On the Error of Approximations in Quantum Mechanics I. General Theory

T. A. Hoffmann

Phil. Trans. R. Soc. Lond. A 1965 **257**, 309-326
doi: 10.1098/rsta.1965.0007

ON THE ERROR OF APPROXIMATIONS IN QUANTUM MECHANICS

I. GENERAL THEORY

By T. A. HOFFMANN*

Aktiebolaget Atomenergi, Studsvik, Sweden

(*Communicated by C. A. Coulson, F.R.S.—Received* 19 *March* 1964)

CONTENTS

General formulas for estimating the errors in quantum-mechanical calculations are given in the formalism of density matrices. Some properties of the traces of matrices are used to simplify the estimating and to indicate a way of obtaining a better approximation. It is shown that the simultaneous correction of all the equations to be fulfilled leads in most cases to a faster convergence than the exact fulfilment of some of the equations and approximating stepwise to some of the others. The corrective formulas contain only direct operations of the matrices occurring and so they are advantageous in computer applications.

In the last section a 'subjective error' definition is given and by taking into account the weight of the errors of the several equations a faster convergence and a single error quantity is obtained. Some special applications of the method will be published later.

1. Introduction

Since the development of quantum mechanics in the second quarter of this century there has been an immense amount of work in applying it to several kinds of electronic systems. The treatment of the simplest systems with one electron was followed by a thorough investigation of the possibilities to calculate a system containing more electrons in a sufficiently good approximation. After the simplest methods the more sophisticated treatments of the various perturbation methods, self-consistent-field calculations and variational processes were developed. The most recent applications of large-scale computers made it possible to perform relatively simply many of those calculations previously desired, but hitherto impossible. It is worth while mentioning that the first modern version of computers, the differential analyser, was constructed and used by Hartree to make self-consistent-field calculations (1928, 1946, 1957).

However, there was a basic difficulty in almost all of these calculations. Namely, if the calculated wave function was not an exact solution of the problem, but only some approximation, then there was in general no measure of estimating the error of the approximation. Most of the approximations used were based on neglecting more or less important parts of the Hamiltonian or making more or less justifiable assumptions regarding the form of the

* On leave of absence from the Research Institute for Telecommunication, Budapest, and from the Central Research Institute for Physics of the Hungarian Academy of Sciences, Budapest.

wave function. But there were in general no means to estimate the error expected before the full calculation was performed and the results compared to some other calculations or to experimental values.

At the same time the lack of a suitable measure of the error had the result in some cases that although the first-order approximations approached the experimental values fairly well, the second-order or higher approximations showed a basic divergence. A lot of work has been undertaken to find an objective criterion for the estimation of the error of any quantum-mechanical calculation (see Temple 1928; Eckart 1930; Shull & Löwdin 1958; Trefftz 1959; Kinoshita 1959; Löwdin 1960; Frost, Kellogg & Curtis 1960; Fröman & Hall 1961; Preuss & Trefftz 1957; Preuss 1958). In general, the choice of the quantity to be minimized to obtain the most rapid convergence was investigated. However, these investigations have shown a large variety of possible cases occurring in various types of approximations without a decisive general answer. Some of the minimization processes show very good convergence in some cases, but fail to converge in other cases. McWeeny (1956*a*,*b*; 1960) has shown some reasons for the divergences of some self-consistent-field calculations and applied a more precisely adapted way of calculation avoiding divergence difficulties.

In the present work a more general way of estimating the error of the approximations is given and therefore a means is provided to control a quantum-mechanical calculation in the case that more errors are to be minimized simultaneously. At the same time, in the development of this paper the applications of modern high-speed computers to perform the calculations sketched below was kept in view.

2. Formulation of the exact problem

Let us suppose that we have made the physical analysis of the problem to be investigated and so we know fairly well how the Hamiltonian of our system is built up of kinetic energy operators and potential energy operators. Assume the potential energy as being a given function of the co-ordinates (including spin co-ordinates) of the electrons, not excluding the possibility that the potential energy is a function in the extended sense, so that it contains differentiations or other operations on the electronic co-ordinates too. Let us denote the exact Hamiltonian constructed by this potential by **H**. The stationary states for a system of electrons are then given by the Schrödinger equation

$$\mathbf{H}\psi_i = E_i\psi_i, \tag{1}$$

where ψ_i is the wave function containing co-ordinates (and eventually spin co-ordinates) of all of the electrons present in the system, E_i is the total energy of the system. Here we disregard the case where E can obtain any value in an interval, the case of a continuous spectrum, and fix ourselves to the case of a discrete spectrum, where only distinct, well-spaced values for E are accessible. The extension of the following arguments is easily made for the continuous case too, although the notation is somewhat more clumsy. The index i refers to the energy state E_i, but we allow that some of the E_i are equal to each other, so introducing the possibility of degenerate states too.

In general there is an infinite number of solutions of equation (1). The real values $E_1, E_2, \ldots, E_i, \ldots$ form the energy spectrum of the system and the corresponding functions $\psi_1, \psi_2, \ldots, \psi_i, \ldots$ span the vector space of the electron wave functions. The exact solutions of (1) have the property that the wave functions are mutually orthogonal to each other if the corresponding E_i values are different from one another. In the case that some of the E_i are equal to each other, there could be performed an orthogonalization process among the corresponding wave functions so that these wave functions would be orthogonal to each other too. Furthermore, since equation (1) is linear in ψ, the normalization of the wave functions can be performed too. We choose the normalization so that

$$\int |\psi_i|^2 \, \mathrm{d}\tau_1 \ldots \mathrm{d}\tau_n = 1, \tag{2}$$

where the integration is performed over the space (and spin space) of every electron, the total number of electrons in the system being n. According to the previous remark we may write, without loss of generality, the general orthogonality relation

$$\int \psi_i^* \, \psi_j \, \mathrm{d}\tau_1 \ldots \mathrm{d}\tau_n = 0 \quad (i \neq j). \tag{3}$$

This last equation seems to denote something less than the usual orthogonality relations, because these latter usually denote the orthogonality of two wave functions of two electrons, and therefore they are orthogonal by integrating over the co-ordinates of only one electron, whereas (3) states only that the integration performed over *all* electrons gives the orthogonality. Naturally, it is sometimes the case that already one or a few of the integrations yield the orthogonality relation. But one has to bear in mind that sometimes (e.g. by taking into account the perfect correlation) the wave function cannot even be separated into simple products of one-electron wave functions and so the result of the integration over one or a few of the electron spaces involves the co-ordinates of the other electrons too, not as multiplicative parameters. The orthogonality relation for one or a few electrons implies the orthogonality relation (3), but not the reverse.

Any of the functions $\psi_1, \psi_2, \ldots, \psi_i, \ldots$ may also be represented in principle in the following way. Imagine the space of the n electrons (together with the spin space) to be divided into sufficiently small volume elements. In each of these volume elements the co-ordinates of the n electrons (together with their spin co-ordinates) are specified. To be able to give the function ψ_i we have only to attach a numerical value (maybe a complex value) to each volume element. If these volume elements are ordered in a fixed way, making a convention of their sequence, the function is given merely by the sequence of these numerical values. Let us denote these numerical values for ψ_i by $\psi_i(x = (1))$, $\psi_i(x = (2))$, $\psi_i(x = (3))$, ..., where the notation implies the convention that $x = (1)$ always means the same specified point in configurational space and spin space, $x = (2)$ another specified, but always the *same* point throughout the calculation. So the entirety of the wave functions may be represented in a matrix form, $\boldsymbol{\psi}$, where the various *columns* denote the various states belonging to $\psi_1, \psi_2, \ldots, \psi_i, \ldots$, respectively, and the various *rows* denote the various conventionally specified points in the space (and spin space) of all electrons (see Löwdin 1955*a*, *b*, 1956):

$$\boldsymbol{\psi} = \begin{pmatrix} \psi_1(x=(1)) & \psi_2(x=(1)) & \psi_3(x=(1)) \ldots & \psi_i(x=(1)) \ldots \\ \psi_1(x=(2)) & \psi_2(x=(2)) & \psi_3(x=(2)) \ldots & \psi_i(x=(2)) \ldots \\ \psi_1(x=(3)) & \psi_2(x=(3)) & \psi_3(x=(3)) \ldots & \psi_i(x=(3)) \ldots \\ \vdots & \vdots & \vdots & \vdots \\ \psi_1(x=(j)) & \psi_2(x=(j)) & \psi_3(x=(j)) \ldots & \psi_i(x=(j)) \ldots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \tag{4}$$

This infinite matrix is only a notation and has as many columns as the number of the different states of the system and as many rows as the subdivision of the n-electron space demands.

