(x_0;x_0;E)=1$, but by the condition

$$\psi_\pm(x_0;x_0;E)=[-\varphi(E)/2F'(E(p)-E_0)]^{1/2}, \tag{5.15}$$

formulas (5.12)—(5.13) take the form

$$\check f(p)=\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty} dx\,\check\psi(x;E(p))f(x),$$
$$f(x)=\frac{1}{(2p)^{1/2}}\int_{-\infty}^{\infty} dp\,\psi(x;E(p))\tilde f(-p), \tag{5.16}$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} dp\,\hat\psi(x;E(p))\check\psi(y;E(p))=\delta(x-y),$$
$$\frac{1}{2\pi}\int_{-\infty}^{\infty} dx\,\check\psi(x;E(p))\check\psi(x;E(q))=\delta(q-p). \tag{5.17}$$

From formula (5.17) we get the Parseval formula

$$\int_{-\infty}^{\infty} dx f^2(x)=\int_{-\infty}^{\infty} dp\tilde f(p)\tilde f(-p),$$
$$\int_{-\infty}^{\infty} dx f(x)g(x)=\int_{-\infty}^{\infty} dp\,[\tilde f(p)\tilde g(-p)-\tilde f(-p)\tilde g(p)]. \tag{5.18}$$

From (5.7) it follows that $\psi(x;E)$ is real along all the gaps, so that for real p,

$$\check\psi(x;E(p))=\overline{\check\psi}(x;E(p)), \tag{5.19}$$

and for a real function $f(x)$, its image $\check f(p)$ has conjugate values at the points p and $-p$. If $f(x)=f_1(x)+if_2(x)$ is a complex-valued function, then, taking $g(x)=f_1(x)-if_2(x)$, we obtain from (5.18) the Parseval formula in its usual form:

$$\int_{-\infty}^{\infty} dx\,|f(x)|^2=\int_{-\infty}^{\infty} dp\,|\check f(p)|^2. \tag{5.20}$$

In formulas (5.12)—(5.17), at all p that are multiples of π', the function $E(p)$ under the integral sign has jumps equal to the length of the corresponding gap, so that $\tilde f(p)$ and $\check f(p)$ have discontinuities of the first kind at points that are multiples of π.

As $p\to\pm\infty$, the ratio $-\varphi(E(p))/2F'(E(p)-E_0)$ is bounded, since the complex zeros of the denominator are located further from the real axis than the zeros of the numerator. Furthermore, the normalizing factor (5.14) tends to 1, as long as p does not approach too close to values that are multiples of π. We recall that for $E\to\infty$

$$\varphi(E)\sim\frac{\sin(E-E_0)}{(E-E_0)^{1/2}},\quad F'(E-E_0)\sim-\frac{\sin(E-E_0)^{1/2}}{2(E-E_0)^{1/2}},$$
$$\psi(x;E(p))\sim\exp(ipx). \tag{5.21}$$

From this and formulas (5.16)—(5.20) it follows that the general theory of the Fourier integral is valid for the transformation (5.16).

6. EXPANSION IN EIGENFUNCTIONS OF A NON-SELF-ADJOINT SCHRÖDINGER OPERATOR WITH COMPLEX PERIODIC POTENTIAL

As in Sec. 5, we use dimensionless variables, but the periodic potential $u(x)=u(x+1)$ is assumed to be complex, as is the case in a medium with energy absorption.

The determination of the Bloch spectrum is independent of whether the potential $u(x)$ is real or complex valued. Just as for a real potential, the Bloch spectrum is the set of points of the spectral plane E for which all solutions of the equation $(H[u]-E)\psi=0$ are bounded on the whole real axis x.

All the definitions and derivations of Sec. 1 up to formula (1.16) inclusive are independent of whether the potential $u(x)$ is real or complex-valued. Also independent of this are, in particular, the rough asymptotic estimate $F(E-E_0)\sim\cos(E-E_0)^{1/2}$ and the asymptotic for $\psi_\pm(x;E)$ in terms of E. The last two statements follow from the derivation of the asymptotic estimate (cf. Ref. 4).

The boundedness of all solutions of the equation $(H[u]-E)\psi=0$ is equivalent to the equality $|\rho_\pm(E)|=1$ or to the reality of the quasimomentum $p(E-E_0)$, i.e., just as in the real case, the Bloch spectrum consists of the interior points of the linear set

$$\mu_u[E;\mathrm{Im}F=0;\ |F|\le 1] \tag{6.1}$$

and thus lies on the curve Γ:

$$\Gamma\{E;\mathrm{Im}F=0\}. \tag{6.2}$$

The distinction between the complex case (as we shall for brevity call the problem with a complex-valued potential) and the real case lies in the absence of symmetry, and possible complication of the structure of the Bloch spectrum and the fundamental domain $\mathcal{G}_0$.

The boundaries of the components of μ are the spectra of the periodic and antiperiodic problems, i.e., in the real case they are the spectra of self-adjoint operators and are therefore real. This is not so in the complex case. In the real case the zeros ξ of the derivative $F'(E)$ are real, $|F(\xi)|>1$, and the lines of constant argument (cf. the Introduction)

$$\arg\rho_-(E)=\arg\rho_-(\xi), \tag{6.3}$$

along which $|\rho(E)|$ drops off as we move away from the points ξ toward the intersection with μ, are just the gaps in the Bloch spectrum. In the complex case the lines (6.3) may not belong to the curve Γ. It is also possible to have values of ξ such that $\xi\in\mu$ and $|F(\xi)|<1$.

In Ref. 7 it was shown in detail that the structure of the fundamental domain $\mathcal{G}_0$ (denoted by F in Ref. 7) is completely determined by the parameters

$$l_\nu+i\theta_\nu=\ln\rho_-(\xi_\nu). \tag{6.4}$$

In the real case $\theta_\nu=0\ (\mathrm{mod}\ \pi)$; in the complex case this may not be so.

In the complex as well as in the real case, the function

$$l=p(E)=\pm i\ln\rho_-(E) \tag{6.5}$$

maps the fundamental region $\mathcal{G}_0$ onto the canonical domain T, which is the upper half-plane $\mathrm{Im}\,l>0$ with vertical cuts of length l_ν and bases at the points θ_ν.

The absence of symmetry in $\mathcal{G}_0$ leads to an absence of symmetry in T_u, but nevertheless one important aspect of this symmetry is preserved. If E and E' denote points on opposite edges of a cut forming the boundary of $\mathcal{G}_0$, while p and p' are the corresponding values $p(E)$ and $p(E')$, then

$$p+p'=2k(p)\pi, \tag{6.6}$$

where $k(p)$ is an integer. Equation (6.6) follows from the fact that the functions $\rho_\pm(E)$ are single-valued in $C_1\setminus\mu$, and the image of any point at which $F^2(E)=1$ is a multiple of κ. The integer-valued function $k(p)$ determines the structure of $\mathcal{G}_0$. For example the identity $k(p)=0$ means that $\mathcal{G}_0$ is the plane with a cut along some semi-infinite branch $[E_0,\infty)$ of the curve Γ, the set μ lies wholly on this branch, the lines (6.3) are the gaps along this branch between the components of μ, and all the zeros of F' lie in these gaps. In this case the canonical domain is symmetric with respect to the imaginary axis. The condition for symmetry of T is not the reality of the potential but rather the much less restrictive condition $k(p)=0$.[16] Let E_0 be some root of the equation $F=1$. One can normalize $k(p)$ so that $p(E_0)=p'(E_0)=0$.