In practice, however, the matrix is not taken as an infinite one, but takes into account only as many states (columns), s, as necessary, and such a fine subdivision of the space of all of the electrons as convenient, i.e. the number of rows will be m^n, where m is the number of volume elements of the space of one electron and n is the number of electrons.

In equation (1), $\mathbf{H}$ transforms a column $\boldsymbol{\psi}$ into another column matrix, so that $\mathbf{H}$ can be represented as a square matrix of size $m^n \times m^n$ the elements of which are operators themselves. We introduce also a diagonal matrix $\mathbf{E}$ having the size $s \times s$ and having for the diagonal elements the values of the energy parameter $E_1, E_2, E_3, \ldots, E_i, \ldots$. So

$$\mathbf{E} = \begin{pmatrix} E_1 & 0 & 0 & \ldots & 0 & \ldots \\ 0 & E_2 & 0 & \ldots & 0 & \ldots \\ 0 & 0 & E_3 & \ldots & 0 & \ldots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \ldots & E_i & \ldots \\ \vdots & \vdots & \vdots & & \vdots & \end{pmatrix}. \tag{5}$$

The well-known rules of matrix multiplication allow $\boldsymbol{\psi}$ to be multiplied by $\mathbf{H}$ from the left-hand side and by $\mathbf{E}$ from the right-hand side only. So the equation

$$\mathbf{H}\boldsymbol{\psi} = \boldsymbol{\psi}\mathbf{E} \tag{6}$$

can be written for the *totality* of equations (1) (i.e. for all values of i).

This equation gives back all the elements of the matrices in a way controlled by equation (1).

Let us denote the general element (an operator itself) of $\mathbf{H}$ by $\mathbf{H}_{jk}$. According to the matrix multiplication rule acting on the column ψ_i of (4) this has the result, given by the use of (1),

$$\sum_k \mathbf{H}_{jk}\, \psi_i(x=(k)) = E_i \psi_i(x=(j)). \tag{7}$$

If we refine the subdivision in the space so that the summation in (7) transforms into an integration, we see that $\mathbf{H}_{jk}$ corresponds now to the kernel of an integral equation, an extended Green function $H(x,y)$. That is, if we introduce the continuous variables x for (j) and y for (k), where both x and y denote the whole ensemble of all electron co-ordinates in the system, we may write (7) in the form

$$\int H(x,y)\, \psi_i(y)\, \mathrm{d}y = E_i \psi_i(x). \tag{8}$$

It is readily seen that the kernel function $H(x,y)$ yields equation (1), if we define

$$H(x,y) = H_0(y)\, \delta(y-x) + H_1(y)\, \delta'(y-x) + H_2(y)\, \delta''(y-x), \tag{9}$$

where $\delta(x)$ is the many-dimensional Dirac function and $\delta'(x)$ and $\delta''(x)$ are its first and second derivatives, respectively. The functions H_0, H_1 and H_2 can be constructed by the proper form of the Hamiltonian in such a way that for any function $f(x)$,

$$\mathbf{H}f(x) \equiv H_2(x)\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} + \left[H_1(x) + 2\frac{\mathrm{d}H_2}{\mathrm{d}x}\right]\frac{\mathrm{d}f}{\mathrm{d}x} + \left[H_0(x) + \frac{\mathrm{d}H_1}{\mathrm{d}x} + \frac{\mathrm{d}^2 H_2}{\mathrm{d}x^2}\right] f(x) \tag{10}$$

identically shall supply the effect of the ordinary Hamiltonian operator. Here x stands symbolically for all of the co-ordinates of all of the electrons and the derivatives mean in the same way symbolically all of the first and second derivatives, respectively.

The matrix for $\mathbf{H}$ given by the elements (9) is now a diagonal matrix. Let us denote the transposed of a matrix $\mathbf{A}$ by $\mathbf{A}^+$ and the complex conjugate of $\mathbf{A}$ by $\mathbf{A}^*$. The transpose of the matrices $\mathbf{H}$ and $\boldsymbol{\psi}$ fulfil clearly the general property of the transposition of a product, namely

$$\boldsymbol{\psi}^+\mathbf{H}^+ = (\mathbf{H}\boldsymbol{\psi})^+. \tag{11}$$

Here we may see easily that from (9)

$$H^+(x,y) = H(y,x) = H_0(y)\,\delta(y-x) - H_1(y)\,\delta'(y-x) + H_2(y)\,\delta''(y-x), \tag{12}$$

where the symmetry properties of the Dirac function were used. Now the operator $\mathbf{H}$ is Hermitian (see, for example, von Neumann 1932) and so we have

$$\mathbf{H}^+ = \mathbf{H}^*, \tag{13}$$

which leads to
$$H_0^*(y) = H_0(y), \tag{14}$$

$$H_1^*(y) = -H_1(y) \tag{15}$$

and
$$H_2^*(y) = H_2(y), \tag{16}$$

which can be checked for every Hermitian Hamiltonian.

Taking now the complex conjugate and transpose of equation (6) we obtain

$$\boldsymbol{\psi}^{+*}\mathbf{H}^{+*} = \mathbf{E}^{+*}\boldsymbol{\psi}^{+*} \tag{17}$$

and by using (13) and taking into account that $\mathbf{E}$ is diagonal and real,

$$\boldsymbol{\psi}^{+*}\mathbf{H} = \mathbf{E}\boldsymbol{\psi}^{+*}. \tag{18}$$

Now multiplying equation (6) from the right-hand side by $\boldsymbol{\psi}^{+*}$ and equation (18) from the left-hand side by $\boldsymbol{\psi}$ we obtain

$$\mathbf{H}\boldsymbol{\psi}\boldsymbol{\psi}^{+*} = \boldsymbol{\psi}\mathbf{E}\boldsymbol{\psi}^{+*} = \boldsymbol{\psi}\boldsymbol{\psi}^{+*}\mathbf{H}. \tag{19}$$

This equation suggests the introduction of the (n-electron) density matrix $\mathbf{R}$ (see, for example, Löwdin 1955*b*; McWeeny 1956*a*), defined by

$$\mathbf{R} = \boldsymbol{\psi}\boldsymbol{\psi}^{+*}, \tag{20}$$

where $\mathbf{R}$ is clearly a square matrix of size $m^n \times m^n$. By this definition

$$\mathbf{R}^+ = (\boldsymbol{\psi}\boldsymbol{\psi}^{+*})^+ = \boldsymbol{\psi}^*\boldsymbol{\psi}^+ = \mathbf{R}^*, \tag{21}$$

therefore $\mathbf{R}$ is Hermitian too.

Equation (19) may now be written

$$\mathbf{HR} = \mathbf{RH}. \tag{22}$$

The matrix multiplication of $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^{+*}$ in the other direction gives clearly the summation over all space elements (spin space included) of all of the electrons, which transforms into an integration through the whole configurational space if the volume elements are chosen suitably small and so according to the normalization and orthogonality conditions (2) and (3)

$$\boldsymbol{\psi}^{+*}\boldsymbol{\psi} = \mathbf{I}_{(s)}, \tag{23}$$

where $\mathbf{I}_{(s)}$ denotes the unit matrix of size $s\times s$. This last equation can be used even in the continuous subdivision of the configurational space and also in the case where only larger parts of the space are taken to be volume elements (e.g. in the l.c.a.o. treatment of molecular calculations).

The definition (20) and equation (23) show now immediately

$$\mathbf{R}^2 = \psi\psi^{+*}\psi\psi^{+*} = \psi\mathbf{I}_{(s)}\psi^{+*} = \psi\psi^{+*} = \mathbf{R}, \tag{24}$$

i.e. $\mathbf{R}$ is idempotent.

It is easy to show that not only does equation (24) follow from the orthonormalization condition (23), but conversely (23) follows from (24) too. That is we may write (24) in the form

$$\psi\psi^{+*}\psi\psi^{+*} = \psi\psi^{+*}. \tag{25}$$

Now let us introduce for a moment the matrix

$$\mathbf{A} = \psi^{+*}\psi \tag{26}$$

of size $s\times s$. Multiplying both sides of the equality (25) by ψ^{+*} from the left and by ψ from the right we may write, using (26),

$$\mathbf{A}^3 = \mathbf{A}^2. \tag{27}$$

Assuming now that the matrix $\mathbf{A}$ *has* an inverse, we obtain

$$\mathbf{A} = \mathbf{I}_{(s)} \tag{28}$$

and so equation (23) is proved. (This roundabout way was followed because the inverses of non-squared matrices, as ψ or ψ^{+*}, are not defined in this sort of calculus, only the inverses of squared matrices.)

One can prove also that equation (1) follows from equation (22) (see, for example, McWeeny 1960) and so our quantum mechanical problem is reduced to equations (22) and (24).

3. Formulation of the approximations

We repeat here the exact equations to be solved (22) and (24), in a form, where the right-hand side is a matrix $\mathbf{0}$

$$\mathbf{HR}-\mathbf{RH} = \mathbf{0} \tag{29}$$

and

$$\mathbf{R}^2-\mathbf{R} = \mathbf{0}. \tag{30}$$

The matrix $\mathbf{0}$ is here a square matrix of the size $m^n\times m^n$, having all elements 0.

In practical calculations it is almost impossible to fulfil both of these equations exactly at the same time. This can be done in a very limited number of cases, the most simple cases, where an exact analytical solution is obtained. In most cases we may allow some deviations from the exact fulfilment of one or both of equations (29) and (30). The matrices obtained in this manner for $\mathbf{R}$ are then only approximations for the exact solution fulfilling both equations (29) and (30). Besides equations (29) and (30), one must take into account also an equation which expresses how many states for the system are required in the calculation. Since the sum of the diagonal elements of $\mathbf{R} = \psi\psi^{+*}$, i.e. $\operatorname{tr}\mathbf{R}$ is the same as that of

$$\mathbf{I}_{(s)} = \psi^{+*}\psi,$$

i.e. $\operatorname{tr}\mathbf{I}_{(s)}$ (see later, equation (36)), we have

$$\operatorname{tr}\mathbf{R}-s = 0. \tag{31}$$

This equation fixes the decomposition of $\mathbf{R}$ into matrices ψ of the due form.

Naturally, the system of equations (29), (30), (31) has various different solutions. There must be undertaken a special investigation to find which solution, **R**, is to be represented by an approximation and what are the conditions, so that we shall approximate to the wanted **R** and not to another one. However, in the present paper we do not deal with this question and suppose throughout that we are sufficiently near to the *required* solution that no other solution is approximated by our process. Let us suppose that we have some approximate matrix for **R** in this sense, say $\mathbf{R}_0$.

The question arises, how could we improve such an approximation. How could we decide whether one approximation is better than the other? We have to introduce a measure for the approximations. In quantum-mechanical approximations there was always the handicap that we did not have any measure of the approximation, so we could not estimate the error of a calculation. Usually, the deviations from the experimental values or from other calculated values were the only means of judging the efficiency of an approximation.

Unfortunately, equations (29), (30) and (31) are all of a different character, so there is always a lot of arbitrariness in judging one or other of these equations as the more important. In some of the calculations in the literature equations (30) and (31) are fulfilled exactly and only equation (29) is approximated. However, a somewhat lengthy computer process (or any other method which is not an exact one, but has a limited error) spoils the validity of equations (30) and (31) if no special care is taken of them. This means that generally we may not take (30) and (31) as exact equations, but consider approximating to them in the same way (but eventually to a different degree of exactness), as (29). The arbitrariness lies in estimating the relative importance of these equations. The measures given for the exactness in the literature (Temple 1928; Bartlett 1955; Löwdin 1960) are mostly for (29) only and do not take into account the deviations of (30) and (31). McWeeny (1956*a*, *b*; 1960) distorts equation (30) in every approximating step, but he uses a special set of steps to recover the original equation, the whole process being thus a double set of approximating sets.

In looking for a measure of the approximation, we have to take into account the following.

(1) The measure of the approximation should be a small number (most advantageously only one) of real positive numbers. This would mean here, that it should not be a matrix or a vector or any set of numbers. The reality is necessary for introducing the smaller—larger relation, i.e. the possibility of judging whether one or another approximation is the better. The inclusion of positivity is advantageous to avoid an ambiguity in the smaller–larger relation in the neighbourhood of 0.

(2) The measure should be constructed so that if the solution is an exact solution, the measure of the approximation should give 0.

(3) No other approximation than the exact one should have the measure 0.

Among real numbers there are several ways to choose a measure fulfilling these requirements. For example, one of the simplest ways is to choose the measure m so that a value b_0 approximates to a value b,

$$m = |b-b_0|; \tag{32}$$

another, maybe more analytical, way is to choose

$$m = (b-b_0)^2, \tag{33}$$

but it is evident that we may choose in general also

$$m = (b^l - b^l_0)^{2k}, \tag{34}$$

where k and l are any fixed positive integers. Only practical reasoning can determine that (32) and (33) are the most suitable forms and not (34) with k and l larger than 1.

In the case of a matrix, we must be sure that in the event of exactness each element of the approximating and of the approximated matrices are exactly the same, therefore their difference is **0**.

A well-known construction of the kind necessary here is connected to the trace of a real symmetric matrix. As the simple rules of matrix multiplication show, the trace of the square of a real symmetric matrix is the sum of the squares of all elements of the matrix. So, the trace forming may be applied to construct the measure. In the case of real matrices, therefore, even the trace of the square of the deviation matrix is acceptable, but we have to extend it to the case of complex matrices too. Therefore, we define as the measure of the deviation of a matrix **B** from the matrix **0**

$$m = \operatorname{tr}(\mathbf{B}^{+*}\mathbf{B}). \tag{35}$$

It is easy to see that the sum of the diagonal elements of $\mathbf{B}^{+*}\mathbf{B}$ is the sum of the squares of the moduli of the elements of **B**. So m can be 0 *only* if $\mathbf{B} = \mathbf{0}$.

This definition fulfils all three requirements stated above. However, here again we have the possibility of choosing some other definitions for m, for instance $\operatorname{tr}(\mathbf{B}^{+*}\mathbf{B}\mathbf{B}^{+*}\mathbf{B})$, etc., but the practical simplicity suggests retaining (35).

At this point we have to introduce some rules for handling traces of matrices. In general the trace of a product of *two* matrices is commutative, i.e.

$$\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA}) \tag{36}$$

for any two matrices. However, this commutativity is somewhat restricted if we have the trace of the product of more than two matrices. In this case the commutativity allows only a cyclic permutation of the matrices. So we can regard the trace of the matrices as allowing a *half-commutative algebra*. To see this we shall take in a completely general manner the trace of the product of k matrices $\mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \ldots, \mathbf{B}^{(k)}$. If the general element of the matrix $\mathbf{B}^{(l)}$ is $b^{(l)}_{ij}$, we have

$$\operatorname{tr}(\mathbf{B}^{(1)}\mathbf{B}^{(2)}\ldots\mathbf{B}^{(l)}\ldots\mathbf{B}^{(k)}) = \sum_{i_1}\sum_{i_2}\ldots\sum_{ik} b^{(1)}_{i_1 i_2}\, b^{(2)}_{i_2 i_3} \ldots b^{(l)}_{i\, i_{l+1}} \ldots b^{(k)}_{i_k i_1}. \tag{37}$$

Since $i_1, i_2, \ldots, i_k$ are all dummy indices, they could be named in any other way too. So, for instance, taking everywhere $i_j \rightarrow i_{j+1}$ $(j<k)$ and at the same time $i_k \rightarrow i_1$ gives as the continuation of equation (37)

$$\begin{aligned}\operatorname{tr}(\mathbf{B}^{(1)}\mathbf{B}^{(2)}\ldots\mathbf{B}^{(k)}) &= \sum_{i_2}\sum_{i_3}\ldots\sum_{i_1} b^{(1)}_{i_2 i_3}\, b^{(2)}_{i_3 i_4}\ldots b^{(k)}_{i_1 i_2} \\ &= \sum_{i_1}\sum_{i_2}\ldots\sum_{i_k} b^{(k)}_{i_1 i_2}\, b^{(1)}_{i_2 i_3}\ldots b^{(k-1)}_{i_k i_1} = \operatorname{tr}(\mathbf{B}^{(k)}\mathbf{B}^{(1)}\mathbf{B}^{(2)}\ldots\mathbf{B}^{(k-1)}).\end{aligned} \tag{38}$$

So the cyclic commutability is established and we can see at the same time why another, not cyclic, commutation does change the value of the trace. Naturally, the cyclic commutability includes the commutation rule in the case of two factors given in (36).

We give here, without any proof, the almost self-evident fact, that the trace of the sum of matrices is the sum of their traces, and the simple fact that the trace of a matrix is the same as that of its transposed.