Formula 21.6.3 remains valid in the complex case, since its derivation is based entirely on the crude asymptotic estimate for $F(E)$ and $\psi_\pm(x;E)$. Therefore, formula (5.11) is also valid. In the inversion formulas (5.12) and formulas (5.13) we must replace $-p$ by p' from (6.6):

$$\tilde f(p)=\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}dx\,\psi(x;E(p))\,f(x),$$

$$f(x)=\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}dp\,\frac{-\varphi(E(p))}{2F'(E(p)-E_0)}\,\psi(x;E(p))\,\tilde f(p'), \tag{6.7}$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}dp\,\frac{-\varphi(E(p))}{2F'(E(p)-E_0)}\psi(x;E(p))\,\psi(y;E(p'))=\delta(x-y), \tag{6.8}$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}dx\,\frac{-\varphi(E(p))}{2F'(E(p)-E_0)}\,\psi(x;E(p'))\,\psi(x;E(q))=\delta(q+p').$$

Since $F'(E)$ is an entire function, the images of the zeros of this function, if they lie within the region of integration in (6.7)–(6.8), are integrable singularities. For such values of E the normalizing factor in (5.15) goes to ∞ and the definition of the function $\tilde f(p)$ according to (5.16) loses its meaning. Nevertheless, the inversion formula is valid:

$$f(x)=\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}dp\,\check\psi(x;E(p))\,\tilde f(p'). \tag{6.9}$$

Formulas (5.17) also remain valid:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}dx\,\check\psi(x;E(p'))\,\check\psi(x;E(q))=\delta(q+p'),$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}dp\,\check\psi(x;E(p))\,\check\psi(y;E(p))=\delta(x-y). \tag{6.10}$$

The Parseval formulas have the form

$$\int_{-\infty}^{\infty}dx\ f^2(x)=\int_{-\infty}^{+\infty}dp\,\tilde f(p)\tilde f(p'),$$

$$\int_{-\infty}^{\infty}dx\ f(x)g(x)=\tfrac12\int_{-\infty}^{\infty}dp\ [\tilde f(p)\tilde g(p')+\tilde f(p')\tilde g(p)]. \tag{6.11}$$

Because of lack of the symmetry (5.19), the Parseval formula cannot be brought to the form (5.20).

The crude asymptotic estimate (5.21) remains valid in the complex case. From it and the formulas of the present section, it follows that the Fourier integral theory remains essentially valid also for the expansion in eigenfunctions of a non-self-adjoint Schrödinger operator with a complex-valued periodic potential. The main new point of principle is the effect of the structure of the set μ and the fundamental domain $\mathcal{G}_0$.

ACKNOWLEDGMENT

I am happy to express my gratitude to B. M. Levitan for many very helpful discussions.

APPENDIX A: THE PRINCIPLE OF INCREASING FLUX

Let $f(z)=u(z)+iv(z)$ be a single-valued analytic function in the simply connected domain $\mathcal{G}$ with positive imaginary part $v(z)>0$. From the fact that $v(z)$ is positive it follows that it determines a generalized function $v(\xi)$ on the boundary $\partial\mathcal{G}$. This generalized function is

the derivative in the harmonic measure $\omega(z_0;d\xi;\mathcal{G})$ of the increasing function

$$\theta_{z_0}(\xi)=\theta[\omega(z_0;\widehat{\xi_0\xi};\mathcal{G})],$$
$$\theta_{z_0}(z)=q(\xi)+\varphi_{z_0}(\xi),\quad \int_{\partial\mathcal{G}} d\theta_{z_0}(\xi)<\infty. \tag{A1}$$

Here $\omega(z_0;\widehat{\xi_0\xi};\mathcal{G})$ is the harmonic measure of the arc $\widehat{\xi_0\xi}\subset\partial\mathcal{G}$ with respect to the domain $\mathcal{G}$ at the point z_0, while $\theta=q+\varphi$ is the Lebesgue decomposition of a function of bounded variation into its components that are singular and completely continuous in the measure ω. In the measure ω the function $v(z)$ has angular limiting values $v(\xi)$ almost everywhere, and

$$\frac{d\varphi_{z_0}(\xi)}{\omega(z_0;d\xi)}=v(\xi),\quad \frac{dq(\xi)}{\omega(z;d\xi)}=0. \tag{A2}$$

The generalized function $v(\xi)$ is defined on the space $C(\omega)$ of functions $F(\xi)$ continuous in the measure ω on $\mathcal{G}$:

$$(F,v)=\int_{\partial\mathcal{G}} F(\xi)\,d\theta[\omega(z_0;\widehat{\xi_0\xi};\mathcal{G})]. \tag{A3}$$

The function $v(z)$ is representable in the region $\mathcal{G}$ in the form of a Poisson–Stieltjes integral

$$v(z)=(P_{z_0}(z;\xi;\mathcal{G}),v)$$
$$=\int_{\partial\mathcal{G}} P_{z_0}(z;\xi)[\omega(z_0;d\xi)+dq(\xi)], \tag{A4}$$

where $P_{z_0}(z;\xi;\mathcal{G})$ is the Poisson kernel

$$P_{z_0}(z;\xi;\mathcal{G})=\frac{\omega(z;d\xi;\mathcal{G})}{\omega(z_0;d\xi;\mathcal{G})}. \tag{A5}$$

The Schwartz kernel $S_{z_0}(z;\xi;\mathcal{G})$ is defined in terms of the Poisson kernel, as the function analytic in $\mathcal{G}$ with real part equal to $P_{z_0}(z;\xi;\mathcal{G})$:

$$\mathrm{Re}S_{z_0}=P_{z_0}. \tag{A6}$$

The kernel S_{z_0} is determined to within an arbitrary pure imaginary addition $ik(\xi)$. This addition can be chosen so that $S_{z_0}\in C(\omega)$.