In the following we shall assume for the sake of simplicity of writing the formulas that all the matrices are real. An extension of the formulas for the complex case is straightforward.

We shall now prove a theorem very important in our later discussions. Let us regard two matrices, $\mathbf{A}$ and $\mathbf{B}$, each of which deviate from the matrix $\mathbf{0}$ in the same measure, i.e.

$$\mathrm{tr}\,(\mathbf{A}^+\mathbf{A}) = \mathrm{tr}\,(\mathbf{B}^+\mathbf{B}). \tag{39}$$

We state that

$$\mathrm{tr}\,(\mathbf{A}\mathbf{B}^+) = \mathrm{tr}\,(\mathbf{A}^+\mathbf{B}) \geqslant -\mathrm{tr}\,(\mathbf{A}^+\mathbf{A}) = -\mathrm{tr}\,(\mathbf{B}^+\mathbf{B}), \tag{40}$$

and the equality sign holds only if

$$\mathbf{B} = -\mathbf{A}. \tag{41}$$

To prove this theorem let us investigate the deviation form $\mathbf{0}$ of the matrix $\mathbf{A}+\mathbf{B}$. Since the measure of deviation from $\mathbf{0}$ is always non-negative, we have

$$\mathrm{tr}\,[(\mathbf{A}+\mathbf{B})^+\,(\mathbf{A}+\mathbf{B})] \geqslant 0, \tag{42}$$

and the equality sign holds only if the matrix itself is $\mathbf{0}$, i.e. in the case if (41) holds.

Expanding the product in the square bracket and making use of the invariance of the trace by transposing and of (39) we obtain

$$\begin{aligned}\mathrm{tr}\,[(\mathbf{A}+\mathbf{B})^+\,(\mathbf{A}+\mathbf{B})] &= \mathrm{tr}\,(\mathbf{A}^+\mathbf{A}+\mathbf{B}^+\mathbf{B}+\mathbf{A}^+\mathbf{B}+\mathbf{B}^+\mathbf{A})\\ &= 2\,\mathrm{tr}\,(\mathbf{A}^+\mathbf{A})+2\,\mathrm{tr}\,(\mathbf{A}^+\mathbf{B}) \geqslant 0,\end{aligned} \tag{43}$$

which supplies already equation (40) and since the equality sign in (42) holds only if (41) is fulfilled, the same is true for (43) and so for (40) too. (In the case of non-real matrices the real part of tr $(\mathbf{A}^{+*}\mathbf{B})$ comes into play, but the same is used also in the equations used later.)

As a consequence of this theorem we obtain the following extension of it. Let $\mathbf{A}$ and $\mathbf{B}$ denote two matrices, which have different deviations from the matrix $\mathbf{0}$, i.e.

$$d = \mathrm{tr}\,(\mathbf{A}^+\mathbf{A}) \neq \mathrm{tr}\,(\mathbf{B}^+\mathbf{B}) = c. \tag{44}$$

Let us define the positive number λ by

$$c/d = \lambda^2. \tag{45}$$

We are interested in finding the largest negative value of tr $(\mathbf{A}^+\mathbf{B})$. According to (44) and (45) the matrices $\lambda\mathbf{A}$ and $\mathbf{B}$ fulfil the requirement (39), since

$$\mathrm{tr}\,[(\lambda\mathbf{A})^+\,(\lambda\mathbf{A})] = \lambda^2\,\mathrm{tr}\,(\mathbf{A}^+\mathbf{A}) = \mathrm{tr}\,(\mathbf{B}^+\mathbf{B}). \tag{46}$$

So, (40), holds for $\lambda\mathbf{A}$ and $\mathbf{B}$ and the equality sign gives the largest negative value of λ tr $(\mathbf{A}^+\mathbf{B})$ in the case, if

$$\mathbf{B} = -\lambda\mathbf{A}, \tag{47}$$

when

$$\mathrm{tr}\,(\mathbf{A}^+\mathbf{B}) = (1/\lambda)\,\mathrm{tr}\,[(\lambda\mathbf{A})^+\,\lambda\mathbf{A}] = -\lambda\,\mathrm{tr}\,(\mathbf{A}^+\mathbf{A}). \tag{48}$$

Further use of (44) and (45) gives now

$$\mathrm{tr}\,(\mathbf{A}^+\mathbf{B}) = -\surd\{\mathrm{tr}\,(\mathbf{A}^+\mathbf{A})\,\mathrm{tr}\,(\mathbf{B}^+\mathbf{B})\}. \tag{49}$$

Returning now to our physical problem, we have stated that equations (29), (30) and (31) supply the exact solution of our problem. However, we have only some approximation of the solution, $\mathbf{R}_0$. The question is what is the deviation of this solution from the exact one, and how could we refine the approximation.

According to the previous discussions, it is an obvious step to take as the measure of the error made by the approximation

$$D = \operatorname{tr}[[(\mathbf{HR}-\mathbf{RH})-(\mathbf{HR}_0-\mathbf{R}_0\mathbf{H})]^+[(\mathbf{HR}-\mathbf{RH})-(\mathbf{HR}_0-\mathbf{R}_0\mathbf{H})]] \tag{50}$$

for equation (29),

$$d = \operatorname{tr}[[(\mathbf{R}^2-\mathbf{R})-(\mathbf{R}_0^2-\mathbf{R}_0)]^+[(\mathbf{R}^2-\mathbf{R})-(\mathbf{R}_0^2-\mathbf{R}_0)]] \tag{51}$$

for equation (30) and

$$\Delta = [(\operatorname{tr}\mathbf{R}-s)-(\operatorname{tr}\mathbf{R}_0-s)]^2 \tag{52}$$

for equation (31). However, since equations (29), (30) and (31) hold for the exact solution, (50), (51) and (52) may be written

$$D = \operatorname{tr}[(\mathbf{HR}_0-\mathbf{R}_0\mathbf{H})^+(\mathbf{HR}_0-\mathbf{R}_0\mathbf{H})], \tag{53}$$

$$d = \operatorname{tr}[(\mathbf{R}_0^2-\mathbf{R}_0)^+(\mathbf{R}_0^2-\mathbf{R}_0)] \tag{54}$$

and

$$\Delta = (\operatorname{tr}\mathbf{R}_0-s)^2. \tag{55}$$

4. Construction of a better approximation

Let us suppose that we have an approximation $\mathbf{R}_0$, which defines, by (53), (54) and (55), the values of D, d and Δ. We investigate what additive change $\delta\mathbf{R}$, in the matrix will make the approximation approach the exact value in the best way. Since the exact matrix $\mathbf{R}$ gives both D and d the value 0, one has to make such a change that *decreases* the non-negative values D and d to the largest extent and similarly for Δ. In the calculations we shall introduce the following notation

$$\mathbf{A}_0 = \mathbf{HR}_0-\mathbf{R}_0\mathbf{H}, \tag{56}$$

$$\mathbf{B}_0 = \mathbf{HA}_0-\mathbf{A}_0\mathbf{H}, \tag{57}$$

$$\mathbf{C}_0 = \mathbf{HB}_0-\mathbf{B}_0\mathbf{H}, \tag{58}$$

$$\mathbf{a}_0 = \mathbf{R}_0^2-\mathbf{R}_0, \tag{59}$$

$$\mathbf{b}_0 = \mathbf{R}_0^+\mathbf{a}_0+\mathbf{a}_0\mathbf{R}_0^+-\mathbf{a}_0, \tag{60}$$

$$\mathbf{c}_0 = \mathbf{R}_0\mathbf{b}_0+\mathbf{b}_0\mathbf{R}_0-\mathbf{b}_0, \tag{61}$$

$$\mathbf{e}_0 = \mathbf{Hb}_0-\mathbf{b}_0\mathbf{H}, \tag{62}$$

$$\mathbf{f}_0 = \mathbf{R}_0\mathbf{B}_0+\mathbf{B}_0\mathbf{R}_0-\mathbf{B}_0, \tag{63}$$

$$\mathbf{g}_0 = \mathbf{B}_0\mathbf{b}_0+\mathbf{b}_0\mathbf{B}_0. \tag{64}$$

With this notation, if we replace $\mathbf{R}_0$ by $\mathbf{R}_0+\delta\mathbf{R}$ in (53), (54) and (55), then the change of D, d and Δ is

$$\delta D = 2\operatorname{tr}[\mathbf{A}_0^+(\mathbf{H}\delta\mathbf{R}-\delta\mathbf{RH})]+\operatorname{tr}[(\mathbf{H}\delta\mathbf{R}-\delta\mathbf{RH})^+(\mathbf{H}\delta\mathbf{R}-\delta\mathbf{RH})], \tag{65}$$