The function $f(z)=u(z)+iv(z)$ is representable as a Schwartz–Stieltjes integral with an undetermined real constant C:

$$f(z)=i(S_{z_0}(z;\xi;\mathcal{G}),v)+C, \tag{A7}$$

or in the form

$$f(z)-f(z_1)=2\,([S_{z_0}(z;\xi;\mathcal{G})-S_{z_0}(z_1;\xi;\mathcal{G})],v),\quad z_1\in\mathcal{G}. \tag{A8}$$

Remark: The relation (A8) is a generalization of what is known in physics as a dispersion relation with one subtraction. Formula (A8), as shown above, is valid if the generalized function $v(\xi)$ induced on $\partial\mathcal{G}$ by the harmonic function $v(z)$ is a Stieltjes measure. The latter is known to be the case if and only if at least one of the integrals

$$\int_{\partial\mathcal{G}_n} v^{\pm}(\xi)\,\omega(z_0;d\xi;\mathcal{G}),\quad v(z)=v^{+}(z)-v^{-}(z), \tag{A9}$$

where $\mathcal{G}_n$ is the approximating family of the region $\mathcal{G}$, is uniformly bounded. From the identity

$$\int_{\partial\mathcal{G}_n}[v^{+}(\xi)-v^{-}(\xi)]\,\omega(z_0;d\xi;\mathcal{G})=v(z_0),$$

the second integral is also uniformly bounded. Thus uniform boundedness of the integrals (A9) is a sufficient condition for the function with imaginary part $v(z)$ and analytic in the region $\mathcal{G}$ to satisfy a generalized dispersion relation with one subtraction.

The once-subtracted generalized dispersion relation (A8) reduces to an ordinary dispersion relation when the singular part $q(\xi)$ is identically zero, i.e., when

$$\lim\int_{\partial\mathcal{G}_n} v(\xi)\,\omega(z_0;d\xi;\mathcal{G}_n)=\int_{\partial\mathcal{G}} v(\xi)\,\omega(z_0;d\xi;\mathcal{G}). \tag{A10}$$

Thus the sufficient condition for the function $f(z)=u(z)+iv(z)$ to satisfy a once-subtracted dispersion relation is the uniform boundedness of the integrals (A9) and the equality (A10).

For the half-plane $\mathrm{Im}z>0$,

$$S_i(z;\xi)=\frac{1}{i}\,\frac{\xi z+1}{\xi-z}=\frac{1}{i}\left(\frac{1+\xi^2}{\xi-z}-\xi\right),\quad \omega(i;d\xi)=\frac{1}{\pi}\,\frac{d\xi}{1+\xi^2} \tag{A11}$$

and the generalized dispersion relation with one subtraction is expressed in the form

$$f(z)-f(z_1)=\int_{-\infty}^{\infty}\left(\frac{1}{\xi-z}-\frac{1}{\xi-z_1}\right)v(\xi)\,d\xi+$$
$$+\int_{-\infty}^{\infty}(z-z_1)\,\frac{1+\xi^2}{(\xi-z)(\xi-z_1)}\,dq(\xi). \tag{A12}$$

Let α be some arc of the boundary $\partial\mathcal{G}$ at whose ends $dq(\xi)=0$. The flux of the function $v(z)$ through the arc α is equal to

$$\int_{\alpha}\frac{\partial v(\xi)}{\partial g(z_0;\xi;\mathcal{G})}\,\omega(z_0;d\xi;\mathcal{G})=u(\xi)\big|_{\alpha}, \tag{A13}$$

where $g(z_0;\xi;\mathcal{G})$ is the Green's function for the domain $\mathcal{G}$ with pole at the point z_0, and on the right is the increase in $u(z)$ along the arc α. If $u(z)$ does not have limiting corner values at the ends of the arc, then we must take for the increase of $u(z)$ along the arc α the limit of its increase along arcs α' at whose ends $u(z)$ does have a definite value.

Let γ be an arc lying in $\mathcal{G}$ and, with the same end points as α, $\mathcal{G}_1$ a subdomain of the domain $\mathcal{G}$, bounded by $\partial\mathcal{G}\setminus\alpha$ and γ, and let $v_1(z)$ be a function harmonic in $\mathcal{G}_1$, inducing zero measure of γ. We denote the corresponding analytic function in the region $\mathcal{G}_1$ by $f_1(z)=u_1(z)+iv_1(z)$.

The principle of increasing flux consists in the statement that the flux of the function $v_1(z)$ through the arc ξ_γ is strictly larger than the flux of $v(z)$ through the arc α:

$$u_1(z)\big|_{\gamma}>u(z)\big|_{\alpha}. \tag{A14}$$

Proof: Without loss of generality we may assume that the region $\mathcal{G}_1$ coincides with the upper half-plane $\mathrm{Im}z>0$, and the arc γ with a segment, possible semi-infinite, $[\xi_1,\xi_2]$. The flux of $v(z)$ across the arc α is equal to its flux across the arc γ, so what we must prove is that

$$u_1(z)\big|_{\gamma}>u(z)\big|_{\gamma}. \tag{A15}$$

We apply formula (A12) to the difference $f(z)-f_1(z)$. Since $v(z)-v_1(z)$ vanishes on $\partial\mathcal{G}\setminus\alpha$,

$$(f(z)-f_1(z))\Big|_{z_1}^{z_2}=\frac{1}{\pi}\int_{\zeta_1}^{\zeta_2}\left(\frac{1}{\zeta-z_2}-\frac{1}{\zeta-\zeta_1}\right)v(\zeta)\,d\zeta. \quad \text{(A16)}$$

Let z_2 and z_1 tend toward real points x_2 and x_1, at which $f(z)$ has limiting values. Separating off the real part, we find

$$u(z)\Big|_{x_1}^{x_2}-u_1(z)\Big|_{x_1}^{x_2}=\frac{1}{\pi}\int_{\zeta_1}^{\zeta_2}\frac{x_2-x_1}{(\zeta-x_2)(\zeta-x_1)}\,v(\zeta)\,d\zeta. \quad \text{(A17)}$$

From the positivity of $v(\zeta)$ on γ it follows that for $x_2=\zeta_2$, $x_1=\zeta_1$, the integral (A17) is strictly less than zero, i.e., the inequality (A14) is valid.

If the segment $[x_1, x_2]$ lies entirely outside $[\zeta_1, \zeta_2]$, the integral (A17) is strictly positive, i.e., we have a theorem which we call the *complement of the principle of increasing flux*:

The flux of the function $v_1(z)$ through the arc $\beta\subset\partial\mathcal{G}\setminus\alpha$ is less than the flux of the function $v(z)$ through that arc:

$$u_1(z)\big|_\beta < u(z)\big|_\beta. \quad \text{(A18)}$$

In the formulation of the principle it was assumed that the arc lies wholly, except for its ends, in $\mathcal{G}$. Actually the arc γ can partially coincide with the arc α; the only important requirement is that γ contain a part lying in $\mathcal{G}$; then $v(z)>0$ on a set of positive measure of the segment $[\zeta_1, \zeta_2]$ in (A17). In this case formula (A17) must be replaced, in accordance with (A12), by

$$u(z)\Big|_{x_1}^{x_2}-u_1(z)\Big|_{x_1}^{x_2}=\int_{\zeta_1}^{\zeta_2}\frac{x_2-x_1}{(\zeta-x_2)(\zeta-x_1)}\times\left[v(\zeta)\frac{d\zeta}{\pi}+(1+\zeta^2)\,dq(\zeta)\right]. \quad \text{(A19)}$$

Since $dq(\zeta)\geq 0$, we again find that the principle is valid when $x_2=\zeta_2$, $x_1=\zeta_1$ and that the complementary principle holds when $[x_1, x_2]\cap[\zeta_1, \zeta_2]=\emptyset$.

The physical interpretation of the principle of increasing flux is so simple and convincing that one may regard it as a short proof of the principle. For this purpose we interpret the analytic function $f(z)=u(z)+iv(z)$ as a complex heat potential with isotherms $v(z)=\text{const}$ and heat flux function $u(z)$. The heat flux through an arc is equal to the increase of $u(z)$ along the arc. The principle of increasing flux consists in the statement that if the temperature in the region is positive and if we push some portion of the boundary into the region while maintaining the piece at zero temperature, then the flux of heat through the arc will increase.

The formal proof given above shows that this sort of physical argument is essentially entirely rigorous and can be used.

Referee's note: A characterization of the manifolds of isospectral periodic potential for the n-band problem can be found in H. P. McKean and P. van Moerbeke, Invent. Math. **30**, 217–74 (1975). Equation (1.21), attributed to V. A. Marchenko, is in fact due to Hochstadt [see W. Magnus and W. Winkler, *Hill's Equation* (Interscience, New York, 1966)]. For eigenfunction expansions see also M. S. P. Eastham, *The Spectral Theory of Periodic Differential Equation* (Scottish Academic Press, Edinburgh, 1973). Some simplification takes place once H is recognized as a direct integral of operators with compact resolvent [see J. Avron, A. Grossman, and R. Rodriguez, Rep Math. Phys. **5**, 113 (1974)].

*The main results of this paper were presented in May 1975 at the Summer School in Zagul'b near Baku.

†*Editor's note*: Because of the author's inability to have his work published in the USSR, this paper was sent to the *Journal of Mathematical Physics*, translated and refereed. Some small corrections were inserted by the referee and some additional comments made which appear in the referee's note.

[1]It is convenient to work with dimensional quantities; in Eq. (0.1) x has the dimensions of cm, t is in cm^3, and $u(x;t)$ is in cm^{-2}. The reader can find the basic properties of the KdV equation and of solitons in the survey paper by Scott, Chu, and McLaughlin [Proc. IEEE **61**, 1443 (1974)].

[2]S. P. Novikov, Funct. Anal. **8**, 54 (1974) [Sov. Math. Funct. Anal. **8**, 236 (1974)].

[3]A. R. Its and V. B. Matveev, Theor, and Math. Phys. **23**, 51 (1975) [Sov. Math. TMP 23, 343 (1975)].

[4]E. C. Titchmarsh, Eigenfunction Expansions (Oxford U. P., Oxford, 1958).

[5]N. N. Meiman, Dokl. Akad. Nauk **130**, 257 (1960) [Sov. Math. Doklady **1**, 41 (1960)].

[6]N. N. Meiman, Dokl. Akad. Nauk **130**, 503 (1960) [Sov. Math. Doklady **1**, 72 (1960)].

[7]N. N. Meiman, Trud. Mosk. Mat. Obshch. **9**, (1960) [Amer. Math. Soc. Transl. Ser. 2 **32**, 350 (1963)].

[8]B. A. Dubrovin and S. P. Novikov, Zh. Eksp. Teor. Fiz. **67**, 2131 (1974) [Sov. Phys. JETP **40**, 1058 (1974)].

[9]B. A. Dubrovin, V. B. Matveev, and S. P. Novikov, Usp. Mat. Nauk **31**, 55 (1976).

[10]V. A. Marchenko, Mat. Sb. **95**, 331 (1974).

[11]N. N. Meiman, Dokl. Akad. Nauk **125**, No. 5, (1959).

[12]V. F. Lazutkin and T. F. Pankratova, Dokl. Akad. Nauk **215**, 1048 (1974) [Sov. Math. Doklady **15-1**, 649 (1974)].

[13]I. V. Stankevich, Dokl. Akad. Nauk **192**, 34 (1970) [Sov. Math. Doklady **11**, 582 (1970)].

[14]N. N. Meiman, Dokl. Akad. Nauk **126**, No. 2 (1959).

[15]N. N. Meiman, Dokl. Akad. Nauk **140**, 755 (1961) [Sov. Math. Doklady **2**, 1261 (1961)].

[16]N. N. Meiman, Zh. Eksp. Teor. Fiz. Pis. Red. **21**, 624 (1975) [Sov. Phys. JETP Lett. **21**, 292 (1975)]; in this article only the case $k(p)=0$ was considered.

Multiquark models and the embedding of groups

B. Gruber*† and T. S. Santhanam*‡

International Centre for Theoretical Physics, Trieste, Italy
(Received 12 March 1976; revised manuscript received 1 June 1976)

Recent experimental results have clearly indicated that the usual quark model must be extended. Several multiquark models are already under consideration. In this paper a classification of *all possible symmetry chains* (group extensions) is given for the embedding of the group SU(3) in semisimple Lie groups for the case that the basic quark fields transform with respect to SU(3) either as $(3+3^*)$, $(3+3)$, $(3+3^*+1)$ or $(3+3+1)$.

I. INTRODUCTION

The recent advances in the domain of high energy physics have clearly indicated[1] the existence of new quantum numbers and the necessarity of extending the usual Gell-Mann—Zweig model[2] even with color.[3] The charm model[4] has also reappeared in the Glashow—Iliopoulos—Maiani mechanism.[5] Six quark models have been proposed by Harari,[6] Barnett,[7] Albright and Oakes[8] and in a different context, earlier by Bacry *et al.*[9] Earlier we had proposed[10] a seven quark model consisting of a triplet, an antitriplet and a singlet. Models of Han-Nambu[11] and of Greenberg[12] involve nine quarks, and gauge models of Pati—Salam[13] require thirty-two quarks. Harari[1] has strongly advocated in favor of a triplet, antitriplet model by analyzing all the available experimental information. The increase in the number of quarks in all these models implies the enlargement of the group SU(3).

In this paper we study the embedding of the group SU(3) in the group SU(6) such that the basic quark states transform like the representations $(3+3)$ and $(3+3^*)$ of SU(3). All symmetry chains leading from SU(6) to these particular SU(3) subgroups are determined, and thus *all* possible models based upon $(3+3)$ and $(3+3^*)$ quarks are found (see Figs. 1 and 2). In addition, an analogous analysis is carried out for the case of quarks which transform like the representations $(1+3+3)$ and $(1+3+3^*)$ of SU(3). That is, all symmetry chains that lead from SU(7) to these particular SU(3) subgroups are determined, yielding *all* possible models that can be built based upon $(1+3+3)$ and $(1+3+3^*)$ quarks (see Figs. 3 and 4).