$$\begin{aligned}\delta d = {} & 2\operatorname{tr}[\mathbf{a}_0^+(\mathbf{R}_0\delta\mathbf{R}+\delta\mathbf{RR}_0-\delta\mathbf{R})] \\ & +\operatorname{tr}[(\mathbf{R}_0\delta\mathbf{R}+\delta\mathbf{RR}_0-\delta\mathbf{R})^+(\mathbf{R}_0\delta\mathbf{R}+\delta\mathbf{RR}_0-\delta\mathbf{R})+2\mathbf{a}_0^+(\delta\mathbf{R})^2] \\ & +2\operatorname{tr}[\mathbf{R}_0\delta\mathbf{R}+\delta\mathbf{RR}_0-\delta\mathbf{R})^+(\delta\mathbf{R})^2]+\operatorname{tr}[(\delta\mathbf{R}^+)^2(\delta\mathbf{R})^2]\end{aligned} \tag{66}$$

and

$$\delta\Delta = 2(\operatorname{tr}\mathbf{R}_0-s)\operatorname{tr}\delta\mathbf{R}+(\operatorname{tr}\delta\mathbf{R})^2. \tag{67}$$

If we regard the zero approximation $\mathbf{R}_0$ as already being near to the exact matrix, $\mathbf{R}$, we may regard $\delta\mathbf{R}$ as a small matrix, i.e. all the elements of which are small, and neglect the products of these small quantities. So, neglecting the second order terms in (65), (66) and (67) we obtain for the corrections

$$\delta_1 D = 2\,\mathrm{tr}\,[\mathbf{A}_0^+(\mathbf{H}\delta\mathbf{R}-\delta\mathbf{R}\mathbf{H})], \tag{68}$$

$$\delta_1 d = 2\,\mathrm{tr}\,[\mathbf{a}_0^+(\mathbf{R}_0\delta\mathbf{R}+\delta\mathbf{R}\mathbf{R}_0-\delta\mathbf{R})] \tag{69}$$

and

$$\delta_1 \Delta = 2\,(\mathrm{tr}\,\mathbf{R}_0-s)\,\mathrm{tr}\,\delta\mathbf{R}. \tag{70}$$

Equation (68) can be transformed by using equations (38) and (57) and taking into account the Hermitian character of $\mathbf{H}$, into

$$\delta_1 D = 2\,\mathrm{tr}\,[\mathbf{B}_0^+\,\delta\mathbf{R}], \tag{71}$$

similarly (69) takes the form by equations (60) and (38)

$$\delta_1 d = 2\,\mathrm{tr}\,[\mathbf{b}_0^+\,\delta\mathbf{R}], \tag{72}$$

and finally (70) may be written

$$\delta_1 \Delta = 2\,\mathrm{tr}\,[(\mathrm{tr}\,\mathbf{R}_0-s)\,\mathbf{I}_{(m^n)}\,\delta\mathbf{R}]. \tag{73}$$

The best approximation for (29) is achieved if $\delta_1 D$ has its largest negative value (since the positive D has to be reduced to 0). According to (47) this can be supplied by choosing

$$\delta\mathbf{R} = -\lambda\mathbf{B}_0, \tag{74}$$

where λ is a non-negative number which can be determined by a minimization process for δD, taking into account the higher order terms too.

The best approximation for (30) is achieved if $\delta_1 d$ has its largest negative value, which is obtained similarly, if

$$\delta\mathbf{R} = -\mu\mathbf{b}_0, \tag{75}$$

where μ is a non-negative number to be determined by minimizing δd, using the second order terms too.

Finally, the best approximation for (31) is obtained if $\delta_1\Delta$ has its largest negative value, which is attained, if

$$\delta\mathbf{R} = -\nu(\mathrm{tr}\,\mathbf{R}_0-s)\,\mathbf{I}_{(m^n)}. \tag{76}$$

Clearly the conditions (74), (75) and (76) are generally incompatible and only in very special cases can they be united. In general, therefore, we have to choose a compromise between the three choices for $\delta\mathbf{R}$. Let us see what is the change if we make a choice

$$\delta\mathbf{R} = -\lambda\mathbf{B}_0-\mu\mathbf{b}_0-\nu(\mathrm{tr}\,\mathbf{R}_0-s)\,\mathbf{I}_{(m^n)}, \tag{77}$$

where λ, μ and ν are non-negative numbers.

The determination of λ, μ and ν can be performed in the following way. Taking the total δD (i.e. without neglecting the higher order terms) in the form (65) and substituting (77) into it, we obtain a form for δD which is quadratic in λ. This form of δD contains μ and ν as *parameters*. The minimization of δD can be made for λ, which gives a linear equation for λ, containing the parameters μ and ν. Similarly, δd given by equation (66) is a polynomial of the fourth order in μ, containing in addition λ and ν as *parameters*. The minimization of δd results in a third-order equation for μ containing λ and ν as parameters.

Finally, $\delta\Delta$ given by equation (67) is a quadratic form again in ν containing λ and μ as *parameters*. Its minimization prescribes the solution of a linear equation again.

These three equations for λ, μ and ν, each containing the other variables as parameters, can be solved as a simultaneous system of equations and we obtain so the values of λ, μ and ν. Performing the calculations sketched above gives for (65), by the substitution of (77) and using the notation (56) to (64),

$$\begin{aligned}\delta D = {} & \lambda^2 \operatorname{tr}[\mathbf{C}_0^+ \mathbf{C}_0] - 2\lambda\{\operatorname{tr}[\mathbf{B}_0^+ \mathbf{B}_0] - \mu \operatorname{tr}[\mathbf{e}_0^+ \mathbf{e}_0]\} \\ & + \mu^2 \operatorname{tr}[\mathbf{e}_0^+ \mathbf{e}_0] - 2\mu \operatorname{tr}[\mathbf{B}_0^+ \mathbf{b}_0] - 2\nu(\operatorname{tr}\mathbf{R}_0 - s)\operatorname{tr}\mathbf{B}_0 \end{aligned} \tag{78}$$

and for (67)

$$\begin{aligned}\delta\Delta = {} & \nu^2 m^2(\operatorname{tr}\mathbf{R}_0 - s)^2 - 2\nu m^n(\operatorname{tr}\mathbf{R}_0 - s)(\operatorname{tr}\mathbf{R}_0 - s - \lambda \operatorname{tr}\mathbf{B}_0 - \mu \operatorname{tr}\mathbf{b}_0) \\ & + \lambda^2(\operatorname{tr}\mathbf{B}_0)^2 + 2\lambda\mu \operatorname{tr}\mathbf{B}_0 \operatorname{tr}\mathbf{b}_0 + \mu^2(\operatorname{tr}\mathbf{b}_0)^2 - 2\lambda(\operatorname{tr}\mathbf{R}_0 - s)\operatorname{tr}\mathbf{B}_0 - 2\mu(\operatorname{tr}\mathbf{R}_0 - s)\operatorname{tr}\mathbf{b}_0, \end{aligned} \tag{79}$$

whereas for (66) we obtain by neglecting terms of higher order than the second

$$\begin{aligned}\delta_2 d = {} & \mu^2\{\operatorname{tr}[\mathbf{c}_0^+ \mathbf{c}_0] + 2\operatorname{tr}[\mathbf{a}_0^+ \mathbf{b}_0^2]\} - 2\mu\{\operatorname{tr}[\mathbf{b}_0^+ \mathbf{b}_0] - \lambda(\operatorname{tr}[\mathbf{f}_0^+ \mathbf{c}_0] \\ & + \operatorname{tr}[\mathbf{a}_0^+ \mathbf{g}_0]) - \nu(\operatorname{tr}\mathbf{R}_0 - s)(\operatorname{tr}[\mathbf{c}_0^+(2\mathbf{R}_0 - \mathbf{I}_{(m^n)})] + 2\operatorname{tr}[\mathbf{a}_0^+ \mathbf{b}_0])\} \\ & + \lambda^2(\operatorname{tr}[\mathbf{f}_0^+ \mathbf{f}_0] + 2\operatorname{tr}[\mathbf{a}_0^+ \mathbf{B}_0^2]) + 2\lambda\nu(\operatorname{tr}\mathbf{R}_0 - s)(\operatorname{tr}[\mathbf{f}_0^+(2\mathbf{R}_0 - \mathbf{I}_{(m^n)})] \\ & + 2\operatorname{tr}[\mathbf{a}_0^+ \mathbf{B}_0]) + \nu^2(\operatorname{tr}\mathbf{R}_0 - s)^2\{\operatorname{tr}[(2\mathbf{R}_0 - \mathbf{I}_{(m^n)})^+ (2\mathbf{R}_0 - \mathbf{I}_{(m^n)})] \\ & + 2\operatorname{tr}\mathbf{a}_0\} - 2\lambda \operatorname{tr}[\mathbf{B}_0^+ \mathbf{b}_0] - 2\nu(\operatorname{tr}\mathbf{R}_0 - s)\operatorname{tr}\mathbf{b}_0. \end{aligned} \tag{80}$$

Instead of (80) we may use the untruncated form of (66), where terms up to the fourth power in λ, μ and ν appear. The solution for λ, μ and ν is then given by two linear and a third-order equations, which can be solved in the specified calculations easily, but since the general discussion of the behaviour of the roots is more simple in the truncated form (80), we shall use this one, remarking that the results are not too different if the initial approximation is already not too bad.

In this case the minimization described above requires the solution of the following system of linear equations in λ, μ and ν:

$$\lambda \operatorname{tr}[\mathbf{C}_0^+ \mathbf{C}_0] + \mu \operatorname{tr}[\mathbf{e}_0^+ \mathbf{e}_0] = \operatorname{tr}[\mathbf{B}_0^+ \mathbf{B}_0], \tag{81}$$

$$\lambda \operatorname{tr}\mathbf{B}_0 + \mu \operatorname{tr}\mathbf{b}_0 + \nu m^n(\operatorname{tr}\mathbf{R}_0 - s) = \operatorname{tr}\mathbf{R}_0 - s, \tag{82}$$

$$\begin{aligned}\lambda(\operatorname{tr}[\mathbf{f}_0^+ \mathbf{c}_0] + \operatorname{tr}[\mathbf{a}_0^+ \mathbf{g}_0]) + \mu(\operatorname{tr}[\mathbf{c}_0^+ \mathbf{c}_0] + 2\operatorname{tr}[\mathbf{a}_0^+ \mathbf{b}_0^2]) \\ + \nu(\operatorname{tr}\mathbf{R}_0 - s)(2\operatorname{tr}[\mathbf{c}_0^+ \mathbf{R}_0] - \operatorname{tr}\mathbf{c}_0 + 2\operatorname{tr}[\mathbf{a}_0^+ \mathbf{b}_0]) = \operatorname{tr}[\mathbf{b}_0^+ \mathbf{b}_0]. \end{aligned} \tag{83}$$

It is readily seen that the functions δD and $\delta\Delta$ given by (78) and (79) have *minima* at the points given by (81) and (82), since in δD the coefficient of λ^2 and in $\delta\Delta$ that of ν^2 are always non-negative. Investigating $\delta_2 d$ given by (80) the situation is already somewhat more complicated, but without a detailed discussion we may conclude that since $\mathbf{b}_0$ and $\mathbf{c}_0$ are according to (60) and (61) of the same order of magnitude as $\mathbf{a}_0$, the term $\operatorname{tr}[\mathbf{c}_0^+ \mathbf{c}_0]$ is dominating the term $2\operatorname{tr}[\mathbf{a}_0^+ \mathbf{b}_0^2]$, if $\mathbf{a}_0$ is not too large and therefore the coefficient of μ^2 is positive. So we may conclude that if equation (30) is not too roughly approximated, (80) has a *minimum* at the point given by (83). If we are taking into account the whole fourth-order function in μ instead of (80), we easily see that the coefficient of μ^4 is in this function tr $[\mathbf{b}_0^{+2}\mathbf{b}_0^2]$, a non-negative number. This means that either δd has one minimum only, or two minima and one maximum. The numerical procedure can then decide which root is to be taken.

To obtain a better insight into the situation we shall give here a simplified geometrical interpretation of the process. To make possible a representation, we shall regard $\mathbf{R}$ as a vector that can be represented in a two-dimensional space. In this space we shall design for the sake of simplicity the equipotential lines of only two of the functions D, d, and Δ given by (53), (54) and (55) respectively. (See figures 1 to 4.)

In each figure, one of the lines 0–0 denotes the line where one of the quantities is 0 and the other line, where the other quantity vanishes. The required solution is, therefore, the crossing point of these lines. The dashed thick lines denote the approximating steps made by applying minimization to only one of the two quantities in every step. It is obvious that there are regions where this kind of approximation gives an uneven approximation, where the various steps one after the other make an oscillation in the error (see figure 2). The same is true for the case where one (or some) of the equations is re-established exactly in every step (see figure 4). The same is true in the case where we choose only such corrections that one (or some) of the equations is identically fulfilled throughout. In this case the sketch is the same as in figure 3, or 4, only two subsequent steps are to be regarded as one simple step. We may see in figure 4 that even in this case it is possible that the individual steps may spoil the approximation instead of correcting it.

In the case of choosing (77) we always have a direction of the corrective vector in a direction between the two gradients and so one may obtain a more uniform, even approximation. (This is partly assured by applying only non-negative λ, μ and ν.)

The method given here can therefore be summed up as taking the gradients (steepest descents) for all equations not fulfilled exactly, then laying parabolas in the directions of the gradients osculating the respective surfaces at the starting point and determining the vertices of these parabolas. The vertices for the several parabolas coincide for a certain set of the values λ, μ, ν, i.e. for only a certain point in the $\mathbf{R}$ plane and that will be the next approximation for $\mathbf{R}$. Then the process can be iterated and in each step the error can be checked by equations (53) to (55).

From the discussion it is clear that there may be cases when one (or some) of the roots of the system of equations (81) to (83) is negative. This can occur if the surfaces represented in figures 1 to 4 have the feature that the direction of one of the *gradients* is *changing to the opposite* if we move along the direction of the other gradient. This means, therefore, no deviation of the originally stated non-negativity of λ, μ and ν, since the matrices multiplied by them change their signs in these cases also.

Generally, if the exact Hamiltonian $\mathbf{H}$ is known and we have by *any method* obtained some solution, ψ, of the problem, which is either a mathematical approximation of the exact problem, or a physical approximation neglecting or transforming some terms in the Hamiltonian, we can construct $\mathbf{R}_0$ from this solution and with this (53), (54) and (55) will supply the errors of the approximation. However, D, d and Δ give some absolute deviation and not the relative error and therefore even their relative magnitude is not comparable (really the dimension of D is (energy)^2, whereas d and Δ are dimensionless). There is, however, no reason to require such a relative error, since the problem is always to compare

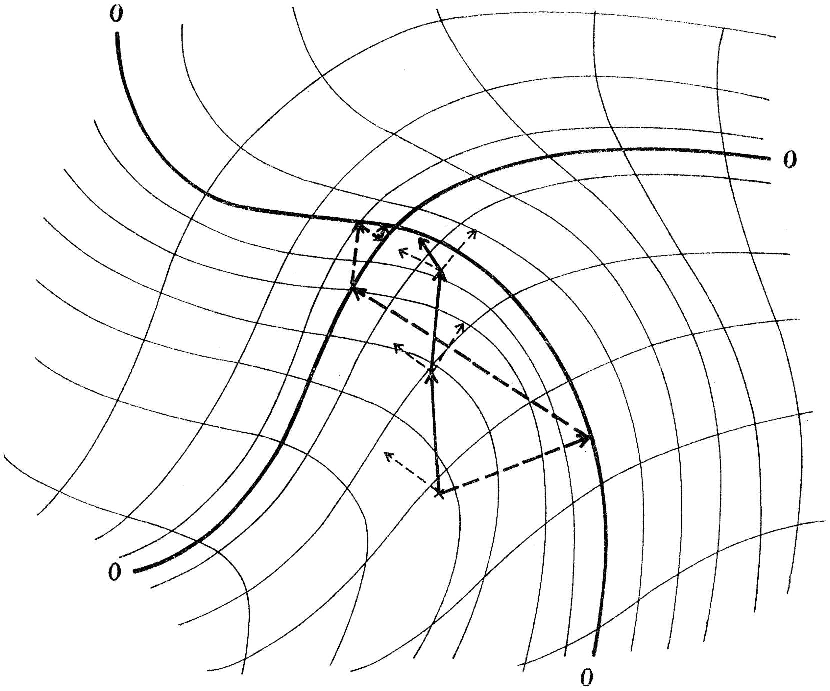

FIGURE 1. — — —, Approximation by minimizing only one quantity in each step; ———, approximation by minimizing both quantities at the same time; - - - - - - -, gradients of the two quantities.