Apart from a complete classification of possible models, interrelationships between the various models (symmetry chains) are obtained. The matrices defining the various embeddings (up to phases) are given, and with it the mapping from the weight space of the group into the weight space of the subgroup (branching of representations under restriction to the subgroup). Of the classes of equivalent defining matrices, corresponding to conjugate subgroups, the ones are chosen which lead to one and the same SU(3) subgroup through all symmetry chains [and not to conjugate SU(3) subgroups].

The approach taken follows that of Dynkin's classification of semisimple Lie groups (algebras) in semisimple Lie group (algebras)[14] and the work of one of the authors (B. G.).[15]

II. EMBEDDING IN SU(6)

In this section we analyze the embedding of the representations $(3+3)$ and $(3+3^*)$ of SU(3) in the defining six-dimensional representation of SU(6). From the Dynkin rules we can see that the chains of embedding are the following.

A. The case of 3 + 3

We use the notation $A_l \sim \mathrm{SU}(l+1)$ (Fig. 1) and the superscript denotes the Dynkin index specifying the nature of embedding.

(a) The chain $A_2^2 \subset A_2^2 + A_1^3 \subset A_5$

The embedding of the diagonal generators $\tilde{H}$ of the subgroups in the diagonal generators H_μ $(\mu = 1, 2, \ldots, 6)$ of A_5 are given by the mapping

$$\tilde{H} = \rho H, \tag{2.1}$$

where the defining matrix ρ is given by

$$\rho(A_2^2 + A_1^3 \subset A_5) = \left[\begin{array}{cccccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ \hline \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{array}\right], \tag{2.2}$$

$$\rho(A_2^2 \subset A_2^2 + A_1^3) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{2.3}$$

From Eqs. (2.2) and (2.3) it follows that

$$\rho(A_2^2 \subset A_5) = \frac{1}{3}\begin{bmatrix} 2 & -1 & -1 & 2 & -1 & -1 \\ -1 & 2 & -1 & -1 & 2 & -1 \\ -1 & -1 & 2 & -1 & -1 & 2 \end{bmatrix}, \tag{2.4}$$

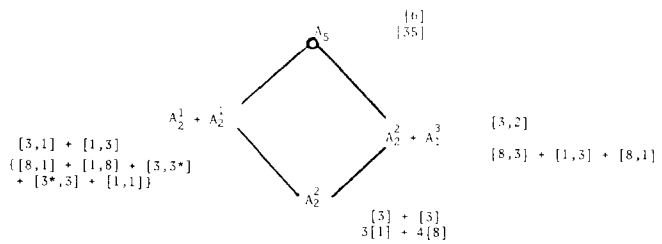


FIG. 1. Embedding of the representation $(3+3)$ of A_2^2 in A_5.

Copyright © 1977 American Institute of Physics

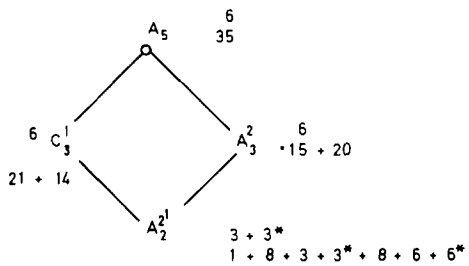


FIG. 2. Embedding of the representation $(3+3^*)$ of $A_2^{2'}$ in A_5.

where we have used the "trace" condition of the diagonal generators of A_5,

$$H_1+H_2+\cdots+H_6=0. \tag{2.5}$$

From Eqs. (2.2), (2.3), and (2.4) we see that the six-dimensional representation branches under this chain as $6\to(3,2)\to 3+3$. This chain is well known in the Gürsey–Radicati–Sakita scheme[16] of SU(6), where the extension from SU(3) to SU(6) is by spin.

(b) The chain $A_2^2\subset A_2^1+A_2^1\subset A_5$

The defining matrices in this case are given by

$$\rho(A_2^1+A_2^1\subset A_5)=\frac{1}{3}\left[\begin{array}{ccc|ccc} 2&-1&-1&&&\\ -1&2&-1&&0&\\ -1&-1&2&&&\\ \hline &&&2&-1&-1\\ &0&&-1&2&-1\\ &&&-1&-1&2 \end{array}\right], \tag{2.6}$$

where use has been made of the "trace" condition of the diagonal generators of SU(3) and

$$\rho(A_2^2\subset A_2^1+A_2^1)=\frac{1}{3}\begin{bmatrix} 2&-1&-1&2&-1&-1\\ -1&2&-1&-1&2&-1\\ -1&-1&2&-1&-1&2 \end{bmatrix}, \tag{2.7}$$

and we see that the defining matrix in the chain $A_5\to A_2^1+A_2^1\to A_2^2$ is the same as before in Eq. (2.4). This chain has been used to study the collinear SU(3)⊗SU(3)[17] and recently by Barnett.[7] The six-dimensional representation branches under this chain as $6\to(3,1)+(1,3)\to 3+3$.

B. The case of 3 + 3*

The embedding of the representation $3+3^*$ of SU(3) in the defining representation of SU(6) goes through two different chains (Fig. 2).

(c) $A_2^{2'}\subset C_3^1\subset A_5$

The defining matrices are

$$\rho(C_3\subset A_5)=\begin{bmatrix} 1&0&0&0&0&-1\\ 0&1&0&0&-1&0\\ 0&0&1&-1&0&0 \end{bmatrix}, \tag{2.8}$$

and

$$\rho(A_2^{2'}\subset C_3)=\frac{1}{3}\begin{bmatrix} 2&-1&-1\\ -1&2&-1\\ -1&-1&2 \end{bmatrix}. \tag{2.9}$$

It should be observed that C_3 is not a regular subalgebra of SU(6) in the sense that the root system of C_3 is not a subsystem of the roots of SU(6). The embedding of $A_2^{2'}$ in A_5 in this chain is therefore furnished by the matrix

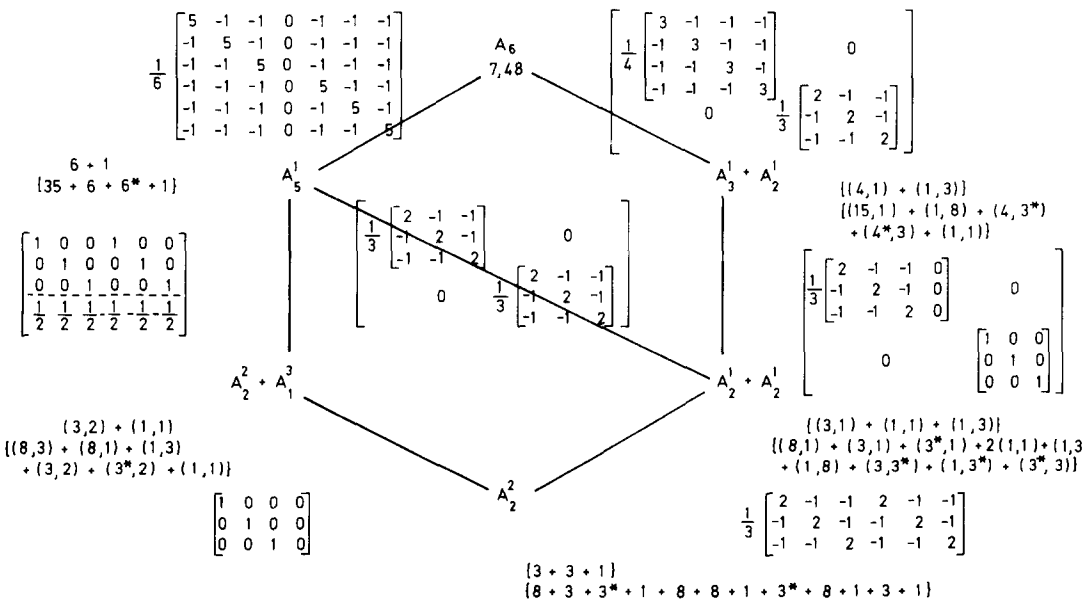


FIG. 3. Embedding of the representation $(1+3+3)$ of A_2^2 in A_6. Among the equivalent conjugate subalgebras we have chosen those for which the embedding is identical in all the chains and the embedding is given by

$$\rho(A_2^2\subset A_6)=\tfrac{1}{3}\begin{bmatrix} 2&-1&-1&0&2&-1&-1\\ -1&2&-1&0&-1&2&-1\\ -1&-1&2&0&-1&-1&2 \end{bmatrix}.$$

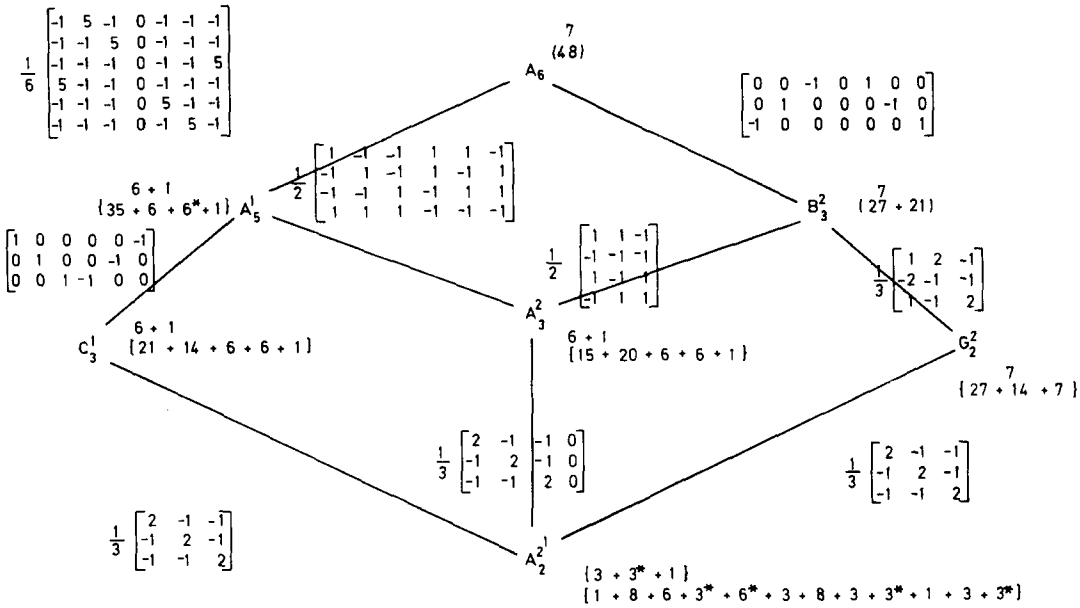


FIG. 4. Embedding of the representation $(1+3+3^*)$ of $A_2^{2'}$ in A_6. Among the equivalent conjugate subalgebras we have chosen those for which the embedding is identical in all the chains and the embedding is given by

$$\rho(A_2^{2'}\subset A_6)=\tfrac{1}{3}\begin{bmatrix} 1 & 2 & -1 & 0 & 1 & -2 & -1 \\ 1 & -1 & 2 & 0 & -2 & 1 & -1 \\ -2 & -1 & -1 & 0 & 1 & 1 & 2 \end{bmatrix}.$$

$$\rho(A_2^{2'}\subset A_5)=\frac{1}{3}\begin{bmatrix} 2 & -1 & -1 & 1 & 1 & -2 \\ -1 & 2 & -1 & 1 & -2 & 1 \\ -1 & -1 & 2 & -2 & 1 & 1 \end{bmatrix} \tag{2.10}$$

from which it is clear that the six-dimensional representation of SU(6) branches under this chain as $6\to 6\to 3+3^*$.

This chain corresponds to the model of Bacry *et al.*,[9] where the embedding is through the symplectic group.

(d) $A_2^{2'}\subset A_3^2\subset A_5$

The defining matrices are given by

$$\rho(A_3^2\subset A_5)=\frac{1}{2}\begin{bmatrix} 1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 \end{bmatrix}, \tag{2.11}$$

and

$$\rho(A_2^{2'}\subset A_3^2)=\frac{1}{3}\begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \end{bmatrix}, \tag{2.12}$$

and therefore it follows that the defining matrix for the embedding of $A_2^{2'}$ is the same as Eq. (2.10). The embedding here too is nonregular. The six-dimensional representation branches in this chain as $6\to 6\to 3+3^*$. This chain corresponds to the model of Albright and Oakes[8] and that of Harari[1] (his 0- scheme). *The embeddings of* A_2^2 and $A_2^{2'}$ *in SU(6) are thus entirely different and they are nonconjugate subalgebras of* A_5. The usefulness of the method can be seen, for instance, by the fact that the embedding of the representation $3+3^*$ of SU(3) in SU(6) *cannot* be through the group SU(3) × SU(3) (U-scheme of Harari[1]).

The trivial chain $A_5\to A_4\to A_3\to A_2$ has not been considered here. Under this, the branching is $6\to 5+1\to 4+1+1\to 3+1+1+1$.

III. EMBEDDING IN SU(7)

In this section we shall study the embedding of the group SU(3) in the group SU(7). In particular, we shall see that there are two completely different types of embeddings, one for the representation $1+3+3$ and the other for $1+3+3^*$. Of course, the embeddings we studied in the previous section can be extended (trivially) as $A_5\to A_6$. However, we shall see that there are nontrivial embeddings.

A. The case of 1 + 3 + 3

Dynkin's method gives the following additional chains (Fig. 3):

$$A_2^2\subset A_2^1+A_2^1\subset A_3^1+A_2^1\subset A_6.$$

The defining matrices for this embedding are given by

$$\rho(A_3^1+A_2^1\subset A_6)=\begin{bmatrix} \frac{1}{4}\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix} & 0 \\ 0 & \frac{1}{3}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \end{bmatrix}, \tag{3.1}$$

$$\rho(A_2^1+A_2^1\subset A_3^1+A_2^1)=\frac{1}{3}\begin{bmatrix}2&-1&-1&0\\-1&2&-1&0\\-1&-1&2&0\end{bmatrix}, \tag{3.2}$$

with A_2^1 remaining the same and with $A_2^2\subset A_2^1+A_2^1$ given earlier in Eq. (2.7) we find that

$$\rho(A_2^2\subset A_6)=\frac{1}{3}\begin{bmatrix}2&-1&-1&0&2&-1&-1\\-1&2&-1&0&-1&2&-1\\-1&-1&2&0&-1&-1&2\end{bmatrix}. \tag{3.3}$$

From this it follows that the seven-dimensional representation of SU(7) branches under this chain as $7\to(4,1)+(1,3)\to(3,1)+(1,1)+(1,3)\to3+1+3$.

B. The case of 1 + 3 + 3*

The following two additional chains (additional to the trivial embedding $A_5\to A_6$) are obtained (Fig. 4):

(i) $A_2^{2'}\subset A_3^2\subset B_3^2\subset A_6$.

The defining matrices in this chain are given by

$$\rho(B_3^2\subset A_6)=\begin{bmatrix}0&0&-1&0&1&0&0\\0&1&0&0&0&-1&0\\-1&0&0&0&0&0&1\end{bmatrix}, \tag{3.4}$$

$$\rho(A_3^2\subset B_3^2)=\frac{1}{2}\begin{bmatrix}1&1&-1\\-1&-1&-1\\1&-1&1\\-1&1&1\end{bmatrix}, \tag{3.5}$$

and with $A_2^{2'}\subset A_3^2$ given by Eq. (2.12) we get

$$\rho(A_2^{2'}\subset A_6)=\frac{1}{3}\begin{bmatrix}1&2&-1&0&1&-2&-1\\1&-1&2&0&-2&1&-1\\-2&-1&-1&0&1&1&2\end{bmatrix} \tag{3.6}$$

from which it follows that the seven-dimensional representation of A_6 branches under this chain as $7\to7\to6+1\to3+3^*+1$.

(ii) $A_2^{2'}\subset G_2\subset B_3^2\subset A_6$

The defining matrices are given by

$$\rho(A_2^{2'}\subset G_2)=\frac{1}{3}\begin{bmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{bmatrix}, \tag{3.7}$$

$$\rho(G_2^2\subset B_3^2)=\frac{1}{3}\begin{bmatrix}1&2&-1\\-2&-1&-1\\1&-1&2\end{bmatrix}, \tag{3.8}$$

and with $(B_3^2\subset A_6)$ given by Eq. (3.4) we find that $(A_2^{2'}\subset A_6)$ is the same as in Eq. (3.6). The seven-dimensional representation of A_6 branches as $7\to7\to7\to3+3^*+1$. This seven quark model has been considered by us for the spectrum of hadrons.[10]

IV. CONCLUSIONS

We have discussed the embedding of the representations (3 + 3) and (3 + 3*) of SU(3) in the larger group SU(6). The resulting chains correspond to *all* the possible distinct schemes for the extension of the Gell-Mann–Zweig quark model to SU(6). We have also worked out the embedding of the representations (1 + 3 + 3) and (1 + 3 + 3*) of SU(3) in SU(7). In the Appendix we explicitly give the embedding of the generators.

ACKNOWLEDGMENTS

The authors are grateful to Professor Abdus Salam, the International Atomic Energy Agency, and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste.

APPENDIX

As mentioned in the text, the defining matrix completely specifies (up to phases) the embedding of the generators. If we denote by H and $\tilde{H}$, the diagonal generators of the group G and its subgroup G we have

$$\rho(\tilde{H}_i)=\rho_{i\mu}H_\mu,\ \cdots,\ i=1,\ldots,\tilde{l},\quad \mu=1,\ldots,l, \tag{A1}$$

where l and $\tilde{l}$ denote the ranks of G and $\tilde{G}$, respectively $(\tilde{l}\le l)$, $\rho_{i\mu}$ is the defining matrix. In the case G or $\tilde{G}$ happens to be unitary groups or G_2 we imbed, respectively, the generators in $(l+1)$- or $(\tilde{l}+1)$-dimensional spaces with the "trace" conditions $\sum_{\mu=1}^{l}H_\mu=0$ and/or $\sum_{i=1}^{\tilde{l}}\tilde{H}_i=0$. We use the following notation. The e_i's are unit vectors. The l simple roots of A_l are given by e_i-e_{i+1}, $i=1,\ldots,l$. We denote the corresponding shift operators by $\tilde{E}_{e_i-e_{i+1}}$ for the subgroup $(i=1,\ldots,\tilde{l})$ and $E_{e_i-e_{i+1}}$ for the larger group $(i=1,\ldots,l)$.

For the embedding of the shift operators of A_2^2 in A_5 we have

$$\begin{aligned}\rho(\tilde{E}_{(1,-1,0)})&=E_{e_1-e_2}+E_{e_4-e_5},\\ \rho(\tilde{E}_{(0,1,-1)})&=E_{e_2-e_3}+E_{e_5-e_6}.\end{aligned} \tag{A2}$$

For the embedding $A_2^{2'}$ in A_5 we have

$$\begin{aligned}\rho(\tilde{E}_{(1,-1,0)})&=E_{e_1-e_2}+E_{e_5-e_6},\\ \rho(\tilde{E}_{(0,1,-1)})&=E_{e_2-e_3}+E_{e_4-e_5}.\end{aligned} \tag{A3}$$

For the embedding A_2^2 in A_6 we find

$$\begin{aligned}\rho(\tilde{E}_{(1,-1,0)})&=E_{e_1-e_2}+E_{e_5-e_6},\\ \rho(\tilde{E}_{(0,1,-1)})&=E_{e_2-e_3}+E_{e_6-e_7}.\end{aligned} \tag{A4}$$

And, finally, for the embedding $A_2^{2'}$ in A_6 one finds that

$$\begin{aligned}\rho(\tilde{E}_{(1,-1,0)})&=E_{e_2-e_3}+E_{e_5-e_6},\\ \rho(\tilde{E}_{(0,1,-1)})&=E_{e_1-e_5}+E_{e_3-e_7}.\end{aligned} \tag{A5}$$

The embedding of the shift operators for any part of the chain can be worked out analogously from the respective defining matrices. We give below an example of nonregular embedding of the shift operators of C_3 in A_5:

$$\begin{aligned}\rho(\tilde{E}_{(1,-1,0)})&=E_{e_1-e_2}+E_{e_5-e_6},\\ \rho(\tilde{E}_{(0,1,-1)})&=E_{e_2-e_3}+E_{e_4-e_5},\\ \rho(\tilde{E}_{(0,0,2)})&=\sqrt{2}\,E_{e_3-e_4},\end{aligned} \tag{A6}$$

where $(0,0,2)$ is the third simple root of C_3.

*Visiting members.

†Physics Department, Southern Illinois University, Carbondale, Illinois 62901 and Physikalisches Institut, Universität Würzburg, D-87 Würzburg, Germany.

‡Permanent address: Matscience, The Institute of Mathematical Sciences, Madras, India.

[1]H. Harari, Ann. Phys. (N.Y.) 94, 391 (1975).

[2]M. Gell-Mann, Pnys. Lett. 8, 214 (1964); G. Zweig, CERN Rep. preprints TH. 401 and TH. 412, 1964 (unpublished).

[3]For the approach for color, see H. Fritsch, M. Gell-Mann, and H. Leutwyler, Phys. Lett. B 47, 365 (1973).

[4]J.D. Bjorken and S.L. Glashow, Phys. Lett. 11, 255 (1964).

[5]S.L. Glashow, J. Iliopoulos, and L. Maiani, Phys. Rev. D 2, 1285 (1970).

[6]H. Harari, Phys. Lett. B 57, 265 (1975).

[7]R.M. Barnett, Phys. Rev. Lett. 34, 41 (1975); Phys. Rev. B 11, 3246 (1975).

[8]C.J. Albright and R.J. Oakes, Proc. Argonne Symp. Hadron Spectrosc., July 1975 (to be published).

[9]H. Bacry, J. Nyuts, and L. Van Hove, Phys. Lett. 9, 279 (1964); Nuovo Cimento 35, 510 (1965).

[10]B. Gruber and T.S. Santhanam, "Seven-Quark Model for the Spectrum of Hadrons", Würzburg preprint, March 1975 (unpublished) and Phys. Rev. D (in press).

[11]M. Han and Y. Nambu, Phys. Rev. 139, B1006 (1965).

[12]O.W. Greenberg, Maryland Report 75-064 (1975) (unpublished).

[13]See, for instance, J.C. Pati and Abdus Salam, Proc. Palermo Conf. (1975) (to be published).

[14]E. Dynkin, Amer. Math. Soc. Transl. Ser 2 6, 111 (1957).

[15]M. Lorente and B. Gruber, J. Math. Phys. 13, 1639 (1972); B. Gruber and M.T. Samuel, in *Group Theory and Its Applications*, edited by E.M. Loebl (Academic, New York, 1975), Vol. III; B. Gruber, in *Group Theory in Non-Linear Problems*, edited by A.O. Barut (Reidel, Dordrecht, Holland, 1974), p. 230.

[16]F. Gürsey and L.A. Radicati, Phys. Rev. Lett. 13, 173 (1964); B. Sakita, Phys. Rev. 136, B1756 (1964).

[17]See, for instance, H. Ruegg, W. Rühl, and T.S. Santhanam, Helv. Phys. Acta 40, 9 (1967).

ERRATA

Erratum: Electromagnetic eigenmode perturbations caused by deformation of a plasma core in a spherical cavity resonator [J. Math. Phys. 18, 312 (1977)]

Arno D. Steiger

Lawrence Livermore Laboratory, University of California, Livermore, California 94550

(Received 25 January 1977)

1. Equation (3.6) should read

$$-\delta\bar{U} = \int da\, \mathbf{e}_r \cdot \mathbf{F}_p\, \delta r.$$

2. The expressions on the right-hand sides of Eqs. (3.8), (3.19), (3.24), and (3.30a) and (3.30b) should appear with the opposite sign.

Copyright © 1977 American Institute of Physics

*Visiting members.

†Physics Department, Southern Illinois University, Carbondale, Illinois 62901 and Physikalisches Institut, Universität Würzburg, D-87 Würzburg, Germany.

‡Permanent address: Matscience, The Institute of Mathematical Sciences, Madras, India.

[1]H. Harari, Ann. Phys. (N.Y.) 94, 391 (1975).

[2]M. Gell-Mann, Pnys. Lett. 8, 214 (1964); G. Zweig, CERN Rep. preprints TH. 401 and TH. 412, 1964 (unpublished).

[3]For the approach for color, see H. Fritsch, M. Gell-Mann, and H. Leutwyler, Phys. Lett. B 47, 365 (1973).

[4]J.D. Bjorken and S.L. Glashow, Phys. Lett. 11, 255 (1964).

[5]S.L. Glashow, J. Iliopoulos, and L. Maiani, Phys. Rev. D 2, 1285 (1970).

[6]H. Harari, Phys. Lett. B 57, 265 (1975).

[7]R.M. Barnett, Phys. Rev. Lett. 34, 41 (1975); Phys. Rev. B 11, 3246 (1975).

[8]C.J. Albright and R.J. Oakes, Proc. Argonne Symp. Hadron Spectrosc., July 1975 (to be published).

[9]H. Bacry, J. Nyuts, and L. Van Hove, Phys. Lett. 9, 279 (1964); Nuovo Cimento 35, 510 (1965).

[10]B. Gruber and T.S. Santhanam, "Seven-Quark Model for the Spectrum of Hadrons", Würzburg preprint, March 1975 (unpublished) and Phys. Rev. D (in press).

[11]M. Han and Y. Nambu, Phys. Rev. 139, B1006 (1965).

[12]O.W. Greenberg, Maryland Report 75-064 (1975) (unpublished).

[13]See, for instance, J.C. Pati and Abdus Salam, Proc. Palermo Conf. (1975) (to be published).

[14]E. Dynkin, Amer. Math. Soc. Transl. Ser 2 6, 111 (1957).

[15]M. Lorente and B. Gruber, J. Math. Phys. 13, 1639 (1972); B. Gruber and M.T. Samuel, in *Group Theory and Its Applications*, edited by E.M. Loebl (Academic, New York, 1975), Vol. III; B. Gruber, in *Group Theory in Non-Linear Problems*, edited by A.O. Barut (Reidel, Dordrecht, Holland, 1974), p. 230.

[16]F. Gürsey and L.A. Radicati, Phys. Rev. Lett. 13, 173 (1964); B. Sakita, Phys. Rev. 136, B1756 (1964).

[17]See, for instance, H. Ruegg, W. Rühl, and T.S. Santhanam, Helv. Phys. Acta 40, 9 (1967).

ERRATA

Erratum: Electromagnetic eigenmode perturbations caused by deformation of a plasma core in a spherical cavity resonator [J. Math. Phys. 18, 312 (1977)]

Arno D. Steiger

Lawrence Livermore Laboratory, University of California, Livermore, California 94550

(Received 25 January 1977)

1. Equation (3.6) should read

$$-\delta\bar{U} = \int da\, \mathbf{e}_r \cdot \mathbf{F}_p\, \delta r.$$

2. The expressions on the right-hand sides of Eqs. (3.8), (3.19), (3.24), and (3.30a) and (3.30b) should appear with the opposite sign.

Copyright © 1977 American Institute of Physics