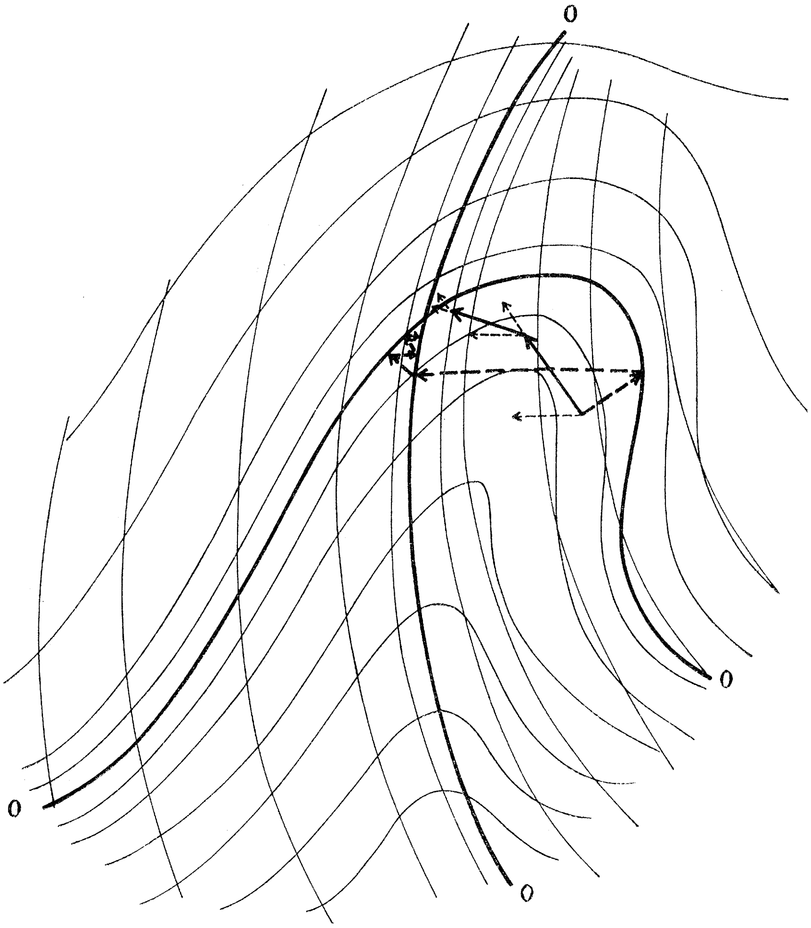

FIGURE 2. The same notation as in figure 1 in a case of an 'unpleasant' shape of the curves.

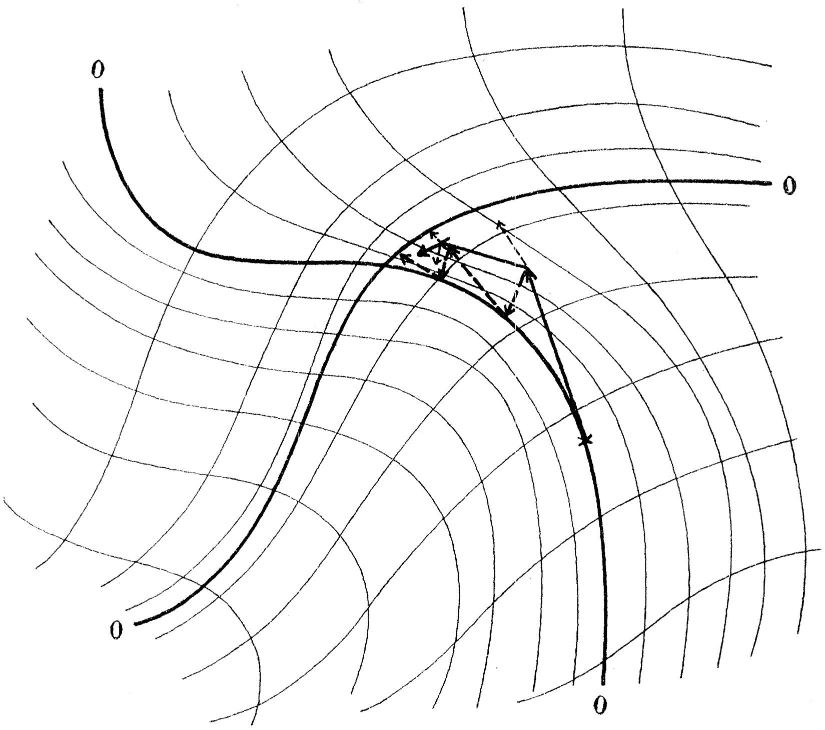

FIGURE 3. The way of approximation if one of the quantities is made to vanish in each step. The notation is the same as in figure 1, but the simultaneous minimization process is begun only after the first common step.

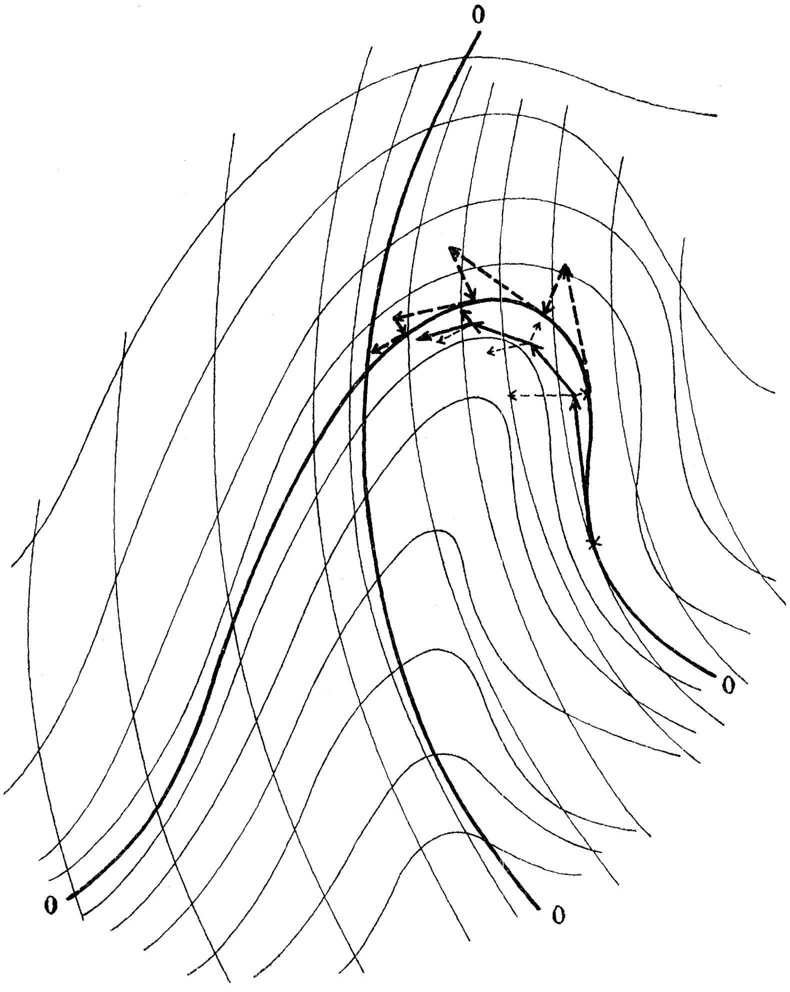

FIGURE 4. The way of approximation if one of the quantities is made to vanish in each step for the case of an 'unpleasant' curve.

calculations for the same system and in this case D, d and Δ may be formed for all concurrent cases. It is, however, possible to normalize D, d and Δ to obtain the relative error; this can be done, arbitrarily, in several different ways, since all the exact solutions give vanishing values for D, d and Δ. This question will be treated in more detail in the next section.

For the sake of completeness we give here another general result (see, for example, McWeeny 1960), stating that any operator, **Q**, has its average value in the state defined by **R**

$$\overline{\mathbf{Q}} = \mathrm{tr}\,(\mathbf{QR}), \tag{84}$$

and so, in particular, the energy of the system is

$$E = \mathrm{tr}\,(\mathbf{HR}). \tag{85}$$

The foregoing discussions can thus be extended to estimating the errors in the expectation values of the various physically interesting operators.

Equations (85) and (53) show an important property worth mentioning here. The energy given by (85) is defined already if we know *only* the diagonal terms in **HR**; however, the D given by (53) depends on *each* element of the matrix **HR**, not only on the diagonal ones. Therefore, in the case that we are interested not only in the energy of the system, the minimization of D in (53) promises better results than the minimization of the energy only.

The suggested method of approximation is also advantageous because it contains only direct matrix operations, thus facilitating application to computers, and further because it is a self-correcting process, since in each step only the quantities of the last step are used.

5. 'Subjective' minimization

The method given in the previous sections is not the best one, because it does not take into account the size of the individual errors and so it minimizes all of them with an equal weight. However, it is quite clear, that if all the errors but one are small, then the most urgent task is to decrease this large error. The reason why we have not done this before is that in general one cannot find any *objective* criterion according to which one shall determine the weights of the various errors. In this section we shall, however, try to give such a determination, but one has to emphasize that the given way of estimating the weights is not unambiguous, it is rather *subjective*. Therefore we call this process the subjective minimization. In some practical cases it has shown exceedingly good results.

First of all we have to normalize the errors in some way. Some of the errors contain only the density matrices, which are dimensionless quantities, but some of them contain **H**, which is a quantity of the dimension of energy. In equation (53) **H** occurs in the second degree. It is therefore advisable (but not unambiguous) to choose as a normalization factor tr$[\mathbf{H}^{+}\mathbf{H}]$, which is a quantity independent of the density matrix and has the same dimensions as (53). So with the notation

$$\alpha = 1/\mathrm{tr}\,[\mathbf{H}^{+}\mathbf{H}], \tag{86}$$

equation (53) should be multiplied by α.

It is sometimes also convenient to make a further normalization for the density matrices, especially if there is a larger number of states included in the calculations and we are interested in the relative errors per state.

Here we make a subjective choice again, giving a normalization factor for all errors

$$\beta = 1/s \tag{87}$$

in the case of the n-electron density matrices, i.e. equations (53) to (55).

The errors being now normalized we have to take into account the size of the errors. This can also be done in several ways too, but the most simple (however subjective) way is to give them weights equal to the normalized errors themselves.

We shall sketch the procedure with equations (53) to (55). With the notation of these equations, we form the following quantity

$$M = \beta[\alpha D\delta D + d\delta_2 d + \Delta\delta\Delta] \tag{88}$$

M is a quadratic function of λ, μ and ν and now we minimize this M according to λ, μ and ν. Instead of equations (81) to (83) we then obtain the following equations:

$$\lambda F_0 + \mu G_0 + \nu(\mathrm{tr}\,\mathbf{R}_0 - s)\,K_0 = \mathrm{tr}\,[\mathbf{B}_0^+\,\mathbf{L}_0], \tag{89}$$

$$\lambda G_0 + \mu M_0 + \nu(\mathrm{tr}\,\mathbf{R}_0 - s)\,N_0 = \mathrm{tr}\,[\mathbf{b}_0^+\,\mathbf{L}_0], \tag{90}$$

$$\lambda K_0 + \mu N_0 + \nu(\mathrm{tr}\,\mathbf{R}_0 - s)\,P_0 = \mathrm{tr}\,\mathbf{L}_0, \tag{91}$$

where the symbols

$$F_0 = \alpha D\,\mathrm{tr}\,[\mathbf{C}_0^+\mathbf{C}_0] + d(\mathrm{tr}\,[\mathbf{f}_0^+\mathbf{f}_0] + 2\,\mathrm{tr}\,[\mathbf{a}_0^+\,\mathbf{B}_0^2]) + \Delta(\mathrm{tr}\,\mathbf{B}_0)^2, \tag{92}$$

$$G_0 = \alpha D\,\mathrm{tr}\,[\mathbf{C}_0^+\mathbf{e}_0] + d(\mathrm{tr}\,[\mathbf{f}_0^+\mathbf{c}_0] + \mathrm{tr}\,[\mathbf{a}_0^+\mathbf{g}_0]) + \Delta\,\mathrm{tr}\,\mathbf{B}_0\,\mathrm{tr}\,\mathbf{b}_0, \tag{93}$$

$$K_0 = d(\mathrm{tr}\,[\mathbf{f}_0^+(2\mathbf{R}_0 - \mathbf{I}_{(m^n)})] + 2\,\mathrm{tr}\,[\mathbf{a}_0^+\mathbf{B}_0]) + m^n\Delta\,\mathrm{tr}\,\mathbf{B}_0, \tag{94}$$

$$L_0 = \alpha D\mathbf{B}_0 + d\mathbf{b}_0 + \Delta(\mathrm{tr}\,\mathbf{R}_0 - s)\,\mathbf{I}_{(m^n)}, \tag{95}$$

$$M_0 = \alpha D\,\mathrm{tr}\,[\mathbf{e}_0^+\mathbf{e}_0] + d\,\mathrm{tr}\,[\mathbf{c}_0^+\mathbf{c}] + 2d\,\mathrm{tr}\,[\mathbf{a}_0^+\mathbf{b}_0^2] + \Delta(\mathrm{tr}\,\mathbf{b}_0)^2, \tag{96}$$

$$N_0 = d(\mathrm{tr}\,[\mathbf{c}_0^+(2\mathbf{R}_0 - \mathbf{I}_{(m^n)})] + 2\,tr\,[\mathbf{a}_0^+\mathbf{b}_0]) + m^n\Delta\,\mathrm{tr}\,\mathbf{b}_0 \tag{97}$$

and

$$P_0 = d(\mathrm{tr}\,[(2\mathbf{R}_0 - \mathbf{I}_{(m^n)})^+(2\mathbf{R}_0 - \mathbf{I}_{(m^n)})] + 2\,\mathrm{tr}\,\mathbf{a}_0) + m^{2n}\Delta \tag{98}$$

were introduced.

Here equation (91) is to be omitted if $\Delta = 0$, since we have divided this equation by the factor $\mathrm{tr}\,\mathbf{R}_0 - s$ which is 0.

A detailed discussion of these equations shows that the solution always gives a minimum in practical cases. Finally, the λ, μ and ν determined from (89) to (91) should be substituted into (77) to obtain the next approximation.

The methods given in this paper are applicable for any quantum-mechanical calculations. In actual cases the explicit evaluation of the matrices appearing in the calculations sometimes give considerable trouble. In the next paper (Hoffmann 1965) some particular applications are given.

I am indebted to Professor C. A. Coulson, F.R.S. for calling my attention to some of the problems presented in this article. I would also express my thanks to Unesco for giving a grant in the field of 'Peaceful uses of atomic energy' which made it possible to prepare this work and to the AB Atomenergi, Studsvik, where this work was performed.

References

Bartlett, J. H. 1955 *Phys. Rev.* **98**, 1067.

Eckart, C. 1930 *Phys. Rev.* **36**, 878.

Frost, A. A., Kellogg, R. E. & Curtis, E. C. 1960 *Rev. Mod. Phys.* **32**, 313.

Fröman, A. & Hall, G. G. 1961 *J. Mol. Spectr.* **7**, 410.

Hartree, D. R. 1928 *Proc. Camb. Phil. Soc.* **24**, 89.

Hartree, D. R. 1946 *Rep. Progr. Phys.* **11**, 113.

Hartree, D. R. 1957 *The calculation of atomic structures.* New York: Wiley.

Hoffmann, T. A. 1965 *Phil. Trans.* A, **257**, 327

Kinoshita, T. 1959 *Phys. Rev.* **115**, 366.

Löwdin, P. O. 1955*a* *Phys. Rev.* **97**, 1474.

Löwdin, P. O. 1955*b* *Phys. Rev.* **97**, 1490.

Löwdin, P. O. 1956 *Advanc. Phys.* **5**, 1.

Löwdin, P. O. 1960 *Amer. Rev. Phys. Chem.* **11**, 107.

McWeeny, R. 1956*a* *Proc. Roy. Soc.* A, **235**, 496.

McWeeny, R. 1956*b* *Proc. Roy. Soc.* A, **237**, 355.

McWeeny, R. 1960 *Rev. Mod. Phys.* **32**, 335.

von Neumann, J. 1932 *Mathematische Grundlagen der Quantenmechanik.* Berlin: Springer.

Preuss, H. & Trefftz, E. 1957 *Phys. Rev.* **107**, 1282.

Preuss, H. 1958 *Z. Naturforsch.* **13***a*, 439.

Shull, H. & Löwdin, P. O. 1958 *Phys. Rev.* **110**, 1466.

Temple, G. 1928 *Proc. Roy. Soc.* A, **119**, 276.

Trefftz, E. 1959 *Z. Naturforsch.* **14***a*, 708.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES