JOURNAL OF
MATHEMATICAL
PHYSICS

VOLUME 13, NUMBER 3 MARCH 1972

A Class of Nonlinear Realizations of the Poincaré Group*

M. Dresden
Institute for Theoretical Physics, State University of New York, Stony Brook, New York 11790

and

A. Albano
Bryn Mawr College, Bryn Mawr, Pennsylvania
(Received 18 August 1971)

A class of nonlinear space-time transformations is exhibited, which forms a nonlinear realization of the Poincaré group. The transformations leave the expression $I = x^2 + f(x^\mu/x^0)$ invariant; f is an arbitrary function of the ratios x^μ/x^0. The infinitesimal generators are constructed as differential operators in the Minkowski space. The transformations are defined only in a restricted region $\mathfrak{N}$ (the allowed region) of the Minkowski space. By introducing auxiliary variables, the transformations can be recast in their usual linear form; this, however, is in general possible only in a region $\mathfrak{L}$ (the linear region) which is different from $\mathfrak{N}$. The region structure is analyzed in general and given explicitly for a special form of the function f. Among the physical ideas suggested by the nonlinear formalism is the notion of "relativity of coincidence." This expresses the fact that events coincident (or having arbitrary small Minkowski separation) in one frame of reference will *not* be coincident (or will have finite Minkowski separation) in a transformed frame.

1. BACKGROUND AND MOTIVATION

In some previous papers, a class of nonlinear space-time transformations was constructed, which turned out to be realizations of the Lorentz group (Refs. 1 and 2). These papers were quite preliminary and tentative in character, certain results were announced without proofs, ideas were suggested, but no organized development was presented. The present paper contains a systematic exposition (with proofs) of a class of nonlinear realizations of the Poincaré group and some suggestions concerning their relevance.

Since both the formal development and the general goals were materially altered during the study, the following comments may help to place this investigation in its proper context.

The initial purpose of this study was the introduction of intrinsic limitations (presumably of a quantum character) of space-time measurements in the theoretical structure. Many such attempts have been made (Halpern and Atkinson, Ref. 3, introduce a change in topology rather than in metric. See Ref. 1 for a very incomplete list of such efforts). The difficulty always is that the experiments neither demand nor exclude such modifications, while the resulting theories tend to become cumbersome and arbitrary. However, the results of high energy scattering do *not* give information about small space-time regions (Ref. 4) in situations where enough energy is available to explore such regions. The participating objects appear to fragment in various pieces (Ref. 5). The resulting picture is at least consistent with an intrinsic limit in the measurement of space-time regions.

Another difficulty is that modifications of the space-time structure in microscopic domains would appear to demand a modification of the Poincaré group, while it is yet required that for large space-time separations (note that even the notion of "large" is not really defined in the context of the Poincaré group) the usual invariance and causality principles are operative. The compatibility of the confinement of such modifications and the existence of an invariance *group* has been a perennial source of trouble. To investigate this point, a particular modification of Lorentz invariance was suggested (Ref. 6). The basic invariant, which replaces x^2, is[7]

$$I = x^2 + f(x^i/x^0). \tag{1}$$

Here $x^2 = (x^0)^2 - (\mathbf{x})^2$; the space-time coordinates are x^μ, $\mu = 0, 1, 2, 3$; $x^0 = t$ is the time coordinate; $i = 1, 2, 3$ denotes the space coordinates, f is a (so far) arbitrary function of the *ratios* x^i/x^0, which in some sense is supposed to express the limitations on the measurability of space time domains. It is not *a priori* obvious that the invariant (1) allows this interpretation. It was shown, however, in Ref. 1 that the transformations leaving (1) invariant form a group. In

general (for arbitrary f) the transformations are *not* defined over all of Minkovski space X(X is the set $-\infty < x^\mu < +\infty$) but only in an allowed region $\mathfrak{A}$. The forbidden or excluded region $\overline{\mathfrak{A}} \equiv X - \mathfrak{A}$ was interpreted as a domain in which no space–time measurements could be carried out. The size of this domain was related to the limits of measurability. The relation between $\mathfrak{A}$ and f was hinted at in previous studies,[8] but the precise connection is more intricate; it is given in this paper, Sec. 3.

The group of transformations which leaves (1) invariant is isomorphic to the homogeneous Lorentz group. Since the transformations are nonlinear, this is an example of a (nonlinear) realization[9] of the Lorentz group by means of nonlinear transformations on a subset $\mathfrak{A}$ of the Minkowski space X. Even though the original purpose was the study of possible modifications of local space–time structure and Lorentz invariance, the implementation of this idea led to nonlinear realizations of the Lorentz group. It is in this sense that the emphasis of the investigation is shifted from the physical implications of modified invariance requirements to the physical significance of nonlinear realizations of Lorentz (or Poincaré) invariant theories. (A deeper analysis might well show that these are in fact not unrelated). This reorientation in the direction of the physical aspects of the nonlinear realizations is especially interesting since nonlinear realizations of internal symmetry groups have been particularly important in particle physics (Ref. 10).

For example, the chiral group $SU(2) \otimes SU(2)$ does *not* possess a linear three-dimensional representation; it does possess a linear four-dimensional representation as well as a nonlinear three-dimensional realization. This suggests a direct physical role for the three-dimensional nonlinear realization as the transformation law for the pion field (Ref. 11). It was noted by Meetz (Ref. 12) that the fundamental three-dimensional nonlinear realization of the chiral group is the set of transformations which leaves the metric of a three-dimensional space of constant curvature invariant. By combining these ideas, it is seen that in this case a direct physical significance can be attributed to a set of nonlinear transformations which are a (nonlinear) realization of a geometrical invariance group. Even though no such interpretation has as yet been given for the nonlinear realizations of the Lorentz group, the analogy is certainly suggestive. In spite of this observation, the role (if any) of the nonlinear realization is still far from clear. One of the problems is that, to obtain an interesting (and testable) theory, a commitment has to be made regarding a physical interpretation of the nonlinear realizations. In order to obtain a sensible physical interpretation, on the other hand, the formalism must be developed sufficiently to provide a framework which allows the convenient formulation of a variety of physical ideas. The formalism itself imposes restrictions on the physical interpretation.

This paper is primarily devoted to the construction of a special class of nonlinear realization of the Lorentz and Poincaré groups. The formalism developed can be handled rather easily and avoids the tedious manipulations of previous papers. Although the development of the formalism is the main concern here, it will be clear throughout this paper that a specific physical interpretation must eventually be made. The considerations in this paper are all preparations to that end—several suggestions and hints scattered through the paper all point in that direction.

Section 2 contains the explicit form of the nonlinear transformations for both the Lorentz group and the Poincaré group. The group property is demonstrated, the explicit form of the infinitesimal generators is given, and it is established that the commutation rules of the infinitesimal generators are those of the Poincaré algebra.

The region structure, their boundaries in particular, are analyzed in Sec. 3. The relationship between the nature of f and the character of the region is obtained in that section. Two fairly simple choices for f are treated in some detail.

The last section, 4, contains a number of disconnected comments, further results, and conjectures. Perhaps most important is the suggestion of the "relativity of coincidence," which asserts that coincidence becomes a *frame-dependent notion*, so that events occurring at the same space-time point in one system would not be coincident in a transformed system. This takes rather careful discussion, but it appears that the formalism allows (in fact demands) this somewhat curious *notion*.

2. THE GROUP CHARACTER OF THE NONLINEAR TRANSFORMATIONS

A. The Homogeneous Group

The question to be discussed in this section is the existence and character of the transformations in the Minkowski space X, which leave I invariant,

$$I = x^2 + f(x^i/x^0). \tag{1}$$

It is sometimes useful to introduce $\alpha^i(x)$ as

$$\alpha^i(x) = x^i/x^0. \tag{2a}$$

α thus represents a particular combination of the space-time coordinates. If an ordinary homogeneous Lorentz transformation Λ is carried out, α^i constructs the same combination of transformed quantities according to

$$\alpha^i(\Lambda x) = (\Lambda x)^i/(\Lambda x)^0. \tag{2b}$$

The function f is thus a function of the α^i variables; it will be assumed to be a real, bounded function of these variables. Although not strictly necessary, it will be assumed further that f depends only on the absolute value of

$$\alpha = \sqrt{[\alpha(x)]^2}. \tag{2c}$$

Certain regularity properties of f are also needed, but it is not necessary to assume that f is everywhere continuous. For physical applications, a form of f where the limits of $f(\alpha)$ as $|\alpha| \to 1$ from above ($|\alpha| > 1$) and below ($|\alpha| < 1$) are different is particularly interesting. Such functions are included in the class considered.

With these general stipulations for the function f, consider the transformations[13] $Q(\Lambda)$:

$$Q(\Lambda): x'^\mu = (\Lambda^\mu_\nu x^\nu) R(x|\Lambda) \equiv (\Lambda x)^\mu R(x|\Lambda), \tag{3a}$$

$$\begin{aligned} R^2(x|\Lambda) &= 1 + \frac{1}{x^2}\left[f(\alpha(x)) - f(\alpha(\Lambda x))\right] \\ &= 1 + \frac{1}{x^2}\left[f\left(\frac{x^i}{x^0}\right) - f\left(\frac{(\Lambda x)^i}{(\Lambda x)^0}\right)\right]. \end{aligned} \tag{3b}$$

The transformation $Q(\Lambda)$ is formally similar to a scale transformation; the scaling factor R depends on both x and Λ as (3b) shows. It follows from the assumed boundedness of f that R^2, as given by (3b), could be singular only on the light cone where $x^2 = 0$. It is easy to show that if f possesses a finite derivative at $\alpha = 1$, the expression for R^2 is finite on the light cone. In the more general case, that $f(\alpha)$ possesses finite discontinuities at $|\alpha| = 1$ (with finite but different right and left derivatives at $|\alpha| = 1$), R^2 is again finite on the light cone. Thus with the restrictions imposed on f the radical R is everywhere nonsingular. The transformations (3) form a group, which leaves I invariant. The proof of the invariance of I is straightforward. (3a) shows immediately that

$$x'^i/x'^0 = (\Lambda x)^i/(\Lambda x)^0. \tag{4a}$$

Constructing $x'^2 = g_{\mu\nu}x'^\mu x'^\nu$ from (3a) gives

$$(x')^2 = x^2\left\{1 + \frac{1}{x^2}\left[f\left(\frac{x^i}{x^0}\right) - f\left(\frac{x'^i}{x'^0}\right)\right]\right\}. \tag{4b}$$

Here (3b) and (4a) have been used. (4b) expresses the invariance of I.

To show that the transformations Q form a group, it is only necessary to verify the group postulates; just the closure property needs some discussion. Let x^μ be subjected to two successive transformations: $Q(\Lambda_1)$ takes $x^\mu \to x'^\mu$, $Q(\Lambda_2)$ takes $x'^\mu \to x''^\mu$. Then application of (3a) yields

$$x''^\mu = (\Lambda_{2\ \nu}^{\mu}\Lambda_{1\ \lambda}^{\nu}x^\lambda)R(x|\Lambda_1)R(x'|\Lambda_2) \tag{5}$$

Substituting (3b) in the expressions for R and recognizing that

$$x''^i/x''^0 = (\Lambda_2\Lambda_1 x)^i/(\Lambda_2\Lambda_1 x)^0 \tag{6}$$

leads to the important relation

$$R(x\,|\,\Lambda_2\Lambda_1) = R(x\,|\,\Lambda_1)R(x'\,|\,\Lambda_2). \tag{7}$$

Combining (7) and (5) and using the properties of the Lorentz matrices gives

$$Q(\Lambda_2\Lambda_1) = Q(\Lambda_2)Q(\Lambda_1). \tag{8}$$

(8) shows that the transformations $Q(\Lambda)$ do indeed provide a realization of the Lorentz group.

There is a standard procedure of constructing the infinitesimal generators of nonlinear coordinate transformations. Let the transformations in an n dimensional space $x_1 \cdots x_n$ be given by

$$x'_i = F_i(x_1 \cdots x_n, \lambda_1 \cdots \lambda_s). \tag{9a}$$

$(\lambda_1 \cdots \lambda_s)$ are parameters of the transformation; then the infinitesimal generators are the operators

$$M_K = \sum_{i=1}^{n}\left(\frac{\partial F_i}{\partial \lambda_K}\right)_{\lambda=0}\frac{\partial}{\partial x_i}, \qquad K = 1 \cdots s. \tag{9b}$$

Applying this prescription to the transformation (3) yields after some calculation

$$M_{\mu\nu} = M_{\mu\nu}^{(0)} - (1/2x^2)(M_{\mu\nu}^0 f)D \tag{10}$$

In (10) the $M_{\mu\nu}^{(0)}$ are the usual infinitesimal generators of the homogeneous Lorentz group:

$$M_{\mu\nu}^{(0)} = x_\nu\frac{\partial}{\partial x^\mu} - x_\mu\frac{\partial}{\partial x^\nu}. \tag{11a}$$

D is the dilatation operator:

$$D = x^i\frac{\partial}{\partial x^i} + x^0\frac{\partial}{\partial x^0}. \tag{11b}$$

The straightforward calculation of the commutators of $M_{\mu\nu}$ and $M_{\rho\sigma}$ shows that the commutation rules of the $M_{\mu\nu}$ have the same structure as those of $M_{\mu\nu}^{(0)}$, so that the Lie *algebra* of the operators $M_{\mu\nu}$ is the same as that of the $M_{\mu\nu}^{(0)}$ (the Lorentz generators).

The arguments just presented establish the group character of the transformations $Q(\Lambda)$ [Eq. (3)]. It is, however, clear from the square root structure [in (3b)] that there may well be points x and transformations Λ such that $R^2(x|\Lambda)$ is negative. In that case, x'^μ is imaginary. Since the x^μ refer to space–time points (they are either directly observable as in classical theory or they label field operators), it is a reasonable requirement that they should be real. Consequently, the definition of the transformations must be *restricted* to those space–time points which will yield real images. Formulated more precisely, define $\mathfrak{A}$ as that subset of Minkovski space such that $R^2(x\,|\,\Lambda)$ is nonnegative for *all* Λ:

$$= \{x^\mu : R^2(x|\,\Lambda) \geq 0 \quad \forall\Lambda\}. \tag{12}$$

The excluded or forbidden region is $\bar{\mathfrak{A}}$, which is that subset of X where $R^2(x|\Lambda)$ is negative for *some* Λ. The transformation Q is thus restricted to $\mathfrak{A}$; it, however, needs to be shown that the restriction of Q to $\mathfrak{A}$ does not destroy the group structure. In particular, it needs to be shown that if $x \in \mathfrak{A}$, $x' \in \mathfrak{A}$ $\forall\Lambda$. To see this, consider Eq. (7). If $x \in \mathfrak{A}$, $R(x\,|\,\Lambda_2\Lambda_1)$ and $R(x\,|\,\Lambda_1)$ are both real for all Λ_1 and Λ_2. It follows from (7) that $R(x'\,|\,\Lambda_2)$ is real for all Λ_1 and all Λ_2. Since x' is the image of x under $Q(\Lambda_1)$, it is seen that $R(x'\,|\,\Lambda_2)$ is real for all Λ_2, hence $x' \in \mathfrak{A}$. It can be shown similarly that if x ranges over all of $\mathfrak{A}$, so does the image x'.

It is possible to introduce auxiliary variables ξ^μ so that the transformations $Q(\Lambda)$ reassume their linear form. Define the transformation[14] $U: x \to \xi$:

$$\xi^\mu = x^\mu\sqrt{1 + f(\alpha(x))/x^2}. \tag{13a}$$

(13) is so designed that

$$\xi^2 = x^2 + f(\alpha(x)) = I. \tag{14}$$

(13) can be easily inverted (U^{-1} takes $\xi \to x$):

$$x^\mu = \xi^\mu\sqrt{1 - f(\alpha(\xi))/\xi^2}. \tag{13b}$$

It can be established by direct calculation that if $Q(\Lambda)$ takes $x \to x'$ the corresponding transformation from $\xi \to \xi'$ is just the Lorentz transformation Λ. It should be noted that [in contrast to $Q(\Lambda)$] the transformation U *can* be singular. Further, the transformation U is defined in a subset $\mathfrak{L}$ of the Minkovski space (the "linearizable" region). As (13) shows

$$\mathfrak{L} = \{x;\ 1 + f(\alpha(x))/x^2 \geq 0\}. \tag{15}$$

The collection of real ξ^μ values obtained via (13a) will be called the set Ξ. In general this set does not contain all ξ^μ values ($-\infty < \xi^\mu < +\infty$). Ξ is the (real) image of X under U. The region structure will be discussed in more detail in Sec. 3.

It is finally interesting to see that if Λ is a pure space rotation [so that $(\Lambda x)^0 = x^0$, $\Lambda r = r$], the last two terms in R^2[in (3b)] cancel, so that the transformations Q become linear again. This is a common feature of the nonlinear realizations in particle physics; when restricted to a subgroup, the nonlinear realizations become linear again.

B. The Inhomogeneous Transformations

It is also possible to construct a set of nonlinear transformations $Q(\Lambda, a)$, which are related to the Poincaré group in the same way that the transformations $Q(\Lambda)$ are related to the Lorentz group. (a is a 4-vector a^μ, characterizing a translation). The transformations $Q(\Lambda, a)$ have the property that, as $f \to 0$, Q reduces to (Λ, a), while (Λ, a) becomes $Q(\Lambda)$ when $a = 0$. Q is given by

$$Q: x^\mu \to x'^\mu = (S^\mu_\nu x^\nu) R(x|S), \tag{16}$$

with

$$S^\mu_\nu x^\nu = \Lambda^\mu_\nu x^\nu + A(x) a^\mu, \tag{16'a}$$

$$A(x) = \sqrt{\frac{x^2}{x^2 + f}}, \tag{16'b}$$

$$R^2(x|S) = 1 + \frac{1}{x^2} f\left(\frac{x^i}{x^0}\right) - \frac{1}{(Sx)^2} f\left(\frac{(Sx)^i}{(Sx)^0}\right). \tag{16'c}$$

The basic result is that the transformations (15) with the definitions (16) form a group, leaving $K(x, y)$ invariant:

$$K(x, y) = [xA^{-1}(x) - yA^{-1}(y)]^2. \tag{17}$$

The proofs are given by verifying the group postulates and by calculating $K(x', y')$. The derivation becomes a routine manipulation once it is recognized that

$$A(x') = R(x|S)A(x), \tag{18a}$$

$$R(x|S_1)R(x'|S_2) = R(x|S_2S_1) \tag{18b}$$

[x' is defined by (15)].

For example,[15] the invariance of K is shown by the following chain of equalities in which (18a) and (15) figure most prominently:

$$\begin{aligned} K(x', y') &= [x'A^{-1}(x') - y'A^{-1}(y')]^2 \\ &= [(Sx)A^{-1}(x) - (Sy)A^{-1}(y)]^2 \\ &= \{[\Lambda(x) + aA]A^{-1}(x) - [\Lambda(y) + aA(y)]A^{-1}(y)\}^2 \\ &= K(x, y). \end{aligned} \tag{19}$$

(18b) is particularly useful for establishing that two successive transformations, $Q(\Lambda_1 a_1)$ takes $x \to x'$, $Q(\Lambda_2 a_2)$ takes $x' \to x''$, can be replaced by a single transformation from $x \to x''$, so that

$$Q(\Lambda_2\Lambda_1, \Lambda_2 a_1 + a_2) = Q(\Lambda_2 a_2)Q(\Lambda_1 a_1). \tag{20}$$

This relationship shows that the transformations (15) form a realization of the Poincaré group. The basic identities (18) facilitate the manipulations involved; they themselves are obtained by substitution of the appropriate formula.

The form of Q_1 shows that if Λ is picked as the *identity* Lorentz transformation, S does not reduce to a translation. The resulting transformation will be called a pseudotranslation T:

$$Tx^\mu = x^\mu + a^\mu A. \tag{21}$$

Since A depends on x [Eq. 16'b)], (21) is not just a translation; it represents a distortion of the space as well. The pseudotranslations form an Abelian subgroup, as (20) shows.

The infinitesimal generators can be obtained following the method outlined in connection with the homogeneous group (9). The calculations get a little longer, but present no difficulty. In addition to the generators $M_{\mu\nu}$ [Eq. (10)], there are now the four generators of the pseudotranslations P_μ:

$$P_\mu = \frac{1}{i}\left(A(x)\frac{\partial}{\partial x^\mu} + \frac{1}{A^2(x)}\frac{\partial A}{\partial x^\mu} D\right) \tag{22}$$

The P_μ are obviously distinct from $(1/i)\partial/\partial x^\mu$. With the explicit operator forms of P_μ and $M_{\mu\nu}$, the various commutation relations may be obtained by explicit calculation. This shows that the algebra of the generators M and P [of the transformations (15)] is precisely the Poincaré algebra,[16] so that the transformations (15) are a nonlinear realization of the Poincaré group.

3. THE REGION STRUCTURE

For the eventual physical utilization of the nonlinear realizations, a knowledge of the characteristics of the regions in which they are defined is essential. The linearizable region $\mathfrak{L}$, defined by (15), is obviously bounded by a surface $B_{\mathfrak{L}}$, given by

$$x^2 + f(\alpha(x)) = I = 0. \tag{23}$$

The boundary is invariant under the nonlinear (homogeneous) transformations; its precise nature depends on the character of f. If f is continuous and satisfies

$$[1/(1 - \boldsymbol{\alpha}^2)]f(\boldsymbol{\alpha}) < 0 \quad \forall \boldsymbol{\alpha}, \tag{24}$$

the surface $B_{\mathfrak{L}}$ is a closed surface enclosing the origin. ($\boldsymbol{\alpha}$ is written frequently instead of α^i.) To prove this statement, write (23) in terms of the α variables; this gives

$$(x^0)^2 = f(\boldsymbol{\alpha})/(1 - \boldsymbol{\alpha}^2). \tag{25}$$

If the condition (24) is satisfied, (25) will yield two finite real roots for x^0 for each value of $\boldsymbol{\alpha}$. Since a fixed $\boldsymbol{\alpha}$ corresponds to a ray through the origin in the Minkovski space ($x^i = \alpha^i x^0$), this states that each ray through the origin intersects the surface $B_{\mathfrak{L}}$ in two real points. This, together with continuity, implies that the surface is closed and encloses the origin. The equation for $B_{\mathfrak{L}}$ can be written in the suggestive form

$$(x^0)^2 + (\mathbf{x})^2 = -[(1 + \boldsymbol{\alpha}^2)/(1 - \boldsymbol{\alpha}^2)]f(\boldsymbol{\alpha}). \tag{26}$$

A. Boundaries

The allowed region $\mathfrak{A}$ was defined by (12); the determination of its boundary is a little less direct than that of $\mathfrak{L}$. Define the (real) numbers M and m by

$$M = \sup_{x^2>0} f(\alpha(x)), \tag{27a}$$

$$m = \inf_{x^2<0} f(\alpha(x)). \tag{27b}$$

M and m are finite by virtue of the conditions imposed on f. If f is assumed to depend on just the absolute value of $\boldsymbol{\alpha}$, M and m are given by

$$M = \sup_{|\alpha|<1} f(\alpha), \tag{28a}$$

$$m = \inf_{|\alpha|>1} f(\alpha). \tag{28b}$$

The surfaces which form the boundary of $\mathfrak{A}$ are

$$x^2 + f(\alpha(x)) = M, \quad x^2 > 0, \tag{29a}$$

$$x^2 + f(\alpha(x)) = m, \quad x^2 < 0. \tag{29b}$$

To derive this result, one observes first that the set of points defined by

$$x^2 > 0, \tag{30a}$$

$$x^2 + f(\alpha(x)) > M \tag{30b}$$

all lie in $\mathfrak{A}$, for consider the expression for $R^2(x|\Lambda)$ for such points,

$$R^2(x|\Lambda) = x^{-2}\{x^2 + f(\alpha(x)) - f(\alpha(\Lambda(x))\} > x^{-2}\{M - f(\alpha(\Lambda x))\} > 0. \quad (31)$$

The first inequality, in (31), follows from (30b) and the positive character of x^2; the second inequality follows from the definition of M. Thus $R^2(x|\Lambda) > 0$ for all points defined by (30); hence, they all lie in $\mathfrak{A}$.

On the other hand, the points for which $x^2 > 0$ and $x^2 + f(\alpha(x)) = M - \epsilon$, $\epsilon > 0$, lie in $\bar{\mathfrak{A}}$. To see this, note that for such points

$$R^2(x|\Lambda) = x^{-2}\{M - \epsilon - f(\alpha(\Lambda(x)))\}. \quad (32)$$

As Λ runs through all Lorentz transformations Λ, α assumes all its values. so that there will be a Λ_0 for which $f(\alpha(\Lambda(x)))$ equals M. For this Λ_0, $R^2(x|\Lambda_0)$ will be negative; hence x belongs to $\bar{\mathfrak{A}}$. The proof of (29b) proceeds in the identical manner. It is very interesting to observe that if f is assumed to be *continuous and* the condition (24) is satisfied (i.e., the boundary surface $B_{\mathfrak{L}}$, encloses the origin), the bounds $m = M = 0$. This, in turn, means by (20) and (23) that the regions $\mathfrak{L}$ and $\mathfrak{A}$ are identical. Consider for simplicity the case where f depends on one variable α. Then (24) shows that

$$f(\alpha) < 0 \quad \text{for } |\alpha| < 1, \quad (33a)$$

$$f(\alpha) > 0 \quad \text{for } |\alpha| > 1, \quad (33b)$$

If f is continuous, one must have $f(1) = 0$. But now it follows from (28) that $M = \sup_{|\alpha|<1} f(\alpha) = 0$ and $m = \inf_{|\alpha|>1}(\alpha) = 0$.

This result requires *both* the continuity of f and the condition (24). A semi-intuitive justification for (24) will emerge from a further consideration of the pseudotranslations.

B. Pseudotranslations

The pseudotranslations were defined by $Q(1, T)$:

$$x^\mu \to x'^\mu = (Tx)^\mu R(x|T), \quad (34a)$$

with

$$T_q x^\mu = x^\mu + q^\mu A \quad (34b)$$

$$R^2(x|T) = 1 + x^{-2}f(\alpha(x)) - (Tx)^{-2}f(\alpha(Tx)). \quad (34c)$$

The reality requirements restrict the translations to the region

$$\mathfrak{A}_1 = \{x;\ R^2(x|T) > 0\ \forall T\}. \quad (35)$$

However, yet another condition has to be satisfied, for $A(x)$ has also to be real to yield real translations. By (16b) this yields

$$1 + x^{-2}f \geq 0. \quad (36)$$

(36) combined with (15) shows that a *necessary* condition for a translation to exist is that $x \in \mathfrak{L}$; the point must lie in the "linearizable region." This is *not* sufficient, for (35) must be satisfied as well to guarantee the possibility of a real translation. Call the set where both (36) and (35) are satisfied $\mathfrak{A}_T$; then the following important result can be obtained: Let[17] a and b be two points in $\mathfrak{A}_T$, with an invariant distance $K(a, b)$ given by (17); then there exists a unique pseudotranslation $T(q)$, which takes $a \to a'$, $b \to b'$, in such a way that

$$K(b, a) = K(b', a') = I(b'). \quad (37)$$

The first equality is obvious from (19); the second is the main result; I is the one point invariant (1).

Proof: Consider (34b) for a sequence of real values of q^μ; by assumption a and b will have real images under $Q(1, T)$. Using (18a), one has for a general value of q

$$a'^\mu A^{-1}(a') = (a^\mu + q^\mu A)R(a|T)[R^{-1}(a|T)A^{-1}(a)] = a^\mu A^{-1}(a) + q^\mu. \quad (38a)$$

Since a^μ is real and $a \in \mathfrak{A}_T$, $A^{-1}(a)$ is real.[18] Hence one can find a real number q_0^μ, such that

$$-q_0^\mu = a^\mu A^{-1}(a). \quad (38b)$$

Study the translation defined by q_0^μ given by (38). It follows directly from (38a) and (38b) that for that translation $a'^\mu A^{-1}(a') = 0$, so that

$$K(b', a') = [a'^\mu A^{-1}(a') - b'^\mu A^{-1}(b')]^2 = (b')^2 + f(b'^i/b'^0). \quad (39)$$

This appears to prove (37). However, a certain amount of care must be exercised since the transformation $Q(1, q_0)$ has certain singular features.

Consider a sequence of numbers q_n^μ which converges to $-q_0^\mu$ given by (38b). Then q_n^μ defines a translation. Define further

$$\epsilon_n^\mu \equiv a^\mu + q_n^\mu A(a). \quad (40)$$

If q_n^μ approaches $q_0^\mu = -a^\mu A^{-1}(a)$, then ϵ_n approaches zero. Conversely, a given ϵ_n^μ defines q_n. Using a sequence of translations, parametrized by q_n^μ(or ϵ_n), yields a sequence of translated points, $x_n'^\mu$, starting from any given point x^μ. (The point x is distinct from a.) The explicit form for $x_n'^\mu$ can easily be written down using (34), and the limit as $\epsilon_n \to 0$ can be taken. This will give the translated point for the particular pseudotranslation characterized by q_0^μ. The result is

$$\lim_{n\to\infty} x_n'^\mu = x'^\mu = x^\mu - a^\mu \frac{A(x)}{A(a)} \times \sqrt{1 + \frac{1}{x^2}f(\alpha(x)) - \left(x - a\frac{A(x)}{A(a)}\right)^{-2} f(\alpha T(x))}. \quad (41)$$

The result is independent of the manner in which the limit $\epsilon_n^\mu \to 0$ (or $q_n^\mu \to q_0^\mu$) is taken. If x does not lie on the light cone, this is a finite well-defined expression. Consequently, the pseudotranslation $T(q_0)$ translates every point in $\mathfrak{A}_T$ to a finite new, unambiguous point. If the same procedure is repeated for the point a itself, it is found that a_n' [which is the point a translated by $T(q_n)$] is again well defined and unambiguous. Taking the limit $\epsilon_n \to 0$ or $q_n^\mu \to q_0^\mu$ (or $\epsilon^\mu \to 0$), requires the evaluation of the quadruple limit

$$a'^\mu = \lim_{\epsilon^\mu \to 0} \sqrt{-\frac{(\epsilon^\mu)^2}{\epsilon^2}} f\left(\frac{\epsilon^i}{\epsilon^0}\right), \quad (42)$$

$$\epsilon^2 = (\epsilon^0)^2 - (\epsilon)^2. \quad (42')$$

Inspection of this limit shows that it is *not* independent of the manner in which the limit is carried out.

For example, if the limit is performed keeping $\epsilon^i/\epsilon^0 = \alpha^i$ fixed and finite, the result is

$$a'^i = \sqrt{-[(\alpha^i)^2/(1-\boldsymbol{\alpha}^2)]f(\boldsymbol{\alpha})} \tag{43a}$$

$$a'^0 = \sqrt{-[1/(1-\boldsymbol{\alpha}^2)]f(\boldsymbol{\alpha})} \tag{43b}$$

[It might be noticed that the condition (24) imposed on f guarantees that a'^i and a'^0 are real.]

On the other hand, the limit could be defined by stipulating that $a'^{(1)}$ is defined by taking the limit $\epsilon^{(1)} \to 0$ first, keeping $\epsilon^{(2)}$, $\epsilon^{(3)}$, $\epsilon^{(0)}$ finite, while $a'^{(2)}$ is obtained by taking the limit $\epsilon^{(2)} \to 0$, first, keeping $\epsilon^{(1)}$, $\epsilon^{(3)}$, $\epsilon^{(0)}$ finite, etc. With this definition of the limiting process, $a'^\mu = 0$. In any case a particular assignment of the limiting process has to be made; the transformation formulas themselves do *not* define a'^μ. The choice made is to define a'^μ as the limit obtained by letting $\epsilon^\mu \to 0$ first, keeping $\epsilon^\nu (\nu \neq \mu)$ finite. With this choice, the translation maps all points in $\mathfrak{A}_T$, to finite points, a' is the new origin in the sense that $K(b', a') = I(a')$; the coordinates a'^μ of a' are $(0, 0, 0, 0)$. This choice means that the region in which the transformations are defined is the set $\mathfrak{A}$ plus the point $(0, 0, 0, 0)$.

C. Two Examples[19]

These general considerations can be illustrated by the use of some examples. It is simplest to work in two dimensions $(x^0, x') = (t, x)$. Consider first

$$f(\alpha) = -(1-\alpha^2)/(1+\alpha^2) = -(t^2-x^2)/(t^2+x^2). \tag{44}$$

This choice for f clearly satisfies (24); f is also continuous. Examination of (44) shows that $m = M = 0$. Consequently, $\mathfrak{L} = \mathfrak{A}$.

The boundary of $\mathfrak{A}$ is given by $I = 0$, or

$$I = (t^2-x^2)[1 - 1/(t^2+x^2)] = 0. \tag{45}$$

Thus the boundary curve is the unit circle with the light cone attached (see Fig. 1). The allowed region $\mathfrak{A}$ is the outside of the unit circle plus the origin; the forbidden region $\overline{\mathfrak{A}}$ is its interior. The quantity A for this case becomes

$$A = \sqrt{1 - 1/(t^2+x^2)}. \tag{46}$$

Clearly, A is real for (x, t) outside the unit circle. Furthermore, $R^2(x|S)$ can be written using (44) and (16c) as

$$R^2(x|S) = 1 - 1/(t^2+x^2) + 1/[(St)^2 + (Sx)^2]. \tag{47}$$

This form of R^2 shows immediately that when

$$t^2 + x^2 > 1, \quad R^2(x|S) > 0 \quad \forall S. \tag{47'}$$

The expression (18b) can be used to show that if x lies outside the unit circle, so does x', for

$$R^2(x|S_1)R^2(x'|S_2) = R^2(x|S_2S_1). \tag{18b}$$

If x lies outside the unit circle, $R^2(x|S_1)$ and $R^2(x|S_2S_1)$ are, by (47a), both positive for all S_1 and S_2; hence $R^2(x'|S_2)$ is likewise positive for all x' and S_2. Thus in this case the allowed region $\mathfrak{A}$, the linearizable region $\mathfrak{L}$, the region where A is real, and the region $\mathfrak{A}_T$ (the region where translations are defined) are all the same; they are all the exterior of the unit circle. It is also simple to introduce the auxiliary variables ξ^μ defined by (13a). Specialized to the present case, the transformation formulas become ($\xi^0 = \tau$, $\xi' = \xi$)

$$U: \quad \xi = x\sqrt{\frac{t^2+x^2-1}{t^2+x^2}}, \quad U^{-1}: \quad x = \xi\sqrt{\frac{\tau^2+\xi^2+1}{\tau^2+\xi^2}},$$

$$\tau = t\sqrt{\frac{t^2+x^2-1}{t^2+x^2}}, \qquad t = \tau\sqrt{\frac{\tau^2+\xi^2+1}{\tau^2+\xi^2}}. \tag{48}$$

All points *outside* (the *open* region) the unit circle in X have a unique image in Ξ; the mapping is one to one with a unique inverse. However, the origin in Ξ is excluded. If one considers as $\mathfrak{A}$ all the points outside and including the unit circle (the closed region), (48) maps $\mathfrak{A}$ into all of Ξ including the origin; however, in that case, the whole unit circle maps into the origin; the mapping is not one to one; the inverse mapping is not unique. So either the mapping U from $X \to \Xi$ is not one to one, or Ξ is a Minkovski space with the origin deleted.

The apparently very similar case

$$f(\alpha) = (1-\alpha^2)/(1+\alpha^2) = (t^2-x^2)/(t^2+x^2) \tag{49}$$

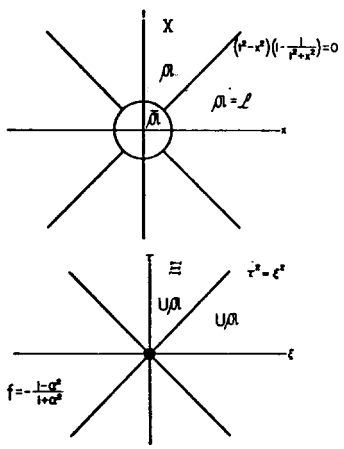


FIG. 1. The region structure for $f = -(1-\alpha^2)/(1+\alpha^2)$.

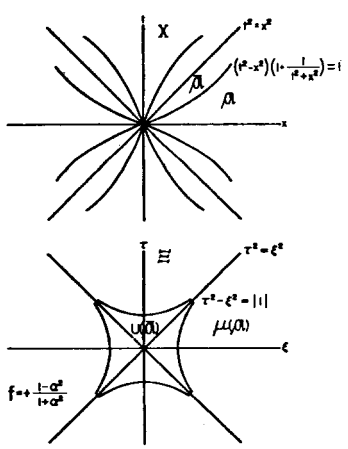


FIG. 2. The region structure for $f = +(1-\alpha^2)/(1+\alpha^2)$.

can be discussed in the same way; the very different results are just recorded. The linearizable region $\mathfrak{L}$ is now all of $X\{1 + [1/(t^2 - x^2)][(t^2 - x^2)/(t^2 + x^2)]$ is always positive$\}$. The region $\mathfrak{N}$ is bounded by the curves

$$(t^2 - x^2)[1 + 1/(t^2 + x^2)] = \pm 1 \tag{50}$$

(see Fig. 2). The boundary is not a closed curve surrounding the origin as was to be expected, since the condition (24) is not satisfied.

4. FURTHER RESULTS, REMARKS

A. The Relativity of Coincidence

One of the interesting and unusual ideas which emerges from the nonlinear realizations (from the pseudotranslations in particular) deals with the coincidence notion. In the first example of Sec. 3C, the set $\mathfrak{N}$ is $x^2 + t^2 \geq 1$, to which is added the single point $(x = 0, t = 0)$, or

$$\mathfrak{N} = \{x, t; x = 0, t = 0 + (x^2 + t^2) \geq 1\}. \tag{51}$$

The origin is surrounded by the unit circle $x^2 + t^2 < 1$, which does not belong to $\mathfrak{N}$. However, by the theorem proven in Sec. 3, one can pick any point a in $\mathfrak{N}$ as origin. What happens to the points in the neighborhood of a and arbitrarily near a (in Euclidean or Minkowski sense) after the pseudotranslation has been carried out? To investigate this, consider, for f given by (44), the extremely simple case of two points $\underline{a}$ and $\underline{b}$ in $\mathfrak{N}$ located on the x axis. It can easily be checked that, if a pseudotranslation is carried out with $q^0 = 0$, $q' \neq 0$, points on the x axis are translated to points which lie on the real axis. Furthermore, the translation is everywhere defined. Thus the calculation to be presented will transform $\underline{a}$ to $\underline{a}'$ which will be the new origin $\underline{a}' = (0, 0)$. Written out in the case to be considered is the translation of the points $\underline{a} = (a^0, a') = (0, a)$ and $\underline{b} = (b^0, b') = (0, b)$. Since a and b lie in $\mathfrak{N}$, $a > 1, b > 1$. However, the Euclidean distance $d_E(\bar{b}, \bar{a}) = b - a$ can be arbitrarily small (or zero). The invariant distance is (17)

$$K_{\bar{b}\bar{a}} = -(\sqrt{b^2 - 1} - \sqrt{a^2 - 1})^2. \tag{52}$$

Evidently, as $\bar{b}$ approaches $\bar{a}$, $K_{\bar{b}\bar{a}}$ can become arbitrarily small; if $b \gg 1$ and $a \gg 1$

$$K_{\bar{b}\bar{a}} \to -[d_E(\bar{b}, \bar{a})]^2, \tag{53}$$

as would be expected.

If now a translation is made, $K_{\bar{b}\bar{a}}$ must remain invariant; however, the Euclidean distance $d_E(\bar{b}', \bar{a}') = \bar{b}' - \bar{a}'$ should always remain larger than 1, for $\bar{b}' \in \mathfrak{K}$, $\bar{a}'$ is the origin, and $\mathfrak{N}$ contains apart from the origin only points outside the unit circle. To see how this comes about, start by finding the pseudotranslation taking $\underline{a}$ to the origin. This, by (38b), is given by

$$q_0^{(1)} = -\sqrt{a^2 - 1}, \tag{54a}$$

$$q_0^{(0)} = -0. \tag{54b}$$

This gives, using (34b),

$$(T_0 b)' = b^{(1)} + q_0^{(1)} A(b) = b\left(1 - \sqrt{\frac{a^2 - 1}{b^2 - 1}}\right), \tag{55a}$$

$$(T_0 b)' = b(1 - \lambda), \tag{55b}$$

$$\lambda \equiv \sqrt{\frac{a^2 - 1}{b^2 - 1}} \tag{55c}$$

Since both a and b lie in $\mathfrak{N}$, λ is real.

The calculation of $R^2(b \mid T_0)$ according to (34c) yields

$$R^2(b \mid T_0) = 1 - b^{-2} + b^{-2}(1 - \lambda)^{-2}. \tag{56}$$

From (55) and (56) and (34a) the coordinates of $\underline{b}'$ can be found as

$$(b^{(1)})' = b(1 - \lambda)\sqrt{1 - b^{-2} + b^{-2}(1 - \lambda)^{-2}}, \tag{57a}$$

$$(b^{(0)})' = 0. \tag{57b}$$

Since $\underline{a}$ is mapped in $\underline{a}' = (0, 0)$, the Euclidean distance between $\underline{a}'$ and $\underline{b}'$ is just

$$d^E(\underline{b}', \underline{a}') = 1 + (b^2 - 1)(1 - \lambda)^2. \tag{58}$$

Since $b > 1$, $(b^2 - 1)(1 - \lambda)^2$ is always positive; hence $d^E(b', a') \geq 1$. The Euclidean distance in the transformed frame is therefore always larger than 1, as required, no matter how close b and a are. The invariant distance should, of course, be the same as a calculation of $K(b', a')$, using (57a), verifies. The situation can perhaps be made clearer by considering a sequence of points b_n on the real axis converging to a:

$$b_n = (1/2^n)b + [(2^n - 1)/2^n]a. \tag{59}$$

Clearly, $b_0 = b$, while $\lim_{n\to\infty} b_n = a$. Under the pseudotranslation described, each b_n maps into b_n', which has a Euclidean distance from a', the origin given by

$$(d^E(b_n', a')) = \sqrt{1 + (b_n^2 - 1)(1 - \lambda_n)^2}, \tag{60a}$$

$$\lambda_n = \sqrt{(a^2 - 1)/(b_n^2 - 1)}. \tag{60b}$$

In the limit as $n \to \infty$, we have $b_n \to a$, $\lambda_n \to 1$, so that

$$\lim_{n\to\infty} d^E(b_n', a') = 1. \tag{60c}$$

Hence, the limit of a sequence converging to a will in the new frame approach the unit circle around the origin. This demonstrates the result announced that events having an arbitrary small Euclidean or Minkovski separation (or coincident events) will appear with a finite Minkovski separation (or no longer coincident) in a transformed frame.

B. Remarks

The following remarks may help to round out the discussion and point to a number of questions.

(i) It was demonstrated in the first example of Sec. 3C that in the Ξ space the transformations Q assume their linear forms. Since there are simple connection formulas between the x and ξ variables, why should one not just work in the Ξ space and forget the X space? It was pointed out that the mapping from $X \to \Xi$ is *not* one to one if the unit circle in X is included. If the unit circle in X is not included, the origin in Ξ must be deleted. The crucial point is

really the physical significance which must be attributed to the x and ξ variables. If the x variables in X are observable and if two distinct points on the unit circle X correspond to distinct physical situations, the mapping U from $X \to \Xi$ obliterates this distinction and the physical difference in question cannot be expressed in the Ξ space. If, however, the x variables are not observable, or if no measurement can as a matter of principle reveal a physical difference between points on the unit circle, the mapping to Ξ, which takes the unit circle to the origin, does not destroy any physical distinctions. In fact, the transformation from $X \to \Xi$ then maps all physical indistinguishable points in X to a single point in Ξ which is highly appropriate.

Although phrased in terms of the first example, the remarks made are of more general validity. For, if one insists that the "physical points" in X shall be in $\mathfrak{A}$ and shall allow translations as well, (36) shows that these points are necessarily in the region $\mathcal{L}$, so that the transformations can be recast in their linear form. If further f is continuous and (24) is satisfied, the situation is exactly as just described. The equivalence or lack of it depends then on the observability of the variables and the different restrictions this imposes. Consider, for example, the equation giving the eigenvalues of the total momentum operator in X:

$$-g^{\mu\nu}P_\mu P_\nu u(x) = m^2 u(x). \tag{61}$$

This equation when transformed to the Ξ space is just

$$g^{\mu\nu}\frac{\partial}{\partial\xi^\mu}\frac{\partial}{\partial\xi^\nu}u(\xi) = -m^2 u(\xi) \tag{62}$$

(m is a constant).

The distinction in the X and Ξ description now can enter through the different boundary conditions which can be imposed on equations. If this is done—for physical reasons—the X and Ξ descriptions will not be mere transcriptions of each other.

Of course, the formalism developed is much broader than that. If one does not require that all points shall be capable of translation, or if f is not continuous, other differences not on an interpretive level appear between X and Ξ spaces.

(ii) It is interesting to note that if the x^μ are interpreted as position variables, and if the nonlinear realizations are appropriate for the physical description, this framework suggests two distinct notions of momenta. One is the *generator* of the translations given explicitly by (22); the other is the quantity conjugate to the variable x [the operator $(1/i)\partial/\partial x$]. The belief that high energy scattering probes small space-time regions rests on the identification of the *physical* momentum with the canonical momentum [$p = (1/i)\partial/\partial x$]. Only then does the usual relation between small distances and large momenta emerge via Fourier transforms. If, on the other hand, the physical momenta are more properly identified with the generators of the translations [P in (22)], the high momentum behavior would not be so directly correlated with the small distance behavior.

(iii) It was noted in remark (i) that the decision of what quantities are observable is crucial for the physical interpretation. As written, the impression might be created that the variables are necessarily positional variables. This, however, need not be the case at all; it would be perfectly possible to carry through the construction of the nonlinear realization in a space where the underlying variables are momentum variables. It would, therefore, appear premature to reject the nonlinear realizations on the grounds that they place undue emphasis on the position variables.

(iv) There is always some subtlety in deciding just what features of an invariance group should have physical significance. Stated differently, is the physics contained in the invariants, the abstract group, the representations, the realizations? The experience with internal symmetry groups indicates that the dimensions of the irreducible representations have a direct physical meaning. The eigenvalues of the Casimir operators of the Poincaré group determine masses and spins. It would be most intriguing if any meaningful physical characteristics could be associated with nonlinear realizations.

---

* Supported in part by A.E.C. Contract AT(30-1) 3668B.

1 M. Dresden and A. Albano, Proc. Natl. Acad. Sci. (U.S.) **58**, 916 (1967).
2 M. Dresden, in *Fundamental Interactions at High Energy*, Coral Gables, 1970 (Gordon and Breach, New York, 1970).
3 M. B. Halpern and D. Atkinson, J. Math. Phys. **8**, 373 (1967).
4 T. T. Wu and C. N. Yang, Phys. Rev. **137**, B708 (1965).
5 J. Benecke, T. T. Chou, C. N. Yang, and E. Yen, Phys. Rev. **188**, 2159 (1969).
6 M. Dresden, in *Noncompact Groups in Particle Physics*, Conference Proceedings, edited by Y. Chow (Benjamin, New York, 1966).
7 The argument leading to (1) is given in Ref. 6. It is generous to call it a heuristic argument; the form was suggested by a number of special examples. For the discussion of this paper, just the fact that f depends on the ratios is important.
8 The statements in Ref. 2 concerning "finite and infinite" excluded regions are saved from being wrong only by being ambiguous. Section 3 contains the definitive results.
9 "Representation" will be used to denote a linear representation; a nonlinear representation will be referred to as a realization.
10 S. Weinberg, Phys. Rev. **166**, 1568 (1968).
11 Julian Schwinger, Phys. Letters **24B**, 473 (1967).
12 K. Meetz, J. Math Phys. **10**, 589 (1969).
13 Primes are used to indicate points transformed according to (3).
14 The transformation U is not a mapping of X into (or onto) itself as is Q, but rather the introduction of a new set of variables on X.
15 In this derivation $x'^\mu = S'^\mu_\nu x^\nu R(x|S)$ is just written as $x' = (Sx)R(x|S)$. The "squares" x^2 in (17) are the usual $g_{\mu\nu}x^\mu x^\nu = x^2$, etc.
16 In particular, the pseudomomenta P_μ commute. In Ref. 1, where the operators were obtained via a limiting process in a de Sitter space, it was erroneously stated that they did not commute. A mistake was made in the limiting process.
17 a is used here to denote a point of coordinates a^μ. This should not be confused with the parameter of the translations previously called a, which now will be called q.
18 It is assumed that a does not lie on the light cone $I(a) \neq 0$.
19 In these two-dimensional examples, x^0, x' are written as t, x; care should be exercised, for x^2 could be $x^2 = (x^0)^2 - (x^{(1)})^2 = t^2 - x^2$, or $(x^{(1)})^2$. The context should define the meaning, so that no confusion should arise.

Hamiltonian Formalism for General Lagrangian Systems in an Exceptional Case

C. Ryan
Theoretical Physics Institute, University of Alberta, Edmonton, Alberta
and
*Mathematical Physics Department, University College, Dublin, Ireland**
(Received 9 August 1971)

It is shown that in the exceptional case of a Lagrangian system, where the highest derivative of a generalized coordinate cannot be expressed in terms of the lower derivatives and the generalized momenta, the procedure to adopt in setting up the Hamiltonian formalism for the system (and consequently the quantization of the system) is to pass to an equivalent Lagrangian which does not have this defect and to use this rather than the original Lagrangian.

1. INTRODUCTION

In recent years intermittent contributions have appeared in the literature concerning the treatment of systems describable in terms of Lagrangians, which depend on time derivatives of the generalized coordinate(s) higher than the first.[1] As has been pointed out in Ref. 2, some of these contributions—those dealing with the setting up of the generalized Hamilton–Jacobi theory for such systems—suffer from the drawback that the authors apparently were not aware of the classic work of Ostrogradsky on this subject.[3] This work shows that in all but one case it is possible to establish a unique consistent Hamiltonian formalism and, consequently, following the usual prescription, a unique quantum theory of such systems.

Clearly the case when the Ostrogradsky approach appears to fail is of some interest, and not long ago a particular instance of this case was focussed on by Hayes.[4] The exceptional case in question is that which occurs when in setting up the Hamiltonian formalism one is unable to eliminate the highest derivative(s) of the generalized coordinate(s) in terms of lower derivatives and momenta. Hayes' solution to this difficulty is to eliminate the highest derivative using the equations of motion, but this is a highly questionable procedure since if it is applied equally to all terms, the Hamiltonian reduces to a quantity which is altogether independent of generalized coordinates and momenta. Actually what Hayes' prescription amounts to in practice is the replacement of a Hamiltonian of the form

$$H(q, p) = (p^2/2m) + V(q) \tag{1.1}$$

by a Hamiltonian

$$H(q, q_1; p, p_1) = (p^2/m) + (p_1^2/m) + \tfrac{1}{2}V(q) + \tfrac{1}{2}V(q_1), \tag{1.2}$$

where the q_1 and p_1 are a new pair of canonically conjugate variables independent of q and p. Clearly, (1.2) is a Hamiltonian which yields the same equations of motion for q as does (1.1); but it is only one of an infinity of such and, besides, added degrees of freedom have been introduced which lead to degeneracy when the system is quantized.

Initially, this is a disquieting prospect since it reduces the traditional quantum mechanical results for systems such as the harmonic oscillator and the hydrogen atom to merely one of an infinity of possible results. However, on closer inspection, one learns that the difficulty of the exceptional Ostrogradsky case, on which all this ambiguity rests, is not a real difficulty at all. We shall show that the correct procedure in such cases is to change from the original Lagrangian to an equivalent one which does not give rise to the exceptional case and then to use this Lagrangian for setting up the Hamiltonian formalism and the consequent quantization procedure.

In Sec. 2 we summarize briefly the Ostrogradsky method and Hayes' instance of an exceptional case where the method breaks down. In Sec. 3 we show that in all instances of the exceptional case, one can define an alternative equivalent Lagrangian which does not give rise to this case and that using this Lagrangian the Hamiltonian formalism and the quantization procedure is unique.

2. THE OSTROGRADSKY METHOD AND AN EXCEPTIONAL CASE

For simplicity we consider first a system described by one generalized coordinate x since the generalization of what follows to systems described by arbitrary numbers of generalized coordinates is quite straightforward, as we show briefly in the end. Given then a system described by a Lagrangian

$$L = L(t, x, Dx, D^2x, \ldots, D^nx), \tag{2.1}$$

where $D = d/dt$, we obtain the Euler–Lagrange equation of motion

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial (Dx)} + \frac{d^2}{dt^2}\frac{\partial L}{\partial (D^2x)} - \cdots + (-)^n \frac{d^n}{dt^n}\frac{\partial L}{\partial (D^nx)} = 0. \tag{2.2}$$

In order to cast this equation in Hamiltonian form, we write

$$q_i = D^{i-1}x, \qquad p_i = \delta L/\delta (D^i x), \qquad i = 1, 2, \ldots, n, \tag{2.3}$$

where

$$\frac{\delta L}{\delta y} \equiv \frac{\delta L}{\delta y} - \frac{d}{dt}\frac{\partial L}{\partial (Dy)} + \frac{d^2}{dt^2}\frac{\partial L}{\partial (D^2y)} - \cdots + (-)^n \frac{d^n}{dt^n}\frac{\partial L}{\partial (D^ny)}, \tag{2.4}$$

and define

$$H = p_1q_2 + p_2q_3 + \cdots + p_{n-1}q_n + p_nD^nx - L, \tag{2.5}$$

where H is supposed to be a function of t, q_i, and p_i, the quantity D^nx being eliminated by use of the equation $p_n = \partial L/\partial (D^nq)$. It then follows[3] that

$$\dot{q}_r = \frac{\partial H}{\partial p_r}, \qquad \dot{p}_r = -\frac{\partial H}{\partial q_r} \tag{2.6}$$

and so the equation of motion (2.2) has been expressed in Hamiltonian form. Quantization of the system can now be carried out in the usual way.

The exception to the above discussion occurs when $D^n x$ cannot be eliminated using the equation $p_n = \partial L/\partial(D^n x)$. We illustrate this situation by reference to the example considered by Hayes.[4] Here the Lagrangian is given by

$$L = -\tfrac{1}{2}mxD^2x - V(x), \tag{2.7}$$

from which we obtain the equation of motion

$$m\ddot{x} + \frac{\partial V}{\partial x} = 0. \tag{2.8}$$

The quantities q_i and p_i defined in Eq. (2.3) are in this case

$$q_1 = x, \quad q_2 = Dx, \quad p_1 = \tfrac{1}{2}mDx, \quad p_2 = -\tfrac{1}{2}mx, \tag{2.9}$$

the Hamiltonian is

$$H = p_1q_2 + p_2D^2x + \tfrac{1}{2}mq_1D^2x + V(q_1) \tag{2.10}$$

and, as we see, we are unable to eliminate D^2x in terms of the q_i and p_i. It might be thought that since due to (2.9), the terms in H involving D^2x cancel out, one can simply omit these terms from H. However, if one does this in a straightforward way, one easily verifies that the canonical equations do not agree with (2.8). What Hayes suggests is to drop the terms in D^2x and rewrite the remaining ones using the third and fourth of Eqs. (2.9) so as to yield

$$H = (p_1^2/m) + \tfrac{1}{4}mq_2^2 + \tfrac{1}{2}V(q_1) + \tfrac{1}{2}V(-(2/m)p_2). \tag{2.11}$$

From this Hamiltonian one obtains canonical equations equivalent to (2.8). If we now make the canonical transformation $-(2/m)p_2 = Q$, $\tfrac{1}{2}mq_2 = P$, we see that Eq. (2.11) is equivalent to the Hamiltonian form given in Eq. (1.2). The shortcomings of such a Hamiltonian have been mentioned already.

3. SOLUTION OF THE DIFFICULTY

In order to circumvent the above type of difficulty, let us enquire into what exactly happens when one is unable to eliminate $D^n x$ using the equation $p_n = \partial L/\partial(D^n x)$. Clearly in such cases the Lagrangian must be of the form

$$L(t, x, Dx, \ldots, D^{n-1}x, D^n x) = D^n x F(t, x, Dx, \ldots, D^{n-1}x) + G(t, x, Dx, \ldots, D^{n-1}x). \tag{3.1}$$

Now when this happens, the Euler–Lagrange equation (2.2) is of order $2n - 2$ at most.

Proof: Only the last two terms on the left-hand side of (2.2) could give rise to terms of order higher than $2n - 2$. Now for these two terms we obtain, using Eq. (3.1),

$$(-)^{n-1}\frac{d^{n-1}}{dt^{n-1}}\frac{\partial L}{\partial(D^{n-1}x)} + (-)^n\frac{d^n}{dt^n}\frac{\partial L}{\partial(D^n x)}$$
$$= (-)^{n-1}\frac{d^{n-1}}{dt^{n-1}}\left(\frac{\partial L}{\partial(D^{n-1}x)} - \frac{dF}{dt}\right)$$
$$= (-1)^{n-1}\frac{d^{n-1}}{dt^{n-1}}\left(D^n x\frac{\partial F}{\partial(D^{n-1}x)} + \frac{\partial G}{\partial(D^{n-1}x)} - \frac{\partial F}{\partial t} - \frac{\partial F}{\partial x}Dx - \cdots - \frac{\partial F}{\partial(D^{n-1}x)}D^n x\right)$$
$$= (-1)^{n-1}\frac{d^{n-1}}{dt^{n-1}}\left(\frac{\partial G}{\partial(D^{n-1}x)} - \frac{\partial F}{\partial t} - \frac{\partial F}{\partial x}Dx - \cdots - \frac{\partial F}{\partial(D^{n-2}x)}D^{n-1}x\right)$$

and this is clearly of order $2n - 2$ at most.

This result raises the suspicion that the system under consideration may actually be describable in terms of a Lagrangian involving derivatives of x up to at most the $(n-1)$th. This is indeed the case.

Proof: Consider the quantity J defined by

$$J(t, x, Dx, \ldots, D^{n-1}x) = \int^{D^{n-1}x} d\alpha F(t, x, Dx, \ldots, D^{n-2}x, \alpha). \tag{3.2}$$

Obviously we have

$$\frac{dJ}{dt} = F(t, x, Dx, \ldots, D^{n-2}x, D^{n-1}x)D^n x + \int^{D^{n-1}x} d\alpha\left(\frac{\partial F(t, x, Dx, \ldots, D^{n-2}x, \alpha)}{\partial t} + \sum_{i=0}^{n-2}\frac{\partial F(t, x, Dx, \ldots, D^{n-2}x, \alpha)}{\partial(D^i x)}\right). \tag{3.3}$$

From this it follows that if we replace the Lagrangian L in Eq. (3.1) by the Lagrangian L' given by

$$L' = -\int^{D^{n-1}x}\left(d\alpha\,\frac{\partial F(t, x, Dx, \ldots, D^{n-2}x, \alpha)}{\partial t} + \sum_{i=0}^{n-2}\frac{\partial F(t, x, Dx, \ldots, D^{n-2}x, \alpha)}{\partial(D^i x)}\right) + G, \tag{3.4}$$

we have obtained a new Lagrangian L', which yields the same equation of motion as L but which contains derivatives of x up to at most the $(n-1)$th. Note that if the foregoing replacement procedure is applied to the Lagrangian in Eq. (2.7), we get the usual form

$$L' = \tfrac{1}{2}m\dot{x}^2 - V(x),$$

which shows the uniqueness of this form within the context of the present discussion.

The lesson of this result is that if L' does not give rise to the exceptional case, it should be used in place of L in setting up the Hamiltonian formalism. If L' also gives rise to the exceptional case, we can reduce still further to a Lagrangian L'' with the $(n-2)$th as the highest derivative and so on until we arrive at a nonexceptional case. There is of course the possibility that all the equivalent Lagrangians in this sequence continue to be exceptional until you have actually arrived at a Lagrangian depending on x alone without any of its derivatives; this is a singular case in which the equation of motion is simply an algebraic equation and clearly no Hamiltonian formulation of such an equation can exist. Hence, barring this singular case, we now see how to construct the Hamiltonian and hence the quantum mechanical formalism for all systems described by a Lagrangian of the form (2.1).

Basically what goes wrong with the Hamiltonian formalism for a Lagrangian which gives rise to the

exceptional case, is a redundancy of generalized coordinates and momenta. The Hamiltonian method outlined in Sec. 2 says equivalently that the value of x at any time t_2 can be derived from a knowledge of $2n$ independent pieces of information about it at any other time t_1, the $2n$ pieces of information being essentially the values of x and of its first $2n-1$ derivatives. Now in the exceptional case also, the Hamiltonian method naturally leads us to introduce the $2n$ quantities q_i and p_i $(i = 1, 2, \ldots, n)$, but we do not need so many since now the Euler–Lagrange equation is only of order $2n-2$. It is this surfeit of generalized coordinates and momenta which causes the trouble and it is equally the reduction of the problem to its proper size (by changing from L to L' as above) which avoids it.

In order to indicate how things go when we have more than one generalized coordinate, we briefly consider the case of a Lagrangian depending on two-generalized coordinates x and y and their first m and n time derivatives, respectively. The exceptional case we have been discussing now corresponds to a Lagrangian of the form

$$L = F(t, x, Dx, \ldots, D^{m-1}x; y, Dy, \ldots, D^n y)D^m x + G(t, x, \ldots, D^{m-1}x; y, Dy, \ldots, D^n y). \quad (3.5)$$

Clearly we can reduce this to an equivalent Lagrangian involving derivatives of x up to at most the $(m-1)$th by adding to it the total derivative dK/dt with K given by

$$K = -\int^{D^{m-1}x} d\alpha F(t, x, Dx, \ldots, D^{m-2}x, \alpha; y, Dy, \ldots, D^n y). \quad (3.6)$$

If this new Lagrangian is also exceptional with respect to $D^{m-1}x$, we reduce it still further and so on until we arrive at an unexceptional Lagrangian as far as x goes. We then do the same with y. The singular case in which an unexceptional Lagrangian is obtained only when all the derivatives of x or of y have been eliminated can occur here too. In general, this case does not have a Hamiltonian formulation though in particular instances it does. Consider, for example, the Lagrangian

$$L = \tfrac{1}{2}m\omega(x\dot{y} - y\dot{x}) + \tfrac{1}{2}m\omega^2(x^2 + y^2). \quad (3.7)$$

The Euler–Lagrange equations are

$$\begin{aligned} -\dot{y} + \omega x &= 0, \\ \dot{x} + \omega y &= 0. \end{aligned} \quad (3.8)$$

Turning now to the Hamiltonian formalism we see that L is exceptional with respect to $\dot{x}$. We therefore pass to the equivalent Lagrangian L',

$$L' = m\omega x\dot{y} + \tfrac{1}{2}m\omega^2(x^2 + y^2). \quad (3.9)$$

This is what we have termed the singular case as far as x goes. The problem is that p_x, the momentum conjugate to x, is zero and hence x is not a canonical variable; it must then be eliminated. It so happens that this can be done because

$$p_y = \frac{\partial L'}{\partial \dot{y}} = m\omega x \quad (3.10)$$

and hence when we set up the Hamiltonian according to the usual rule

$$H = p_y\dot{y} - L'. \quad (3.11)$$

Using (3.9) and (3.10), we find

$$H = (1/2m)p_y^2 + \tfrac{1}{2}m\omega^2 y^2. \quad (3.12)$$

The canonical equations derived from this equation are equivalent to the Euler–Lagrange equations (3.8). We stress, however, that this case is not typical and that generally what we have called singular cases have no Hamiltonian formulation. Notice, by the way, that this example is analogue in particle mechanics of the Lagrangian and Hamiltonian formalism for a Dirac field.[5]

In summary then, we have shown that the exceptional case which arises when the highest derivative in a Lagrangian cannot be expressed as a function of the lower derivatives and the generalized momenta can be handled in general by replacing the original Lagrangian by an equivalent one which does not have this difficulty. The use of the equivalent Lagrangian then allows the setting up in a unique way of the Hamiltonian formalism and hence of the quantization of the system.

ACKNOWLEDGMENTS

I am indebted to Mr. John Browne for bringing this problem to my attention and to Professor Y. Takahashi for a very helpful discussion. The hospitality of Professor Takahashi and of the other members of the Theoretical Physics Institute at the University of Alberta is gratefully acknowledged.

---

* Permanent address.

1 M. Borneas, Am. J. Phys. **27**, 265 (1959); Nuovo Cimento **16**, 806 (1960); J. Koestler and J. Smith, Am. J. Phys. **33**, 140 (1965).
2 J. Kruger and D. Callebaut, Am. J. Phys. **36**, 557 (1968).
3 M. Ostrogradsky, Mém. Acad. St.-Pét. **6**, 385 (1850). For a modern presentation see E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies* (Cambridge, U.P., Cambridge, 1965), p. 266.
4 C. F. Hayes, J. Math. Phys. **10**, 1555 (1969).
5 See, for example, G. Wentzel, *Quantum Theory of Fields* (Interscience, New York, 1949), pp. 170-71, or any book on quantum field theory.

Colliding Plane Gravitational Waves

P. Szekeres
Department of Mathematical Physics, University of Adelaide, Adelaide, South Australia
(Received 12 July 1971)

The equations governing the collision of two plane gravitational waves are derived. The general exact solution representing this situation when both waves are linearly polarized are found, and some special solutions of possible physical interest are discussed in detail.

1. INTRODUCTION

A fundamental problem in gravitation theory is the collisional interaction of gravitational waves. The problem is interesting in that the nonlinear features of general relativity should show up explicitly in the failure of the superposition principle and the precise nature of the diffusion of independent gravitational fields through each other should be susceptible to investigation. Apart from the possible guide to the quantization program which such an analysis could provide (this is, after all, the unquantized equivalent of the graviton–graviton interaction) it does appear particularly appropriate to investigate this problem at this time in view of the fact that gravitational waves are now on the brink of observational experience.[1] The interpretation of Weber's reported fluxes has rested almost entirely on the linearized approximation, and there are manifest difficulties in reconciling them with physically reasonable assumptions about the mechanism for generating them.[2] It is currently most popular to regard these waves as arising from the gravitational collapse of objects at the center of the galaxy, and there is some observational evidence for this.[3] If the objects are distributed randomly throughout the central region of the galaxy, and are "popping off" at the rate of about one a day, a typical wave will experience about 10^5 collisions with other waves before emerging from the central region (assumed to be about 10^3 light years across). It is by no means evident that the cumulative effect of such a large number of nonlinear interactions will leave the emergent wave unscathed as regards the linearized approximation. Furthermore, if our galaxy is at all typical, other galaxies should also be strong sources of gravitational waves, giving rise to very large numbers of collisions in the intergalactic space. An understanding of the space–time curvature resulting from the collision of gravitational waves is therefore of considerable importance to cosmology.

The present study is restricted to colliding plane waves, which represent to some approximation the fields far from radiating sources. The results are in no instance to be regarded as holding without severe modifications for realistic waves having curved wavefronts. However, the nonlinearity is taken fully into account and the solutions discussed should certainly act as a guide to a corresponding discussion for more physically realistic situations.

Exact solutions representing a collision of two plane waves have recently been given by the author[4] and independently by Kahn and Penrose[5], the latter having solved the problem with two impulsive waves. In this paper, I will give a more detailed derivation of these metrics. The field equations representing two arbitrary plane waves in collision will be established in Secs. 2, 3, and 4, and will be solved first approximately by power series in Sec. 5, then exactly for the case of linearly polarized waves in Sec. 6. Finally in Sec. 7 there is a discussion of a physically more interesting situation where the incoming waves each consist of a pair of impulse waves in close succession. Such waves have a finite energy flux and may be compared with the bursts of radiation observed by Weber.

2. GRAVITATIONAL PLANE WAVES

In harmonic coordinates the metric of a *plane wave* is[6]

$$ds^2 = -Udu^2 + 2dudv - (dx^2 + dy^2), \tag{2.1}$$

where

$$U = f(u)(x^2 - y^2) + 2g(u)xy. \tag{2.2}$$

The plane wave is said to have *constant polarization* if $g(u)$ vanishes. By performing a suitable coordinate transformation, the metric may be put in the Rosen form[7]

$$ds^2 = 2e^{-M}dudv - g_{ij}dx^i dx^j, \tag{2.3}$$

where

$$M = M(u), \quad g_{ij} = g_{ij}(u), \quad i = 2, 3.$$

The condition for constant polarization is equivalent to the condition that g_{ij} be diagonalizable by a linear transformation on the x^i.

For the problem of two colliding waves the Rosen coordinate system has the following immediate advantages over the harmonic coordinates:

(i) Both u and v are null coordinates, so that both waves may be simultaneously represented in the same coordinate patch.

(ii) In harmonic coordinates the Riemann tensor components are functions of $f(u)$ and $g(u)$ not involving their derivatives. Hence sandwich waves involving discontinuities of the curvature tensor cannot be represented in these coordinates if one insists on the generally accepted Lichnerowicz conditions[8] (continuity in the metric and its first derivatives). In Rosen's coordinates such discontinuities may be represented since the Riemann tensor involves second derivatives of g_{ij}.

(iii) Plane waves have in general a five-parameter group of symmetries. There is a two-parameter Abelian subgroup of symmetries acting like planar translations in the spacelike 2-surfaces u = const, v = const. These symmetries, which will be seen to be preserved by the interaction, are evident at once from the Rosen form, but not from the harmonic coordinates, by the independence of $g_{\mu\nu}$ on x^i in (2.3).

It is worth noting, however, that in general it is not possible to cover a plane sandwich wave with a single coordinate patch of the form (2.3), as coordinate singularities will invariably appear.[9] This property is of considerable geometrical interest as it is related to a curious and important focussing property of plane waves[10] which has a significant bearing on the development of singularities in the collision problem.

In order to discuss the collision problem, consider the

space–time pictured in Fig. 1. u and v are null coordinates

$$u_{,\mu}v_{,\nu}g^{\mu\nu} = v_{,\mu}v_{,\nu}g^{\mu\nu} = 0,$$

and x^1 and x^2 are spacelike coordinates. In region I ($v < 0$, $u < 0$) the space–time is flat Minkowski space.

$$ds^2 = 2dudv - 2\delta_{ij}dx^i dx^j.$$

In region II ($v < 0$, $u > 0$) it is a plane wave with metric of the form (2.3), while in region III ($v > 0$, $u < 0$) it is again a plane wave of form (2.3) but with $M = M(v)$ and $g_{ij} = g_{ij}(v)$. The junction across the I–II and I–III boundaries must be smooth in the sense of Lichnerowicz. The metric in the interaction zone IV is determined by a characteristic initial value problem with data specified on the pair of null hypersurfaces $u = 0$, $v = 0$ intersecting in a spacelike 2–surface. According to a theorem of Penrose,[11] if this data is well set, the Einstein field equations will uniquely determine the geometry in IV.

3. THE COORDINATE SYSTEM

Figure 1 represents a "head-on" collision of two plane waves when viewed by any geodesic timelike observer with world line $x^2 = \text{const}, x^3 = \text{const}, u = kv$, $k > 0$. Other timelike observers will see the two wavefronts approaching each other not head-on but at an angle; hence this more general situation is also taken care of by the present considerations. From an intuitive point of view it is clear that since the incoming waves have no dependence on the coordinates x^i, the metric in the interaction zone IV should also show no explicit dependence on these coordinates. The validity of such an assumption will follow from its ultimate success in determining the metric in the interaction zone and Penrose's theorem.

We assume then that throughout the space–time there exist a pair of commuting spacelike Killing vectors ξ_1^μ, ξ_2^μ,

$$[\xi_1, \xi_2] = 0.$$

Under these conditions it is a straightforward matter to show that there exists a coordinate system x^μ, $\mu = 0, 1, 2, 3$, such that

$$\xi_1^\mu = \delta_2^\mu, \quad \xi_2 = \delta_3^\mu$$

and $g_{\mu\nu} = g_{\mu\nu}(x^0, x^1)$. Furthermore, at each point of the manifold there exist just two null directions orthogonal to the 2-space spanned by ξ_1 and ξ_2. Let l^μ and n^μ be null vectors in these two directions chosen such that $l_\mu n^\mu = 1$. l_μ and n_μ are determined up to a scaling factor, and they are clearly both of the form $(A, B, 0, 0)$ in our coordinates. Since $g_{\mu\nu} = g_{\mu\nu}(x^0, x^1)$, it is possible to choose the scaling factor such that $A = A(x^0, x^1)$ and $B = B(x^0, x^1)$, whence it follows that there exist integrating factors $\psi(x^0, x^1)$, $\varphi(x^0, x^1)$ such that

$$\psi l_\mu = u_{,\mu}, \qquad \varphi n_\mu = v_{,\mu}, \tag{3.1}$$

where

$$\varphi\psi = u_{,\mu}v_{,\nu}g^{\mu\nu} \tag{3.2}$$

Hence u and v may be taken as the remaining coordinates

$$x^0 = u, \qquad x^1 = v.$$

Let m^μ be a complex null vector spanned by ξ_1^μ and ξ_2^μ, satisfying

$$m^\mu m_\mu = 0, \quad m^\mu \bar{m}_\mu = 1.$$

l^μ, n^μ, m^μ, and $\bar{m}^\mu$ constitute a null tetrad,

$$g^{\mu\nu} = l^\mu n^\nu + n^\mu l^\nu - m^\mu \bar{m}^\nu - m^\nu \bar{m}^\mu, \tag{3.3}$$

whose components in these coordinates may be set to be

$$\begin{aligned} l_\mu &= (\psi^{-1}, 0, 0, 0), \quad & l^\mu &= (0, \varphi, Y^2, Y^3), \\ n_\mu &= (0, \varphi^{-1}, 0, 0), \quad & n^\mu &= (\psi, 0, X^2, X^3), \\ m^\mu &= (0, 0, \xi^2, \xi^3), \end{aligned} \tag{3.4}$$

where

$$X^i = X^i(u, v), \qquad Y^i = Y^i(u, v), \qquad \xi^i = \xi^i(u, v).$$

The following tetrad and coordinate freedoms remain:

(1) Scale transformations:

$$l^{\mu'} = Al^\mu, \qquad n^{\mu'} = A^{-1}n, \qquad A = A(u, v). \tag{3.5}$$

The scale functions ψ and φ transform as follows under such a transformation

$$\psi' = A^{-1}\psi, \qquad \varphi' = A\varphi.$$

(2) Spatial rotations:

$$m^{\mu'} = e^{iC}m^\mu, \qquad C = C(u, v). \tag{3.6}$$

The ξ^i transform as follows

$$\xi^{i'} = e^{iC}\xi^i.$$

(3) Relabeling of null hypersurfaces:

$$u' = f(u), \qquad v' = g(v). \tag{3.7}$$

This induces a transformation of ψ and φ

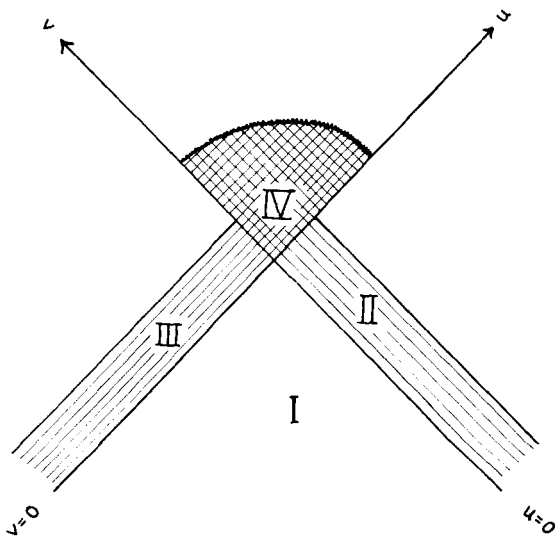


FIG. 1. Colliding plane waves. $u = \text{const}$, $v = \text{const}$ are null hypersurfaces. In region I the metric is Minkowski space. Regions II and III represent incoming plane waves which interact in region IV. A singularity eventually develops along the boundary represented by the jagged line.

$$\psi' = \psi \frac{df}{du}, \quad \varphi', = \varphi \frac{dg}{dv}.$$

(4) Spatial coordinate transformations:

$$u' = u, \quad v' = v, \quad x^{i\prime} = x^i + f^i(u, v). \tag{3.8}$$

These induce transformations

$$Y^{i\prime} = Y^i + \varphi \frac{\partial f^i}{\partial v}, \quad X^{i\prime} = X^i + \psi \frac{\partial f^i}{\partial u}.$$

(5) Linear coordinate transformations:

$$x^{i\prime} = a^i{}_j x^j, \quad a^i{}_j = \text{consts}. \tag{3.9}$$

4. FIELD EQUATIONS

It will be convenient to express the field equations in terms of the Newman–Penrose spin coefficients,[12] defined as

$$\begin{aligned}
\kappa &= l_{\mu;\nu} m^\mu l^\nu, & \pi &= -n_{\mu;\nu}\bar{m}^\mu l^\nu,\\
\rho &= l_{\mu;\nu} m^\mu \bar{m}^\nu, & \lambda &= -n_{\mu;\nu}\bar{m}^\mu \bar{m}^\nu,\\
\sigma &= l_{\mu;\nu} m^\mu m^\nu, & \mu &= -n_{\mu;\nu}\bar{m}^\mu m^\nu,\\
\tau &= l_{\mu;\nu} m^\mu n^\nu, & \nu &= -n_{\mu;\nu}\bar{m}^\mu n^\nu\\
\epsilon &= \tfrac{1}{2}(l_{\mu;\nu} n^\mu l^\nu - m_{\mu;\nu}\bar{m}^\mu l^\nu),\\
\alpha &= \tfrac{1}{2}(l_{\mu;\nu} n^\mu \bar{m}^\nu - m_{\mu;\nu}\bar{m}^\mu \bar{m}^\nu),\\
\beta &= \tfrac{1}{2}(l_{\mu;\nu} n^\mu m^\nu - m_{\mu;\nu}\bar{m}^\mu m^\nu),\\
\gamma &= \tfrac{1}{2}(l_{\mu;\nu} n^\mu n^\nu - m_{\mu;\nu}\bar{m}^\mu n^\nu),
\end{aligned}$$

and Weyl tensor components

$$\begin{aligned}
\Psi_0 &= -C_{\alpha\beta\gamma\delta} l^\alpha m^\beta l^\gamma m^\delta,\\
\Psi_1 &= -C_{\alpha\beta\gamma\delta} l^\alpha n^\beta l^\gamma m^\delta,\\
\Psi_2 &= -\tfrac{1}{2}C_{\alpha\beta\gamma\delta}(l^\alpha n^\beta l^\gamma n^\delta + l^\alpha n^\beta m^\gamma \bar{m}^\delta),\\
\Psi_3 &= -C_{\alpha\beta\gamma\delta} n^\alpha l^\beta n^\gamma \bar{m}^\delta,\\
\Psi_4 &= -C_{\alpha\beta\gamma\delta} n^\alpha \bar{m}^\beta n^\gamma \bar{m}^\delta.
\end{aligned}$$

In our case it is clear that all spin coefficients are functions of u and v alone, and hence the differential operators $D \equiv l^\mu \partial/\partial x^\mu$, $\Delta \equiv n^\mu \partial/\partial x^\mu$, $\delta \equiv m^\mu \partial/\partial x^\mu$ reduce simply to

$$D \equiv \varphi \frac{\partial}{\partial v}, \quad \Delta \equiv \psi \frac{\partial}{\partial u}, \quad \delta \equiv 0$$

when applied to spin coefficients.

The commutation relations [NP(4.4)] give at once

$$\kappa = \nu = 0, \quad \rho = \bar{\rho}, \quad \mu = \bar{\mu}, \quad \beta = \bar{\alpha}, \quad \bar{\tau} = \pi = 2\alpha$$

and

$$D\psi = -(\epsilon + \bar{\epsilon})\psi, \tag{4.1}$$

$$\Delta\varphi = (\gamma + \bar{\gamma})\varphi, \tag{4.2}$$

$$\Delta Y^i - DX^i = (\gamma + \bar{\gamma})Y^i + (\epsilon + \bar{\epsilon})X^i - 4\bar{\alpha}\bar{\xi}^i - 4\alpha\xi^i, \tag{4.3}$$

$$D\xi^i = \sigma\bar{\xi}^i + (\rho + \epsilon - \bar{\epsilon})\xi^i, \tag{4.4}$$

$$\Delta\xi^i = -\bar{\lambda}\bar{\xi}^i - (\mu + \bar{\gamma} - \gamma)\xi^i. \tag{4.5}$$

Before writing down the field equations, it is convenient to introduce modified "scale-invariant" spin coefficients:

$$\begin{aligned}
&\rho^0 = \rho\varphi^{-1}, \quad \mu^0 = \mu\psi^{-1},\\
&\sigma^0 = \sigma\varphi^{-1}, \quad \lambda^0 = \lambda\psi^{-1}\\
&E^0 = i(\bar{\epsilon} - \epsilon)\varphi^{-1}, \quad G^0 = i(\bar{\gamma} - \gamma)\psi^{-1},\\
&Y^{0i} = Y^i\varphi^{-1}, \quad X^{0i} = X^i\psi^{-1}\\
&\Psi^0 = \Psi_0\varphi^{-2}, \quad \Psi_1^0 = \Psi_1\varphi^{-1}, \quad \Psi_2^0 = \Psi_2\varphi^{-1}\psi^{-1},\\
&\Psi_3^0 = \Psi_3\psi^{-1}, \quad \Psi_4^0 = \Psi_4\psi^{-2}.
\end{aligned}$$

All these quantities as well as α, ξ^i, and the product $\varphi\psi$ are all invariant under scale transformations (3.5). Equations (4.1) and (4.2) may be written

$$\begin{aligned}
\epsilon + \bar{\epsilon} &= -\varphi(\log\psi)_{,v},\\
\gamma + \bar{\gamma} &= \psi(\log\varphi)_{,u}.
\end{aligned}$$

Thus the real parts of ϵ and γ cannot be made scale-invariant. On the other hand, these real parts never make an explicit appearance in the remaining commutation relations and field equations when written in terms of scale-invariant spin coefficients; hence these two equations are of no consequence and may be regarded as merely expressing $\epsilon + \bar{\epsilon}$ and $\gamma + \bar{\gamma}$ in terms of φ and ψ.

Equation (4.3) can be rewritten in terms of scale-invariant expressions

$$Y^{0i}{}_{,u} - X^{0i}{}_{,v} = -4e^{-M}(\bar{\alpha}\bar{\xi}^i + \alpha\xi^i), \tag{4.6}$$

where

$$M = \log(\varphi\psi). \tag{4.7}$$

This equation tells us that $\alpha = 0$ is a necessary and sufficient condition for there to exist a spatial coordinate transformation (3.8) which makes X^i and Y^i simultaneously zero. Finally, the remaining commutation relations (4.4), (4.5) and the field equations NP(4.2a)–(4.2r) reduce to the following set of equations:

$$\xi^i{}_{,v} = \sigma^0\bar{\xi}^i + (\rho^0 + iE^0)\xi^i, \tag{4.8}$$

$$\xi^i{}_{,u} = -\bar{\lambda}^0\bar{\xi}^i - (\mu^0 - iG^0)\xi^i, \tag{4.9}$$

$$\rho^0{}_{,v} = (\rho^0)^2 - \rho^0 M_{,v} + \sigma^0\bar{\sigma}^0, \tag{4.10}$$

$$\rho^0{}_{,u} = -2\mu^0\rho^0 - 4e^{-M}\alpha\bar{\alpha} \tag{4.11}$$

$$\mu^0{}_{,v} = 2\rho^0\mu^0 + 4e^{-M}\alpha\bar{\alpha}, \tag{4.12}$$

$$\mu^0{}_{,u} = -(\mu^0)^2 - \mu^0 M_{,u} - \lambda^0\bar{\lambda}^0, \tag{4.13}$$

$$\sigma^0{}_{,v} = \sigma^0(2\rho^0 - M_{,v} + 2iE^0) + \Psi_0^0, \tag{4.14}$$

$$\sigma^0{}_{,u} = (2iG^0 - \mu^0)\sigma^0 - \bar{\lambda}^0\rho^0 - 4e^{-M}\bar{\alpha}^2, \tag{4.15}$$

$$\lambda^0{}_{,v} = \lambda^0(\rho^0 - 2iE^0) + \bar{\sigma}^0\mu^0 + 4e^{-M}\alpha^2, \tag{4.16}$$

$$\lambda^0{}_{,u} = -\lambda^0(2\mu^0 + M_{,u} + 2iG^0) - \Psi_4^0, \tag{4.17}$$

$$\alpha_{,v} = \alpha(3\rho^0 - iE^0) + \bar{\sigma}^0\bar{\alpha}, \tag{4.18}$$

$$\alpha_{,u} = -\alpha(3\mu^0 + iG^0) - \lambda^0\bar{\alpha}, \tag{4.19}$$

$$\mu^0\rho^0 - \lambda^0\sigma^0 = \tfrac{1}{2}M_{,uv} + \tfrac{1}{2}i(G^0{}_{,v} - E^0{}_{,u}) - 12e^{-M}\alpha\bar{\alpha}, \tag{4.20}$$

$$\Psi_1^0 = 2\rho^0\bar{\alpha} - 2\alpha\sigma^0, \tag{4.21}$$

$$\Psi_2^0 = \mu^0\rho^0 - \lambda^0\sigma^0, \tag{4.22}$$

$$\Psi_3^0 = 2\mu^0\alpha - 2\bar{\alpha}\lambda^0. \tag{4.23}$$

Equations (4.8)–(4.23) are all entirely in terms of scale-invariant spin coefficients; thus the scale transformation freedom (3.5) has effectively been eliminated.

Returning to Fig. 1, we have flat space in region I; hence $\alpha = 0$ there. From Eqs. (4.18) and (4.19) and the uniqueness theorem of ordinary differential equations it follows that α must vanish in regions II and III. Applying the uniqueness theorem again in IV gives that α vanishes throughout the space–time (*N.B.*: the integrability conditions for (4.18) and (4.19) are satisfied automatically as a consequence of the other field equations). Hence, if Y^i and X^i can be simultaneously transformed away by a transformation (3.8) in region I, it is a consequence of the field equations that they can be transformed away everywhere.

Assuming this to be done, we have that the metric is, by (3.3), (3.4), and (4.7),

$$ds^2 = 2e^{-M}dudv + g_{ij}dx^i dx^j,$$

where

$$g^{ij} = -(\xi^i\xi^j + \bar{\xi}^i\xi^j).$$

Writing $g_{ij}dx^i dx^j$ in the form

$$-e^{-U}(e^V \cosh W(dx^2)^2 + e^{-V}\cosh W(dx^3)^2 - 2\sinh W dx^2 dx^3),$$

where

$$U = -\log(\det g_{ij}),$$

we have

$$\xi^2 = e^{(U-V)/2}\sqrt{\tfrac{1}{2}\cosh W}e^{i\theta}$$

$$\xi^2 = e^{(U+V)/2}\sqrt{\tfrac{1}{2}\cosh W}e^{i\varphi},$$

where

$$\cos(\theta - \varphi) = \tanh W.$$

By means of a spatial rotation (3.6) it is clearly possible to achieve that $\theta = \frac{1}{2}\pi - \varphi = \frac{1}{2}\sin^{-1}(\tanh W)$. From (4.8) and (4.9) all remaining spin coefficients may be expressed in terms of the functions U, V, W:

$$\rho^0 = \tfrac{1}{2}U_v, \qquad \mu^0 = -\tfrac{1}{2}U_u,$$

$$E^0 = -\tfrac{1}{2}V_v \sinh W, \qquad G^0 = \tfrac{1}{2}V_u \sinh W,$$

$$\sigma^0 = \tfrac{1}{2}iW_v - \tfrac{1}{2}V_v \cosh W,$$

$$\lambda^0 = \tfrac{1}{2}iW_u + \tfrac{1}{2}V_u \cosh W,$$

where subscripts u, v refer to partial derivatives taken with respect to these variables and 2θ has been chosen in the first or fourth quadrant.

Equations (4.11), (4.12) give

$$U_{uv} - U_u U_v \equiv e^U(e^{-U})_{uv} = 0;$$

hence

$$U = -\log[f(u) + g(v)]. \tag{4.24}$$

A coordinate transformation (3.7) could be used to make $f(u) = u$ and $g(v) = v$, but we refrain from doing this at this stage since it will turn out that in the case of colliding waves such a coordinate transformation becomes singular on the junction surfaces $u = 0$ and $v = 0$.

Equations (4.10), (4.13), (4.20), (4.15), and (4.16) give

$$2U_{vv} - U_v^2 + 2U_v M_v = W_v^2 + V_v^2 \cosh^2 W, \tag{4.25}$$

$$2U_{uu} - U_u^2 + 2U_u M_u = W_u^2 + V_u^2 \cosh^2 W, \tag{4.26}$$

$$2M_{uv} + U_{uv} = W_u W_v + V_u V_v \cosh^2 W, \tag{4.27}$$

$$2W_{uv} - U_u W_v - U_v W_u = 2V_u V_v \cosh W \sinh W, \tag{4.28}$$

$$2V_{uv} - U_u V_v - U_v V_u = -2(V_v W_u + V_u W_v)\tanh W, \tag{4.29}$$

while the components of the Weyl tensor are computed from the remaining field equations

$$\Psi_0^0 = -\tfrac{1}{2}[V_{vv}\cosh W + 2V_v W_v \times \sinh W - V_v(U_v - M_v)\cosh W] + \tfrac{1}{2}i[W_{vv} - W_v(U_v - M_v) - V_v^2 \cosh W \sinh W], \tag{4.30}$$

$$\Psi_4^0 = -\tfrac{1}{2}[V_{uu}\cosh W + 2V_u W_u \times \sinh W - V_u(U_u - M_u)\cosh W] - \tfrac{1}{2}i[W_{uu} - W_u(U_u - M_u) - V_u^2 \cosh W \sinh W], \tag{4.31}$$

$$\Psi_2^0 = \tfrac{1}{2}M_{uv} - \tfrac{1}{2}i(V_u W_v - V_v W_u)\cosh W,$$

$$\Psi_1^0 = \Psi_3^0 = 0. \tag{4.32}$$

The task of solving Eqs. (4.25)–(4.29) is not as daunting as may appear at first sight, for it turns out that Eqs. (4.28) and (4.29) are just the integrability conditions for Eqs. (4.25)–(4.28). More precisely, if V and W satisfy Eqs. (4.28) and (4.29), then, apart from the exceptional case where U, V and W are functions of u or v alone, Eqs. (4.26)–(4.28) are automatically satisfied for some function M. Thus we may concentrate on simply solving (4.28) and (4.29), obtaining M from (4.26) and (4.27) by simple integration.

5. COLLIDING PLANE WAVES—APPROXIMATE SOLUTIONS

Returning to the situation of Fig. 1 we have, throughout the space–time, a metric of the form

$$ds^2 = 2e^{-M}dudv - e^{-U}(e^V \cosh W(dx^2)^2 + e^{-V}\cosh W(dx^3)^2 - 2\sinh W dx^2 dx^3), \tag{5.1}$$

where M and U are given by (4.7) and (4.24) and M, U, V and W satisfy Eqs. (4.25)–(4.29).

In region I (Minkowski space–time) we may put

$$f(u) = g(v) = \tfrac{1}{2} \quad \text{(i.e., } U = 0\text{)}, \qquad M = V = W = 0.$$

In region II we have a plane u-wave, i.e.,

$$g = \tfrac{1}{2}, \qquad M = M(u), \qquad V = V(u), \qquad W = W(u),$$

while in III there is a plane v-wave

$$f = \tfrac{1}{2}, \qquad M = M(v), \qquad V = V(v), \qquad W = W(v).$$

From (4.30), (4.31), and (4.32) it follows that the Riemann tensor is of Petrov type N in both regions:

$$\Psi_0^0 = \Psi_1^0 = \Psi_2^0 = \Psi_3^0 = 0, \quad \Psi_4^0 = \Psi_4^0(u) \quad \text{in II}$$

and

$$\Psi_0^0 = \Psi_0^0(v), \quad \Psi_1^0 = \Psi_2^0 = \Psi_3^0 = \Psi_4^0 = 0 \quad \text{in III.}$$

The condition for constant polarization in either of these regions, $W = 0$ [g_{ij} diagonalizable by a linear coordinate transformation (3.9)], is equivalent to the condition that Ψ_4^0 or Ψ_0^0 is real in II and III, respectively.

The Lichnerowicz conditions (continuity of the metric and its first derivatives) imply that at the I–II boundary ($u = 0, v < 0$)

$$f = \tfrac{1}{2}, \quad M = V = W = f_u = M_u = V_u = W_u = 0.$$

From Eq. (4.26) it follows then that also $f_{uu} = 0$ at this boundary, and differentiating this equation with respect to u gives $f_{uuu} = 0$. On differentiating again, however, one finds

$$f_{uuuu} = -(V_{uu}^2 + W_{uu}^2)$$

and it is not necessary that V_{uu} or W_{uu} vanish at the I–II boundary since Ψ_4^0 may have a discontinuity there. In fact, assuming power series solutions at $u = 0$ (although this restriction is unnecessarily strong and may easily be weakened), we see that for $u > 0$

$$f = \tfrac{1}{2} + a_1u^4 + \cdots,$$
$$V = b_1u^2 + \cdots,$$
$$W = c_1u^2 + \cdots,$$
$$M = d_1u^2 + \cdots,$$

where

$$a_1 = -\tfrac{1}{6}(b_1^2 + c_1^2)$$

and

$$\Psi_4^0 = -b_1 - ic_1 + \cdots.$$

A similar argument at the I–III boundary ($v = 0$, $u < 0$) gives

$$g = \tfrac{1}{2} + a_2v^4 + \cdots,$$
$$V = b_2v^2 + \cdots,$$
$$W = c_2v^2 + \cdots,$$
$$M = d_2v^2 + \cdots,$$
$$\Psi_0^0 = -b_2 + ic_2 + \cdots,$$

where

$$a_2 = -\tfrac{1}{6}(b_2^2 + c_2^2).$$

In the interaction zone IV it is clear from the required continuity of U that $f(u)$ and $g(v)$ will have just the same functional forms that they take in II and III, respectively. At the II–IV boundary we must have $V_v = W_v = M_v = 0$, while at the III–IV boundary $V_u = W_u = M_u = 0$ in order to comply with the Lichnerowicz conditions. If power series are assumed for all functions in IV, it is easy to compute the leading terms from Eqs. (4.25)–(4.29):

$$V = b_1u^2 + \cdots + b_2v^2 + \cdots - \tfrac{1}{2}(c_1c_2b_2 + c_2^2b_1 + a_2b_1)u^2v^4 - \tfrac{1}{2}(c_1c_2b_1 + c_1^2b_2 + a_1b_2)u^4v^2 + \cdots,$$

$$W = c_1u^2 + \cdots + c_2v^2 + \cdots + \tfrac{1}{2}(b_2b_1c_2 - a_2c_1)u^2v^4 + \tfrac{1}{2}(b_1b_2c_1 - a_1c_2)u^4v^2 + \cdots,$$

$$M = d_1u^2 + \cdots + d_2v_2 + \cdots + \tfrac{1}{2}(c_1c_2 + b_1b_2)u^2v^2 + \cdots,$$

$$\Psi_0^0 = -b_2 + ic_2 + \cdots,$$
$$\Psi_4^0 = -b_1 - ic_1 + \cdots,$$
$$\Psi_2^0 = [c_1c_2 + b_1b_2 - 2i(b_1c_2 - b_2c_1)]uv + \cdots,$$
$$\Psi_1^0 = \Psi_3^0 = 0.$$

Further terms are easily calculated, but the basic structure of the interaction is already apparent, at least in the neighborhood of the collision plane $u = v = 0$. In linear theory Ψ_0^0 and Ψ_4^0 would remain unchanged in IV (principle of superposition) and there would be no Ψ_2^0 term. In the full nonlinear theory a Ψ_2^0 term quickly develops and the Ψ_0^0 and Ψ_4^0 become modified (higher terms in the expansion must be computed to see this).

However, in order to understand the collision process in greater detail, and in particular to discuss the nature of the solution at points far from $u = v = 0$, it will be necessary to obtain some exact solutions.

6. COLLIDING PLANE WAVES—EXACT SOLUTIONS

In order to obtain exact solutions, it is necessary to solve the pair of equations (4.28) and (4.29) for V and W. The remaining equations are then integrable for M. The situation is considerably simplified if it is assumed that both incoming waves are linearly polarized, i.e., assume $W = 0$ in II and III. W satisfies the hyperbolic differential equations (4.28), and in IV we have an initial characteristic value problem with W vanishing on the characteristics $u = 0, v = 0$. It follows[13] that W must vanish throughout the interaction zone IV. Another way of expressing this is to say that if g_{ij} can be simultaneously diagonalized in II and III, then it can be diagonalized throughout the space-time. There now remains just the single linear equation in V

$$2V_{uv} - U_uV_v - U_vV_u = 0.$$

If a change of variables is made to $f = f(u)$, $g = g(v)$, this equation becomes

$$L[V] = 2(f + g)V_{fg} + V_f + V_g = 0, \tag{6.1}$$

an Euler–Darboux equation.

It is clear, however, from the considerations of the previous section that this coordinate transformation is singular at $u = 0$ ($f = \frac{1}{2}$), and at $v = 0$ ($g = \frac{1}{2}$), since $u = O((\frac{1}{2} - f)^{1/4})$, $v = O((\frac{1}{2} - g)^{1/4})$ at these surfaces. Nevertheless, we may still apply this coordinate transformation in the interiors of regions II, III, and IV separately, assuming that $V \approx \pm [6(\frac{1}{2} - f)]^{1/2}$ at the I–II boundary, $V \approx \pm [6(\frac{1}{2} - g)^{1/2}$ at the I–III boundary, and

$$V \approx (\tfrac{1}{2} - f)^{1/2}G(g) + (\tfrac{1}{2} - g)^{1/2}F(f) + H(f, g)$$

at the boundaries of IV, where F, G, and H are regular

functions. Once the equations are solved, we may return to variables u and v by a coordinate transformation in order to connect the different patches up in a smooth manner. If it is not required that Ψ_4^0 and Ψ_0^0 have actual discontinuities at the I–II and I–III boundaries, respectively (or perhaps that they suffer something more drastic there, such as a δ-function discontinuity), one may assume $f \approx \frac{1}{2} - a^2u^n$ in which case Eq. (4.26) implies

$$V \approx \pm\,[8(1 - 1/n)(\tfrac{1}{2} - f)]^{1/2}.$$

Given $V = V_1(f)$ in II, $V = V_2(g)$ in III, $V(\frac{1}{2}, \frac{1}{2}) = 0$, it is possible to solve (6.1) in region IV explicitly by Riemann's method.[13,14] A specific Riemann–Green function, satisfying

$$\begin{aligned} \tilde{L}[R] &= R_{fg} - (R/2(f+g))_{,f} - (R/2(f+g))_{,g} = 0, \\ R_f - R/2(f+g) &= 0 \quad \text{at } g = g_0, \\ R_g - R/2(f+g) &= 0 \quad \text{at } f = f_0, \\ R(f_0, g_0) &= 1 \end{aligned}$$

is[14]

$$R(f,g;f_0,g_0) = \left(\frac{f+g}{f_0+g_0}\right)^{1/2} P_{-1/2}\left(1 + 2\,\frac{(f-f_0)(g-g_0)}{(f+g)(f_0+g_0)}\right),$$

where $P_{-1/2}$ is the Legendre function of order $-\frac{1}{2}$,

$$\begin{aligned} P_{-1/2}(1+2z) &= F(\tfrac{1}{2}, \tfrac{1}{2}; 1; -z) \\ &= 1 - (\tfrac{1}{2})^2 z + \left(\frac{1\cdot 3}{2\cdot 4}\right)^2 z^2 - \left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)^2 z^3 + \cdots. \end{aligned} \tag{6.2}$$

In region IV ($f < \frac{1}{2}, g < \frac{1}{2}$) the solution of (6.1) is obtained by integrating $RL[V] - V\tilde{L}[R]$ over the rectangle $PNML$ of Fig. 2:

$$\begin{aligned} V(f_0,g_0) = (RV)_M &- \int_L^M R(f,g;f_0,g_0)[V_f + V/2(f+g)]df \\ &- \int_N^M R(f,g;f_0,g_0)[V_g + V/2(f+g)]dg, \end{aligned}$$

i.e.,

$$\begin{aligned} V(f,g) = &-(f+g)^{-1/2}\left[\int_f^{1/2} P_{-1/2}\left(1 + \frac{2(\xi - f)(\frac{1}{2} - g)}{(\xi + \frac{1}{2})(f+g)}\right)\frac{d}{d\xi}\right. \\ &\times [\sqrt{\tfrac{1}{2} + \xi}\,V_1(\xi)]d\xi + \int_g^{1/2} P_{-1/2}\left(1 + \frac{2(\frac{1}{2} - f)(\eta - g)}{(\eta + \frac{1}{2})(f+g)}\right) \\ &\left.\times \frac{d}{d\eta}[\sqrt{\tfrac{1}{2} + \eta}\,V_2(\eta)]d\eta\right]. \end{aligned} \tag{6.3}$$

If $V(f)$ and $V(g)$ are regular functions for $f > -\frac{1}{2}$, $g > -\frac{1}{2}$, then in the region $f + g > 0$ the integrands are regular. For $P_{-1/2}(1 + 2z)$ becomes singular only at $z = -1$ which occurs at $\xi = -g$ which is less than f, and $\eta = -f$ which is less than g. To investigate the nature of the solution as $f + g \to 0$, we must know the behavior of $P_{-1/2}(1 + 2z)$ as $z \to \infty$. This cannot be obtained from the power series expansion (6.2) which converges only for $|z| < 1$, but can be obtained from the Laplace integral formula[15]

$$P_{-1/2}(z) = \frac{1}{\pi}\int_0^\pi d\phi\,[z + (z^2 - 1)^{1/2}\cos\phi]^{1/2}. \tag{6.4}$$

This can be written as

$$m(z^2 - 1)^{-1/2}\pi^{-1}K(m),$$

where

$$m = 2(z^2 - 1)^{1/2}/[z + (z^2 - 1)^{1/2}]$$

and $K(m)$ is the standard elliptic integral having the asymptotic behavior[16]

$$\lim_{m\to 1}[K(m) - \log 4 + \tfrac{1}{2}\log(1 - m)] = 0.$$

Hence

$$\begin{aligned} P_{-1/2}(1 + 2z) = \pi^{-1}[&z^{-1/2}\log z \\ &+ 4z^{-1/2}\log 2 + O(z^{-3/2}\log z)] \quad \text{as } z \to \infty. \end{aligned} \tag{6.5}$$

Thus from (6.3) we have that for $\frac{1}{2} > f > -\frac{1}{2}$, $\frac{1}{2} > g > -\frac{1}{2}$, as $f + g \to 0$,

$$V(f,g) = H(f,g)\log(f+g),$$

where $H(f,g)$ is regular at $f + g = 0$. Using $U = -\ln(f+g)$ and Eqs. (4.25)–(4.32), we have that the behavior of M and Ψ^0's at $f + g = 0$ may be computed:

$$\begin{aligned} M &\approx \tfrac{1}{2}(1 - H^2)\log(f+g), \\ \Psi_0^0, \Psi_4^0 &\approx -\tfrac{1}{4}H(1 - H^2)(f+g)^{-2}, \\ \Psi_2^0 &\approx -\tfrac{1}{4}(1 - H^2)(f+g)^{-2}. \end{aligned}$$

Thus $f + g = 0$ is a real curvature singularity of the space-time which cannot be eliminated by any coordinated transformation. It is a curious fact that arbitrarily weak incoming waves *must* produce curvature singularities in the interaction zone, proving an earlier conjecture of the author.[4]

Although (6.3) is an explicit solution of (6.1) for arbitrary incoming wave functions $V_1(f)$ and $V_2(g)$, it is in general very difficult to perform the integrations in (6.3). A specific solution which has been found more or less by trial and error, and which has

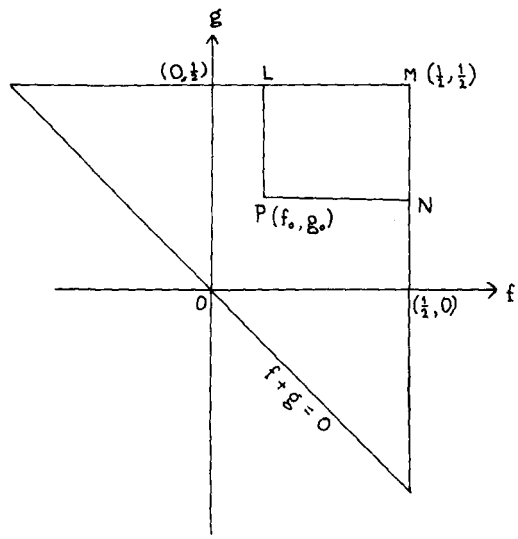


FIG. 2. The f–g picture. In coordinates $u' = f(u)$, $v' = g(v)$ the interaction zone IV is the region inside the triangle. The sloping line $f + g = 0$ is the singular boundary, while the sides $g = \frac{1}{2}$, $f = \frac{1}{2}$ are the II–IV and III–IV boundaries, respectively. Given $V = V_0(f)$ on $g = \frac{1}{2}$ and $V = V_1(g)$ on $f = \frac{1}{2}$, $V(f,g)$ may be obtained explicitly at any point P in IV by a suitable integration around the rectangle $PNML$.

the desired behavior at $f = \frac{1}{2}$ and $g = \frac{1}{2}$ described above, is

$$V = k_1 \tanh^{-1}\left(\frac{\frac{1}{2} - f}{\frac{1}{2} + g}\right)^{1/2} + k_2 \tanh^{-1}\left(\frac{\frac{1}{2} - g}{\frac{1}{2} + f}\right)^{1/2}. \tag{6.6}$$

In region II we must have $V = V_1(f) = k_1 \tanh^{-1}(\frac{1}{2} - f)^{1/2}$, and in region III $V = V_2(g) = k_2 \tanh^{-1}(\frac{1}{2} - g)^{1/2}$. The complete solution is expressed most compactly by putting

$$p = \sqrt{\tfrac{1}{2} - f}, \quad q = \sqrt{\tfrac{1}{2} - g}, \quad r = \sqrt{\tfrac{1}{2} + f},$$
$$w = \sqrt{\tfrac{1}{2} + g}, \quad t = \sqrt{f + g},$$

where

$$f = \tfrac{1}{2} - (au)^{n_1}\theta(u), \quad g = \tfrac{1}{2} - (bv)^{n_2}\theta(v),$$

($n_i \geq 1$) and $\theta(x)$ is the Heaviside step function.

Then

$$V = k_1 \tanh^{-1} p/w + k_2 \tanh^{-1} q/r,$$
$$V_1 = k_1 \tanh^{-1} p, \quad V_2 = k_2 \tanh^{-1} q,$$

where

$$k_i^2 = 8(1 - 1/n_i)$$

in order to comply with the junction conditions across the I–II and I–III boundaries. Integration of Eqs. (4.25) and (4.26) results in

$$M = [1 - k_1 k_2 - \tfrac{1}{4}(k_1 - k_2)^2]\log t + \tfrac{1}{4}k_1^2 \log w + \tfrac{1}{4}k_2^2 \log r + \tfrac{1}{2}k_1 k_2 \log(pq + rw).$$

(i) Putting $n_1 = n_2 = 4$, $k_1 = k_2 = -\sqrt{6}$ gives

$$e^M = (wr)^{3/2}(pq + rw)^3 t^{-5},$$
$$\Psi_4^0 = \sqrt{6}a^2\theta(u)t^{-4}r^{-3}(pq + rw)[1 - (p - q)^2 + 8pqrw],$$
$$\Psi_0^0 = \sqrt{6}b^2\theta(v)t^{-4}w^{-3}(pq + rw)[1 - (p - q)^2 + 8pqrw],$$
$$\Psi_2^0 = 2a^2b^2uv\theta(u)\theta(v)t^{-4}r^{-1}w^{-1}[3(pq + rw)^2 - 2pqrw],$$

which is the earlier solution given by the author[4] representing two colliding shock waves. In region II, where $v < 0$, only Ψ_4^0 is nonvanishing, and it becomes singular at $u = a^{-1}$, where $r = 0$. The behavior in region III is similar.

(ii) Putting $n_1 = n_2 = 2$, $k_1 = k_2 = -2$ gives

$$e^M = rwt^{-3}(pq + rw)^2,$$
$$\Psi_4^0 = a^2[w^{-1}\delta(au) + 3\theta(u)t^{-4}r^{-2}wq(pq + rw)],$$
$$\Psi_0^0 = b^2[r^{-1}\delta(bv) + 3\theta(v)t^{-4}w^{-2}rp(pq + rw)],$$
$$\Psi_2^0 = ab\theta(u)\theta(v)t^{-4}r^{-1}w^{-1}[(pq + rw)^2 - pqrw],$$

which is the solution of Penrose and Kahn[5] representing two colliding impulse waves. Ψ_4^0 and Ψ_0^0 suffer a δ-function discontinuity at ($u = 0, v < 0$) and ($u < 0$, $v = 0$), respectively. Both solutions have the singular behavior discussed above at $f + g = 0$, where $t = 0$.

It should be noted however that the exact solutions discussed in this section are still not the most general ones representing two linearly polarized waves in collision. It has been tacitly assumed that the polarization of the two waves match up in such a way that the metric may be simultaneously diagonalized in regions I, II, and III. In general, there will be a constant phase difference between the waves making it impossible to do this.

7. COLLIDING WAVEPACKETS OF FINITE ENERGY

The solution (ii) above representing two colliding impulse waves cannot be regarded as the limiting case of a physically realistic situation since the energy content of the incoming waves is apparently infinite. This is certainly true if the energy flux is calculated in the linearized approximation in region II by the Landau–Lifschitz pseudotensor[17]

$$ct^{01} = \frac{c^5}{8\pi G}\left|\int^u \Psi_4^0(u)du\right|^2 = \frac{c^5}{8\pi G}a^2\theta(u)$$

for $\Psi_4^0 = a\delta(u)$. Thus there is a constant energy flux continuing indefinitely after $u = 0$. In the full nonlinear theory two possible definitions of gravitational energy flux have been given by Penrose[18] in terms of the focusing power of gravitational fields. One of these definitions leads again to an infinite energy flux while the other gives zero flux. In either case impulse waves cannot be regarded as physically reasonable solutions. Following a suggestion of Penrose,[5] let us now assume that each incoming wave consists of a pair of equal and opposite impulse waves in succession (Fig. 3),

$$\Psi_4^0 = a\delta(u) - a\delta(u - u_0), \quad u_0 > 0. \tag{7.1}$$

In this case both the linearized pseudotensor approximation and Penrose's definition give the same finite total energy per unit area

$$\sqrt{2}\frac{c^4}{8\pi G}a^2u_0,$$

the total time for this burst of energy being $\sqrt{2}c^{-1}u_0$.

The exact solution having the curvature profile (7.1) is best found by putting

$$F = e^{(V-U)/2}, \quad G = e^{(U+V)/2}, \quad M = 0,$$

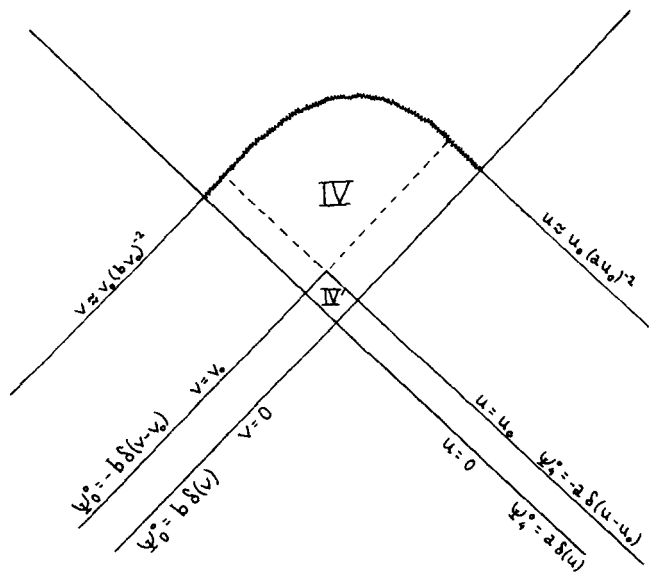


FIG. 3. Incoming wave packets of finite energy. If each incoming wave consists of a pair of equal and opposite impulse waves in close succession at $u = 0$ and $u = u_0$, $v = 0$ and $v = v_0$, respectively, they will both have finite total energy flux. The interaction region IV again has a singular boundary.

so that in II

$$ds^2 = 2dudv - F^2dx^2 - G^2dy^2.$$

The field equations (4.26) and (4.31) read

$$F_{uu}G + FG_{uu} = 0,$$
$$\Psi_4^0 = -F_{uu}/F.$$

By substituting (7.1) the solution of these equations is seen to be

$$F = 1 - a^2u_0(u - u_0) - au_0,$$
$$G = 1 - a^2u_0(u - u_0) + au_0.$$

Hence

$$f = FG - \tfrac{1}{2} = \tfrac{1}{2} - a^2u^2\theta(u) + (u - u_0)^2a^2(1 + a^2u_0^2)\theta(u - u_0),$$

and $U = -\log(f + \frac{1}{2})$ becomes singular at $f = -\frac{1}{2}$, where

$$u = (1/a^2u_0)(1 + a^2u_0^2 - au_0) \approx [u_0/(au_0)^2] \quad \text{for } au_0 \ll 1.$$

The last approximation appears to be reasonable for bursts such as those reported by Weber having a total flux $\approx 6 \times 10^9$ ergs/cm^2, and duration $\sqrt{2}cu_0 \approx 0.5$ sec. For then, $u_0 \approx 10^{10}$ cm, $a \approx 5 \times 10^{-25}$ cm^{-1}, $au_0 \approx 5 \times 10^{-15}$. If the event is occurring at the center of the galaxy, using the fact that a is inversely proportional to distance from the source, we see that the approximation would break down at distances $\lesssim 1,500$ km, where $au_0 \approx 1$. Thus for two sources which are very close to each other the waves will interfere in a highly nonlinear manner.

The function $V = -\log F/G$ when expressed in terms of f in region II is

$$V(f) = \begin{cases} -2\tanh^{-1}(\frac{1}{2} - f)^{1/2} & \text{for } \frac{1}{2} > f > \frac{1}{2} - a^2u_0^2 \\ -2\tanh^{-1}[1 + (au_0)^{-2}(\frac{1}{2} + f)^{-1/2}] & \text{for } -\frac{1}{2} < f < \frac{1}{2} - a^2u_0^2. \end{cases} \quad (7.2)$$

Assuming a similar form for the wave profile in III,

$$\Psi_0^0 = b\delta(v) - b\delta(v - v_0),$$

we can calculate $V(f, g)$ in region IV from Eq. (6.3).

In region IV′ ($u < u_0$, $v < v_0$) where only the first pulses of the incoming waves have interacted the solution is exactly given by solution (ii) of Sec. 6. The most interesting feature of the solution in this region is that it is dominated by Ψ_2^0 which jumps discontinuously across the boundaries $u = 0, v = 0$ from zero to a finite value

$$\Psi_0^0 = ab[1 + O(abuv)].$$

Ψ_4^0 and Ψ_0^0, however, remain small (assuming $au_0 \ll 1, bv_0 \ll 1$),

$$\Psi_4^0 \approx 3a^2bv[1 + O(b^2v^2)],$$
$$\Psi_0^0 \approx 3b^2au[1 + O(a^2u^2)].$$

In the region $v > v_0$, $u > u_0$ where both pulses have interacted, it is not possible to evaluate the integral (6.3) explicitly. However, by using the power series (6.2) and the asymptotic form (6.5) for $P_{-1/2}(1 + 2z)$ and substituting (7.2) into (6.3), the following approximate results emerge, valid for $au_0, bv_0 \approx O(\epsilon) \ll 1$:

$$\Psi_2^0 = -abh_1h_2(f + g)^{-2}au_0bv_0[1 + O(\epsilon^2)],$$
$$\Psi_4^0 = a^4u_0^2bv_0[3h_1^{-3} - (f + g)^{-2}(1 + 2h_1^{-1}) + O(\epsilon^2)],$$
$$\Psi_0^0 = b^4v_0^2au_0[3h_2^{-3} - (f + g)^{-2}(1 + 2h_2^{-1}) + O(\epsilon^2)],$$

where h_1 and h_2, $0 < h_i < 1$, are defined by

$$u = (1 - h_1)(a^2u_0)^{-1}, \quad v = (1 - h_2)(b^2v_0)^{-1},$$

and

$$f + g = h_1^2 + h_2^2 - 1 + O(\epsilon^2) > 0.$$

Thus in the interaction zone Ψ_2^0 jumps back to small but nonvanishing values $\approx O(abau_0bv_0) = abO(\epsilon^2)$. On the other hand $\Psi_4^0 = a^2O(\epsilon^3)$, $\Psi_0^0 = b^2O(\epsilon^3)$ are smaller by one order of magnitude in ϵ, and the solution is dominated by the Ψ_2^0 term. The most impressive feature is that Ψ_2^0 shows no sign of dying off the further one goes into the interaction zone. Indeed the tendency is to increase indefinitely as $h_1^2 + h_2^2 \to 1$, which is just the singular behavior at $f + g = 0$ discussed in Sec. 6. It is, however, very unlikely that this singular behavior is characteristic of real gravitational waves. It is almost certainly a feature due to the very high symmetry of plane waves and will be lost when curved wavefronts are considered. However, the possible presence of a small but persistent curvature in the wake of two colliding gravitational waves even when the wavefronts are curved is by no means ruled out. This could prove to be of some interest in cosmology if, as seems likely on the basis of Weber's measurements, the galaxies are strong emitters of gravitational waves which will be in constant collision with each other.

Finally there is another aspect of the plane wave solutions which may carry over to some extent in the curved wavefront case. In the case of single impulse waves, solution (ii) of Sec. 6 shows that the amplitude of the u-pulse is aw^{-1} which $\to \infty$ as $g \to -\frac{1}{2}$, i.e., as $v \to a^{-1}$. Thus the u-pulse is amplified by the v-pulse and eventually becomes singular, and the v-pulse is similarly amplified by the u-pulse. For the double impulse waves considered in this section, a detailed calculation shows a similar behavior. To first approximation, neglecting finite non-δ-function contributions.

$$\Psi_4^0 \approx (\tfrac{1}{2} + g)^{-1/2}a[\delta(u) - \delta(u - u_0)] \approx h_2^{-1}a[\delta(u) - \delta(u - u_0)],$$
$$\Psi_0^0 \approx h_1^{-1}b[\delta(v) - \delta(v - v_0)].$$

Thus again the waves are amplified by a factor which becomes infinitely large as $h_i \to 0$, i.e., $u \to (a^2u_0)^{-1}$, $v \to (b^2v_0)^{-1}$. The eventual singular behavior is just another aspect of Penrose's result[10] that plane gravitational waves act as a perfect astigmatic lens. It is certainly false for waves with curved fronts, but such waves may still act as imperfect lenses providing a certain degree of focussing and amplification for each other. In view of the distressingly high energy fluxes reported by Weber, it is not impossible to rule out the possibility that such amplification as this is occurring

to produce apparent fluxes which are considerably higher than would be expected on the basis of the linearized approximation. Clearly a better understanding of the interaction of gravitational waves with more realistic wavefronts is a problem of considerable importance.

[1] J. Weber, Phys. Rev. Letters **22**, 1320 (1969).
[2] G. W. Gibbons and S. W. Hawking, to be published.
[3] J. Weber, Phys. Rev. Letters **25**, 180 (1970).
[4] P. Szekeres, Nature **228**, 1183 (1970).
[5] K. Kahn and R. Penrose, Nature **229**, 185 (1971).
[6] J. Ehlers and W. Kundt, in *Gravitation, An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[7] N. Rosen, Phys. Z. Sowjet. **12**, 366 (1937).
[8] A. Lichnerowicz, *Théories relativistes de la gravitation et de l'electromagnetisme* (Masson, Paris, 1955).
[9] J. Synge, *Relativity, The General Theory* (North-Holland, Amsterdam, 1960).
[10] R. Penrose, Rev. Mod. Phys. **37**, 215 (1965).
[11] R. Penrose in ARL Technical Documentary Report No. 63-56 (1963).
[12] E. Newman and R. Penrose, J. Math. Phys. **3**, 902 (1962), referred to as NP in this paper.
[13] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1961), Vol. II, p. 450.
[14] A. G. Mackie, *Boundary Value Problems* (Oliver and Boyd, London 1965), Chap. 6.
[15] E. Whittaker and G. Watson, *A Course of Modern Analysis* (Cambridge U.P., Cambridge, 1965), p. 314.
[16] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965), p. 591.
[17] L. Landau and E. Lifshitz, *The Classical Theory of Fields* (Pergamon, London, 1961), p. 352.
[18] R. Penrose, in *Perspectives in Geometry and Relativity*, edited by B. Hoffmann (Indiana U.P., Bloomington, 1966).

Random Walks in Polytype Structures*

M. F. Thorpe†
Physikalisches Institut der Universität, 7 Stuttgart, W. Germany

(Received 11 August 1971)

It is shown that the total number of walks, starting and ending at the same point and having the same number of steps, is the same for all polytype structures in the fcc, hcp series and in the zincblende, wurtzite series and is independent of the starting point. This result is proved by showing that the eigenvalues of a simple Hamiltonian are the same within the two series considered. A relation is found between random walks in the two series of structures that is useful in extending currently available tables of random walks for the zincblende structure.

I. INTRODUCTION

The study of random walks on lattices is of interest in itself and plays an important role in the statistical mechanics of solids.[1] Many problems concerned with the thermodynamics of crystals involve the summing of diagrams on lattices which can often be related to the diagrams in a simple random walk process (e.g., the study of the Ising model above the transition temperature[1]). This paper is concerned with random walks on two of the simplest series of polytype structures. The simplest structures occur in the fcc, hcp series, where the close packed layers are arranged, ABCABC$\cdots$ in fcc, and ABABAB$\cdots$ in hcp. An infinite number of other possibilities may be imagined —some with a repeating pattern and some without. All these possible structures belong to the fcc, hcp polytype series. A general discussion of polytype structures is given by Verma and Krishna[2]: "Polytypism may be defined, in general, as the ability of a substance to crystallize in a number of different modifications, in all of which two dimensions of the unit cell are the same while the third is a variable integral multiple of a common unit. The different polytypic modifications can be regarded as built up of layers of structure stacked parallel to each other at constant intervals along the variable dimension. The two unit-cell dimensions parallel to these layers are the same for all modifications. The third dimension depends on the stacking sequence, but is always an integral multiple of the layer spacing. Different manners of stacking these layers may result in structures having not only different morphologies but even different lattice types and space groups." As well as the fcc, hcp polytype series, we also discuss the zincblende, wurtzite polytype series in this paper. Each structure in this series may be derived by placing two atoms at each site in the fcc, hcp series to form two interpenetrating sublattices. Zincblende itself has zinc on one sublattice and sulphur on the other whilst diamond is an example of a crystal having only one type of atom and the zincblende structure.

We write down a simple Hamiltonian, containing a single state at each site, that permits hopping between nearest neighbors only. The lth moment of the density of states formed from the eigenvalues of this Hamiltonian is just the number of ways of returning to a starting point from a walk of l steps. The theorem is proved by using a unitary transformation that shows that the density of states is identical for all structures within a polytype series. It is also shown that the number of returns is independent of the starting point.

In Sec. VI, a connection is found between the density of states for the fcc, hcp series and the zincblende, wurtzite series. That such a connection exists is not surprising as each structure in the latter series may be derived from a structure in the former series. An integral expression for the density of states of the fcc and zincblende structures is used to derive a relationship between the number of returns to a starting point for the two polytype series. Tables are given for the total number of returns for walks with up to nine steps in the fcc, hcp series and up to 18 steps in the zincblende, wurtzite series.

II. THE HAMILTONIAN

We define a Hamiltonian

$$H = V \sum_{\langle ij \rangle} |\phi_i\rangle \langle \phi_j|, \tag{1}$$

where V is the overlap between states $|\phi_i\rangle$ and $|\phi_j\rangle$ on

to produce apparent fluxes which are considerably higher than would be expected on the basis of the linearized approximation. Clearly a better understanding of the interaction of gravitational waves with more realistic wavefronts is a problem of considerable importance.

[1] J. Weber, Phys. Rev. Letters **22**, 1320 (1969).
[2] G. W. Gibbons and S. W. Hawking, to be published.
[3] J. Weber, Phys. Rev. Letters **25**, 180 (1970).
[4] P. Szekeres, Nature **228**, 1183 (1970).
[5] K. Kahn and R. Penrose, Nature **229**, 185 (1971).
[6] J. Ehlers and W. Kundt, in *Gravitation, An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[7] N. Rosen, Phys. Z. Sowjet. **12**, 366 (1937).
[8] A. Lichnerowicz, *Théories relativistes de la gravitation et de l'electromagnetisme* (Masson, Paris, 1955).
[9] J. Synge, *Relativity, The General Theory* (North-Holland, Amsterdam, 1960).
[10] R. Penrose, Rev. Mod. Phys. **37**, 215 (1965).
[11] R. Penrose in ARL Technical Documentary Report No. 63-56 (1963).
[12] E. Newman and R. Penrose, J. Math. Phys. **3**, 902 (1962), referred to as NP in this paper.
[13] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1961), Vol. II, p. 450.
[14] A. G. Mackie, *Boundary Value Problems* (Oliver and Boyd, London 1965), Chap. 6.
[15] E. Whittaker and G. Watson, *A Course of Modern Analysis* (Cambridge U.P., Cambridge, 1965), p. 314.
[16] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965), p. 591.
[17] L. Landau and E. Lifshitz, *The Classical Theory of Fields* (Pergamon, London, 1961), p. 352.
[18] R. Penrose, in *Perspectives in Geometry and Relativity*, edited by B. Hoffmann (Indiana U.P., Bloomington, 1966).

Random Walks in Polytype Structures*

M. F. Thorpe†
Physikalisches Institut der Universität, 7 Stuttgart, W. Germany

(Received 11 August 1971)

It is shown that the total number of walks, starting and ending at the same point and having the same number of steps, is the same for all polytype structures in the fcc, hcp series and in the zincblende, wurtzite series and is independent of the starting point. This result is proved by showing that the eigenvalues of a simple Hamiltonian are the same within the two series considered. A relation is found between random walks in the two series of structures that is useful in extending currently available tables of random walks for the zincblende structure.

I. INTRODUCTION

The study of random walks on lattices is of interest in itself and plays an important role in the statistical mechanics of solids.[1] Many problems concerned with the thermodynamics of crystals involve the summing of diagrams on lattices which can often be related to the diagrams in a simple random walk process (e.g., the study of the Ising model above the transition temperature[1]). This paper is concerned with random walks on two of the simplest series of polytype structures. The simplest structures occur in the fcc, hcp series, where the close packed layers are arranged, ABCABC$\cdots$ in fcc, and ABABAB$\cdots$ in hcp. An infinite number of other possibilities may be imagined —some with a repeating pattern and some without. All these possible structures belong to the fcc, hcp polytype series. A general discussion of polytype structures is given by Verma and Krishna[2]: "Polytypism may be defined, in general, as the ability of a substance to crystallize in a number of different modifications, in all of which two dimensions of the unit cell are the same while the third is a variable integral multiple of a common unit. The different polytypic modifications can be regarded as built up of layers of structure stacked parallel to each other at constant intervals along the variable dimension. The two unit-cell dimensions parallel to these layers are the same for all modifications. The third dimension depends on the stacking sequence, but is always an integral multiple of the layer spacing. Different manners of stacking these layers may result in structures having not only different morphologies but even different lattice types and space groups." As well as the fcc, hcp polytype series, we also discuss the zincblende, wurtzite polytype series in this paper. Each structure in this series may be derived by placing two atoms at each site in the fcc, hcp series to form two interpenetrating sublattices. Zincblende itself has zinc on one sublattice and sulphur on the other whilst diamond is an example of a crystal having only one type of atom and the zincblende structure.

We write down a simple Hamiltonian, containing a single state at each site, that permits hopping between nearest neighbors only. The lth moment of the density of states formed from the eigenvalues of this Hamiltonian is just the number of ways of returning to a starting point from a walk of l steps. The theorem is proved by using a unitary transformation that shows that the density of states is identical for all structures within a polytype series. It is also shown that the number of returns is independent of the starting point.

In Sec. VI, a connection is found between the density of states for the fcc, hcp series and the zincblende, wurtzite series. That such a connection exists is not surprising as each structure in the latter series may be derived from a structure in the former series. An integral expression for the density of states of the fcc and zincblende structures is used to derive a relationship between the number of returns to a starting point for the two polytype series. Tables are given for the total number of returns for walks with up to nine steps in the fcc, hcp series and up to 18 steps in the zincblende, wurtzite series.

II. THE HAMILTONIAN

We define a Hamiltonian

$$H = V \sum_{\langle ij \rangle} |\phi_i\rangle \langle \phi_j|, \tag{1}$$

where V is the overlap between states $|\phi_i\rangle$ and $|\phi_j\rangle$ on

neighboring sites, the summation in (1) being restricted so that nearest neighbor pairs are counted only once. The states are normalized so that

$$\langle \phi_i | \phi_j \rangle = \delta_{ij}.$$

All the structures considered in this paper may be regarded as having M planes stacked vertically, each plane containing N atoms. The density of states $n(E)$ per atom for the Hamiltonian (1) is given by

$$n(E) = -(1/\pi NM)\mathrm{Im}\sum_i \langle \phi_i | 1/(E-H) | \phi_i \rangle, \qquad (2)$$

where the energy E has a small positive imaginary part. The lth moment of the density of states is

$$\int_{-\infty}^{+\infty} E^l n(E) dE = (1/NM)\sum_i \langle \phi_i | H^l | \phi_i \rangle = V^l r_l, \qquad (3)$$

where r_l is the number of returns to a given starting point in a walk of l steps; an average being taken over all possible starting points in the structure and the thermodynamic limit $N, M \to \infty$ taken to eliminate surface effects. In fact it is unnecessary to take an average in (3) as it is shown at the end of Sec. IV, that $\langle \phi_i | H^l | \phi_i \rangle$ is independent of i for all structures considered in this paper. For the moment, however, we are content with an average. From (3) we see that it is a necessary and sufficient condition that the density of states $n(E)$ be independent of the stacking sequence for the r_l also to be independent of the stacking sequence and therefore identical for all the members of the polytype series.

III. THE FCC, HCP POLYTYPE SERIES

This series consists of close packed layers, as shown in Fig. 1, stacked vertically in the z direction one on top of the other. There are three types of layers—A, B, C all equivalent to within a horizontal displacement. The fcc sequence of layers is $ABCABC\cdots$ and hcp is $ABABAB\cdots$. Each close packed layer is a Bravais lattice and so it is useful to use Bloch's theorem to define states within each plane

$$|\mathbf{k}, n\rangle = (1/\sqrt{N})\sum_i e^{i\mathbf{k}\cdot\mathbf{i}} |\phi_{i,n}\rangle, \qquad (4)$$

where $\mathbf{k}$ is a two-dimensional vector parallel to the plane. We use an additional label n for each plane where $1 \leq n \leq M$ and the summation over i is only over the N states $|\phi_{i,n}\rangle$ within the plane labeled by n. The origin of coordinates is chosen arbitrarily for the first plane and in a manner to be prescribed for subsequent planes. We see from (1) and (4) that

$$\begin{aligned}\langle \mathbf{k}, n | H | \mathbf{k}, n \rangle &= V\sum_\delta e^{i\mathbf{k}\cdot\delta} \\ &= 2V(\cos k_x a + 2\cos\tfrac{1}{2}k_x a \cos\tfrac{1}{2}\sqrt{3}k_y a) \\ &= \alpha_{\mathbf{k}} \end{aligned} \qquad (5)$$

where the summation over δ is over the six nearest neighbors within the plane. It is important to note that $\alpha_{\mathbf{k}}$ is real and that H only connects states with the same $\mathbf{k}$ vector in accordance with Bloch's theorem. ($\mathbf{k}$ is defined to be within the first Brillouin zone to give exactly N states for each plane; the possibility of two states differing by a reciprocal lattice vector is therefore discounted.)

We now place the next layer on top of the first layer. Each atom in the first layer has three nearest neighbors in the second. We choose the atom in the second plane that lies nearest to the origin in the first plane and is in the yz plane (see Fig. 1) as the origin for the second plane. This is a unique prescription and we obtain the following matrix elements between adjacent planes:

$$\begin{aligned}&\langle \mathbf{k}, n | H | \mathbf{k}, n+1 \rangle \\ &\quad = \theta_{\mathbf{k}} \text{ for cyclic sequence } AB, BC, CA, \\ &\quad = \theta_{\mathbf{k}}^* \text{ for anticyclic sequence } BA, CB, AC, \end{aligned} \qquad (6)$$

where

$$\theta_{\mathbf{k}} = V(1 + 2\cos\tfrac{1}{2}k_x a \exp i\tfrac{1}{2}\sqrt{3}k_y a). \qquad (7)$$

Because of the translation group perpendicular to the z axis, the Hamiltonian only connects states with the same $\mathbf{k}$ vector in adjacent planes. From Eqs. (5) and (6), we see that by using Bloch's theorem perpendicular to the z axis, the Hamiltonian is split into N blocks each one being $M \times M$ and characterised by a $\mathbf{k}$ vector. We can consider the eigenvalues of each block separately. For hcp we have an $ABABAB\cdots$ stacking sequence and the structure of one of the blocks is

$$\begin{array}{cccccccccc} \alpha & \theta & & & & & & & & \\ \theta^* & \alpha & \theta^* & & & & & & & \\ \theta & \alpha & \theta & & & & & & & \\ & \theta^* & \alpha & \theta^* & & & & & & \\ & & \theta & \alpha & \theta & & & & & \\ & & & \theta^* & \alpha & \theta^* & & & & \\ & & & & \theta & - & - & & & \\ & & & & & - & - & - & & \\ & & & & & & - & - & - & \\ & & & & & & & - & & \end{array}, \qquad (8)$$

where the $\mathbf{k}$ label on α, θ has been dropped and the only nonzero elements lie on and just above and below the diagonal. The layers are stacked alternatively cyclically, AB and anticyclically, BA and so we get a sequence $\theta\ \theta^*\ \theta\ \theta^*\ \theta\ \theta^* \cdots$ above the diagonal. The

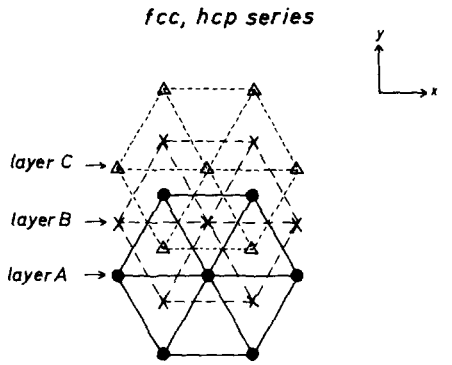
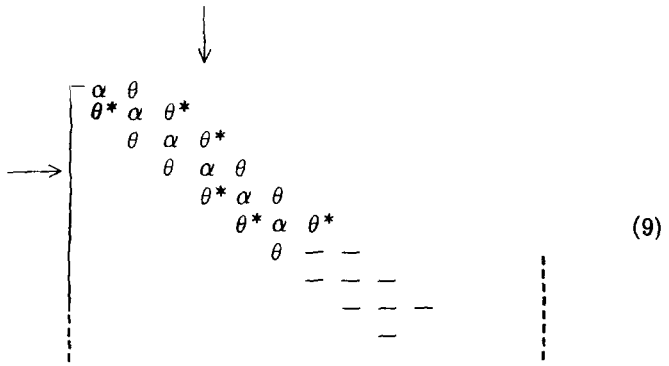


FIG. 1. The three possible stacking positions for the fcc, hcp series. The x, y axes lie in the plane perpendicular to the stacking axis z. The distance between nearest neighbors in the plane is a and the distance between successive layers is $c = \sqrt{\frac{2}{3}}\,a$. A polytype structure may be constructed from any sequence of the three layers A, B, and C provided only that the same layer does not occur twice in adjacent layers, e.g., $ABCBABC$ is allowed whereas $ABBACBA$ is not.

eigenvalues of (8) are unaltered by any unitary transformation. The row and column indicated by arrows in (8) are multiplied by θ/θ^* and θ^*/θ, respectively. The transformed matrix becomes



(9)

and the stacking sequence that leads to (9) is $ABACAB\cdots$, i.e., the second B has been changed to C. It is clear, therefore, that we can have either B or C between two A's, A or B between two C's, or C or A between two B's, and the Hamiltonian (1) will have the same eigenvalues. A little thought will convince the reader that any desired sequence may be generated by this procedure. For example, we can go from hcp to fcc in the following way:

$$
\begin{array}{l}
A\ B\ A\ B\ A\ B\ A\ B\ A\ B\ \cdots \\
A\ B\ C\ B\ A\ B\ A\ B\ A\ B\ \cdots \\
A\ B\ C\ B\ C\ B\ A\ B\ A\ B\ \cdots \\
A\ B\ C\ A\ C\ B\ A\ B\ A\ B\ \cdots \\
A\ B\ C\ A\ C\ B\ C\ B\ A\ B\ \cdots \\
A\ B\ C\ A\ C\ A\ C\ B\ A\ B\ \cdots \\
A\ B\ C\ A\ B\ A\ C\ B\ A\ B\ \cdots \\
A\ B\ C\ A\ B\ A\ C\ B\ C\ B\ \cdots \\
A\ B\ C\ A\ B\ A\ C\ A\ C\ B\ \cdots \\
A\ B\ C\ A\ B\ A\ B\ A\ C\ B\ \cdots \\
A\ B\ C\ A\ B\ C\ B\ A\ C\ B\ \cdots
\end{array}
\qquad (10)
$$

It is clear that we are generating an fcc sequence. We are also generating a CBA for every ABC, and will therefore finish up with a structure half of which is $ABCABCABC\cdots$ and the other half $CBACBACBA\cdots$. There will be an interface between these two sequences; however it will have negligible weight in the thermodynamic limit. We could generate hcp from fcc by reversing the above procedure used in (10).

If we count a cyclic sequence AB, BC, CA as $+1$ and an anticyclic sequence BA, CB, AC as -1, then the sum of all these numbers for a given structure may be called the cyclicity C of that structure and is conserved under the unitary transformation described in this section. Thus the hcp structure with an odd number of layers has $C = 0$, whereas the fcc structure has $C = M - 1$, where M is the number of layers. A structure with cyclicity $-C$ can be formed from a structure with cyclicity $+C$ by reflection in the xy plane. Therefore by inserting an appropriate small number of interfaces (one in the case of hcp to fcc), it is possible to generate any structure and the interfaces will have negligible weight in the thermodynamic limit.

The equivalence of the number of returns in a random walk on the fcc and hcp lattices was first demonstrated by C. Domb and M. F. Sykes.[3] Their approach involves the detailed consideration of lattice diagrams and it is not easy to see how it would generalize to all close packed structures.

IV. THE ZINCBLENDE, WURTZITE POLYTYPE SERIES

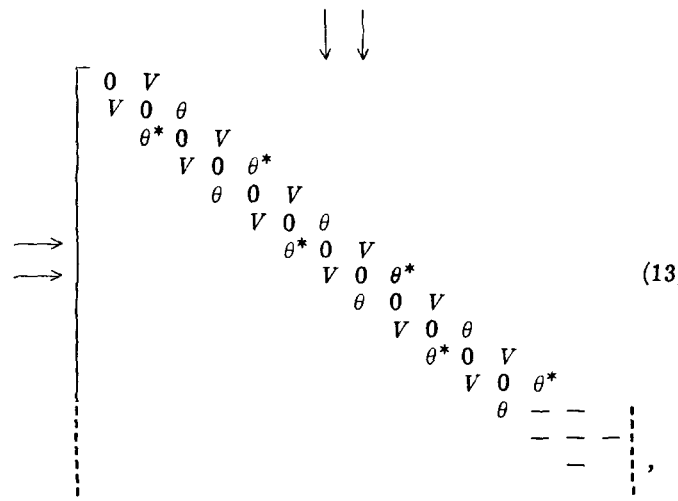
The zincblende, wurtzite series is rather more complex than the fcc, hcp series and contains two kinds of atoms that may be called a and b. We define each atom in the fcc, hcp series to be of type a. A structure in the zincblende, wurtzite series can be formed from one in the fcc, hcp series by placing an atom of type b a distance $\frac{3}{4}c$ above each a type atom in the z direction (see Fig. 1). The structure now consists of two identical interpenetrating sublattices with each atom in one sublattice having four nearest neighbors in the other sublattice at the corners of a tetrahedron.[2,4] We may visualize this structure as consisting of puckered planes, stacked vertically, and containing a type atoms from the $(n+1)$th layer with b type atoms from the nth layer a distance $\frac{1}{4}c$ below and displaced horizontally so that the nearest neighbor bonds form hexagonal chains.[2,4] Each atom has three nearest neighbors in this puckered plane and each a type atom has a nearest neighbor b in the plane above. The translational symmetry of the structure perpendicular to the z axis can again be used to define a two-dimensional $\mathbf{k}$ vector. We now have two kinds of states for the nth layer, $|\mathbf{k},n,a\rangle$ for sublattice a and $|\mathbf{k},n,b\rangle$ for sublattice b. We thus have

$$\langle \mathbf{k},n,a\,|H|\,\mathbf{k},n,b\rangle = V \qquad (11)$$

for the vertical bonds and

$$
\begin{aligned}
&\langle \mathbf{k},n,b\,|\,H\,|\mathbf{k},n+1,a\rangle \\
&\quad = \theta_{\mathbf{k}} \text{ for cyclic sequence } AB, BC, CA, \\
&\quad = \theta_{\mathbf{k}}^* \text{for anticyclic sequence } BA, CB, AC,
\end{aligned}
\qquad (12)
$$

where $\theta_{\mathbf{k}}$ is defined in (7) and the origin of coordinates in each a type layer is chosen in the way described in the previous section. If we start with the wurtzite structure which is derived from hcp and therefore stacked $ABABAB\ldots$ each of the N blocks of H characterized by a two-dimensional $\mathbf{k}$ vector looks like



(13)

where we have suppressed the $\mathbf{k}$ label on θ and the basis states for the matrix are ordered $\cdots|\mathbf{k},n,a\rangle\,|\mathbf{k},n,b\rangle\,|\mathbf{k},n+1,a\rangle\,|\mathbf{k},n+1,b\rangle\cdots$. We perform

the unitary transformation in which the two rows indicated are multiplied by θ/θ^* and the two columns are multiplied by θ^*/θ. The sequence above the diagonal becomes $V\theta\, V\theta^*\, V\theta^*\, V\theta\, V\theta\, V\theta^*$ which corresponds to $ABACABA\ldots$. This is the same change in sequencing that we achieved for the fcc, hcp series in the previous section and the argument, therefore, proceeds in an identical manner from this point to show that the density of states is the same for any desired sequence.

As a footnote to the preceding two sections, it can be seen that the diagonal elements of H^l are equal within each of the M blocks in the Hamiltonian in either of the two polytype series, provided that end affects are neglected of course. This is because in the diagonal elements of H^l, a given θ is always multiplied by its conjugate θ^*. Therefore $\langle\phi_i|H|\phi_i\rangle$ is independent of i and so the random walk can begin at any site and the number of returns will be the same. This important result allows us to remove the $(1/NM)\sum_i$ in Eq. (3). Any site is "typical" in respect to random walks as long as it is sufficiently far from the surface that the walk does not reach the surface. This result is perhaps a little surprising as all the structures considered in this paper, except fcc, contain sites not related by a vector in the translation group of the structure.

V. DISCUSSION

It is amusing to notice that the two structures considered in Secs. III and IV, which have the interesting property with respect to random walks, are the closest packed (12 neighbors) and loosest packed (4 neighbors) of the three-dimensional structures. The main interest is in the equivalence of the random walks; however, the equivalence between the density of states for the Hamiltonian (1) is also of some interest.

We have found it convenient in this work *not* to define a wave vector k_z in the direction of the random stacking. To do so is misleading even in the case of structures periodic in the z direction as the essential simplicity is obscured. As an illustration of this we consider the fcc and hcp structures in the 001 direction. Defining a wave vector k_z, the dispersion E_{001} along 001 for the hcp, where the separation between adjacent planes is $c = (\sqrt{2/3})\,a$, is given by the eigenvalues of a 2×2 matrix,

$$\begin{bmatrix} 6V & 3V(1 + e^{i2k_zc}) \\ 3V(1 + e^{i2k_zc}) & 6V \end{bmatrix}, \tag{14}$$

$$\text{i.e., } E_{001} = 6V(1 \pm \cos k_zc) \tag{15}$$

These two branches are shown in Fig. 2. The Brillouin zone has a hexagonal cross section perpendicular to the k_z axis and a height $2\pi/2c$ in the k_z direction. It is possible to describe the fcc lattice in a similar manner although it is more usual to exploit the cubic symmetry to its greatest extent and have only one atom in each unit cell, so that z axis becomes a 111 direction. However, we can define a k_z vector and the dispersion E_{001} along 001 is now given by the eigenvalues of a 3×3 matrix,

$$\begin{bmatrix} 6V & 3V & 3Ve^{3ik_zc} \\ 3V & 6V & 3V \\ 3Ve^{-3ik_zc} & 3V & 6V \end{bmatrix}, \tag{16}$$

$$E_{001} = 6V(1 + \cos k_zc) \quad \text{and} \quad 6V[1 - \cos(k_zc \pm \pi/3]. \tag{17}$$

These three branches are also shown in Fig. 2. The Brillouin zone again has a hexagonal cross section perpendicular to the k_z direction but its height is only $2\pi/3\,c$ corresponding to a repeat distance of $3c$ for the fcc structure as opposed to $2c$ for hcp. This zone is only $\frac{1}{3}$ the size of the Brillouin zone that is usually used for the fcc structure; however, every k vector has three modes associated with it instead of just one. It can be seen immediately from Fig. 2 that the energies are the same for the two cases as required by the result of Sec. 3; however, the k space description is quite different; in particular the zero energy mode is at the zone center in hcp and at the zone boundary in fcc. As the period in the z direction gets longer, the height of the Brillouin zone shrinks and the number of modes associated with each k vector increases until in the limit of no periodicity in the z direction, k_z becomes zero and there are M modes associated with each k vector in the plane perpendicular to the z axis.

In the theory of magnetic insulators, the Heisenberg Hamiltonian with nearest neighbor ferromagnetic coupling is often used.[5] Within the states of one-spin deviation, this Hamiltonian reduces to the simple hopping Hamiltonian (1) apart from a constant. Thus the density of spin waves states is the same for all structures in the fcc, hcp series and in the zincblende, wurtzite series.

It is probable that the work of this paper can be extended to more complex Hamiltonians than Eq. (1) provided that the hopping is *restricted to nearest neighbors* only. Recently, Thorpe and Weaire[6] have shown that the density of states for a model of Si or Ge containing four states per atom (when spin is not considered) in a tetrahedrally coordinated structure can be related to the density of states of the Hamiltonian (1) by a *structure independent transformation*. Thus for this model the density of states for Si and Ge in the valence and conduction bands is independent of the structure within the zincblende, wurtzite polytype

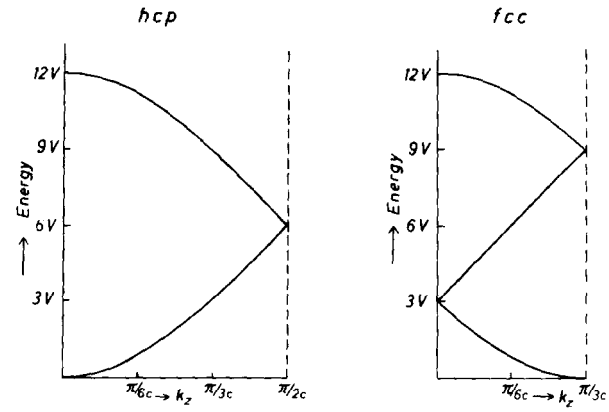


FIG. 2. The dispersion relation along the (001) direction for the hcp and fcc structures. The Brillouin zone is hexagonal in cross section perpendicular to the z axis and has a repeat distance of $2\pi/2c$ for hcp and $2\pi/3c$ for fcc along the z direction. This corresponds to the conventional zone for hcp, but the zone for fcc is $\frac{1}{3}$ of the conventional zone for that structure. The figure illustrates that the eigenvalues are the same for the two structures even though the labeling of the states in k space is quite different.

series. This leads to the import result that the band gap is also structure independent. The problem of the difference in the band structure for the zincblende and wurtzite structures when the hopping is not restricted to nearest neighbors and when two atomic species are present, has been considered by Birman.[4]

VI. RELATION BETWEEN THE TWO SERIES

In this section we derive a simple relationship between the number of returns r_l in random walks on the two polytype series considered in this paper. It is convenient to fix our attention upon the simplest member of each series, i.e., fcc and zincblende. It is not surprising that the r_l for these two lattices are related as zincblende is the fcc lattice with two atoms at each lattice point.

The fcc lattice has one atom in the unit cell and this has 12 nearest neighbors which we may put at $a(\pm1, \pm1, 0), a(\pm1, 0, \pm1), a(0, \pm1, \pm1)$. The eigenvalues $E_{\mathbf{k}}^{\mathrm{fcc}}$ of the Hamiltonian (1) are given by

$$\begin{aligned} E_{\mathbf{k}}^{\mathrm{fcc}} &= V\sum_{\delta} e^{i\mathbf{k}\cdot\delta} \\ &= 4\,V\alpha_{xyz} \end{aligned} \tag{18}$$

where

$$\alpha_{xyz} = \cos k_x a \cos k_y a + \cos k_y a \cos k_z a + \cos k_z a \cos k_x a. \tag{19}$$

The wave vector $\mathbf{k}$ is of course three-dimensional and the sum over δ is over the 12 nearest neighbors. Via (3), the r_l for the fcc lattice (denoted by r_l^{fcc}) are given by

$$\begin{aligned} r_l^{\mathrm{fcc}} &= \frac{1}{V^l N M}\sum_{\mathbf{k}} (E_{\mathbf{k}}^{\mathrm{fcc}})^l \\ &= 4^l \frac{1}{(2\pi)^3}\int_{-\pi}\int\int^{\pi} \\ &\quad \times (\cos x\cos y + \cos y\cos z + \cos z\cos x)^l dx dy dz. \end{aligned} \tag{20}$$

The integrals in (20) can easily be evaluated for small values of l using

$$\left.\begin{aligned} \frac{1}{2\pi}\int_{-\pi}^{\pi}\cos^{2n}x dx &= \frac{(2n)!}{4^n (n!)^2} \\ \frac{1}{2\pi}\int_{-\pi}^{\pi}\cos^{2n+1}x dx &= 0 \end{aligned}\right\}, \tag{21}$$

where n is an integer. The r_l^{fcc} for 1 up to 9 are given in Table I. They agree with the values given by Domb[1] for the fcc and hcp lattices.

The zincblende structure has two atoms at each lattice point of an fcc lattice. Each atom on one sublattice has four nearest neighbors on the other sublattice with may put at $a/2(1, 1, 1), a/2(1, -1, -1), a/2(-1, 1, -1)$ and $a/2(-1, -1, 1)$. The eigenvalues $E_{\mathbf{k}}^{\mathrm{zb}}$ for the zincblende structure are obtained from the 2×2 matrix

$$\begin{bmatrix} 0 & V\beta_{\mathbf{k}} \\ V\beta_{\mathbf{k}}^* & 0 \end{bmatrix},$$

where $\beta_{\mathbf{k}} = 1 + e^{i(k_x + k_y)a} + e^{i(k_y + k_z)a} + e^{i(k_z + k_x)a}$

and so

$$E_{\mathbf{k}}^{\mathrm{zb}} = \pm\, 2V\sqrt{1 + \alpha_{xyz}}, \tag{22}$$

where α_{xyz} also occurs in the dispersion relation for the fcc lattice and is given by (19). The number of returns r_l^{zb} for the zincblende are obtained from (3) and (22):

$$\begin{aligned} r_{2l}^{\mathrm{zb}} &= \frac{1}{V^{2l}NM}\sum_{\mathbf{k}} (E_{\mathbf{k}}^{\mathrm{zb}})^{2l} \\ &= 4^l \frac{1}{(2\pi)^3}\int_{-\pi}\int\int^{\pi} (1 + \cos x\cos y + \cos y\cos z \\ &\quad + \cos z\cos x)^l dx dy dz, \end{aligned} \tag{23}$$

where l is an integer. There are no returns involving walks with an odd number of steps. The summation over $\mathbf{k}$ is over both branches of the spectrum (22).

It is clear that from (20) and (23) we can derive a simple relation between r_l^{fcc} and r_l^{zb}

$$r_{2m}^{\mathrm{zb}} = \sum_{m_l=0}^{l} 4^{m-l}\binom{m}{l} r_l^{\mathrm{fcc}}. \tag{24}$$

We have used (24) to derive r_l^{zb} for l up to 18. This extends the table given by Domb[1] for the diamond structure which goes to $l = 12$. The results are shown in Table I. Notice that there are no returns consisting of an odd number of steps. This is because all the members of the zincblende, wurtzite series can be split into two interpenetrating sublattices as described in Sec. IV.

TABLE I. The number of returns to the origin r_l for the first few walks on the fcc, hcp series [calculated from Eq. (20)] and for the zincblende, wurtzite series [calculated from Eq. (24)].

fcc, hcp polytype series		zincblende, wurtzite polytype series	
l	r_l	l	r_l
0	1	0	1
1	0	2	4
2	12	4	28
3	48	6	256
4	540	8	2 716
5	4 320	10	31 504
6	42 240	12	387 136
7	403 200	14	4 951 552
8	4 038 300	16	65 218 204
9	40 958 400	18	878 536 624

The density of states for the Hamiltonian (1) for the fcc lattice has been calculated numerically by Frikkee[7] and expressed in terms of elliptic integrals by Joyce.[8] Using (18) and (22), the density of states for the zincblende structure can be found by a simple transformation on the density of states for the fcc lattice.[6]

VII. CONCLUSIONS

We have shown that the density of states for a simple hopping Hamiltonian is the same for all structures in the fcc, hcp polytype series and in the zincblende, wurtzite polytype series. This result is then used to show that the number of returns to the origin in a walk starting from *any* point and having a given number of steps is the same for all members of each of the two polytype series. A relationship is given between random walks on the two sets of structures.

Note added in proof: I would like to thank M. E. Fisher for informing me of his unpublished proof of the equivalence of random walks on the lattices of the fcc, hcp polytype structures. This rather elegant proof utilizes generating functions (see Ref. 1) and its extension to the zincblende, wurtzite polytype series is discussed briefly in the Appendix of J. F. Nagle, J. Math. Phys. 7, 1484 (1966).

ACKNOWLEDGMENTS

I should like to thank J. Birman, W. Brenig, and R. J. Elliott for critical comments and J. Nagle and D. Weaire for continuing discussions in the general area of this work. I should also like to thank H. Pick and R. Weber for their hospitality at the University of Stuttgart.

* This work was supported in part by the U.S. Atomic Energy Commission.
† Permanent address: Becton Center, Yale University, New Haven, Conn. 06520.
1 C. Domb, Advan. Phys. 9, 245 (1960).
2 A. R. Verma and P. Krishna, *Polymorphism and Polytypism in Crystals* (Wiley, New York, 1966), Chap. 4.
3 C. Domb and M. F. Sykes, Proc. Phys. Soc. London B70, 896 (1957).
4 J. L. Birman, Phys. Rev. 115, 1493 (1959).
5 See, for example, C. Kittel, *Quantum Theory of Solids* (Wiley, New York, 1963), Chap. 4.
6 M. F. Thorpe and D. Weaire, Phys. Rev. B4, 3518 (1971).
7 E. Frikkee, J. Phys. C 2, 345 (1969).
8 G. S. Joyce, J. Phys. C 4, L53, (1971).

Moments and Correlation Functions of Solutions of Some Stochastic Matrix Differential Equations

J. A. Morrison
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received 20 September 1971)

The (not necessarily linear) vector differential equation

$$\frac{d\mathbf{u}}{dz} = \mathbf{f}(\mathbf{u}(z), M(z), z), \qquad \mathbf{u}(0) = \mathbf{g}[M(0)]$$

is first considered, where $M(z)$ is a finite-state Markov process which has, in general, a nonstationary transition mechanism. The joint process $\{\mathbf{u}(z), M(z)\}$ is a Markov process, and forward and backward Kolmogorov equations are derived for the transition probability density functions. Attention is then turned to the linear matrix differential equation

$$\frac{d\mathbf{W}}{dz} = \mathbf{A}(M(z), z)\mathbf{W}(z), \qquad \mathbf{W}(0) = \boldsymbol{\gamma}[M(0)],$$

where $\mathbf{W}$ and $\boldsymbol{\gamma}$ are $n \times m$ matrices and $\mathbf{A}$ is an $n \times n$ matrix. The forward equations for the corresponding probability density functions are used to obtain two different, but equivalent, formulations for the calculation of the moments of any given order, and of the correlation functions, of the solution. The calculation of the moments and correlation functions is reduced to the solution of systems of linear ordinary differential equations, with prescribed initial conditions. The inhomogeneous matrix equation

$$\frac{d\mathbf{Y}}{dz} = A(M(z), z)\mathbf{Y}(z) + \mathbf{B}(M(z), z), \qquad \mathbf{Y}(0) = \boldsymbol{\gamma}[M(0)]$$

is also considered. Some applications, in particular to the calculation of the average modal powers in randomly coupled transmission lines, will be given elsewhere.

1. INTRODUCTION

This paper is concerned with the calculation of the moments and correlation functions of the solutions of the stochastic matrix differential equation

$$\frac{d\mathbf{W}}{dz} = \mathbf{A}(M(z), z)\mathbf{W}(z), \tag{1.1}$$

satisfying the initial condition

$$\mathbf{W}(0) = \boldsymbol{\gamma}[M(0)], \tag{1.2}$$

where $M(z)$ is a continuous parameter, finite-state Markov chain[1] which has, in general, a nonstationary transition mechanism. Here $\mathbf{W}(z)$ and $\boldsymbol{\gamma}(\cdot)$ are $n \times m$ matrices and $\mathbf{A}(\cdot, \cdot)$ is an $n \times n$ matrix valued function of its arguments. Application of the results to the calculation of the average power in each of two randomly coupled modes, traveling in the same direction in a transmission line, will be made in another paper.

The results of this paper generalize and extend those of a recent paper by McKenna and Morrison,[2] in which equations were obtained for the first- and second-order moments of the solutions of the two pairs of equations

$$\frac{du_m}{dz} = v_m, \qquad \frac{dv_m}{dz} = -\beta_0^2[1 + \eta f(M(z))]u_m, \qquad m = 1, 2, \tag{1.3}$$

satisfying the nonstochastic initial conditions

$$u_1(0) = 1 = v_2(0), \qquad u_2(0) = 0 = v_1(0). \tag{1.4}$$

The system (1.3) and (1.4) may be written in the matrix form (1.1) with

$$\mathbf{W} = \begin{bmatrix} u_1 & u_2 \\ v_1 & v_2 \end{bmatrix}. \tag{1.5}$$

In an earlier paper by McKenna and Morrison[3] the solutions of (1.3) were considered in the special case $f(M(z)) = T(z)$, where $T(z)$ is the random telegraph process,[4] and the correlation functions were calculated, in addition to the moments.

The starting point of the present analysis is the real (not necessarily linear) vector differential equation

$$\frac{d\mathbf{u}}{dz} = \mathbf{f}(\mathbf{u}(z), M(z), z), \tag{1.6}$$

subject to the initial condition

$$\mathbf{u}(0) = \mathbf{g}[M(0)]. \tag{1.7}$$

Here $\mathbf{u}(z)$ is a column vector with components $u_i(z)$, $i = 1, \ldots, L$, and $\mathbf{g}(\cdot)$ and $\mathbf{f}(\cdot, \cdot, \cdot)$ are L-vector valued functions of their arguments. The sample functions of the Markov process $M(z)$ can take on only the values a_p, $p = 1, \ldots, N$, and the paths are assumed to

ACKNOWLEDGMENTS

I should like to thank J. Birman, W. Brenig, and R. J. Elliott for critical comments and J. Nagle and D. Weaire for continuing discussions in the general area of this work. I should also like to thank H. Pick and R. Weber for their hospitality at the University of Stuttgart.

* This work was supported in part by the U.S. Atomic Energy Commission.
† Permanent address: Becton Center, Yale University, New Haven, Conn. 06520.
1 C. Domb, Advan. Phys. 9, 245 (1960).
2 A. R. Verma and P. Krishna, *Polymorphism and Polytypism in Crystals* (Wiley, New York, 1966), Chap. 4.
3 C. Domb and M. F. Sykes, Proc. Phys. Soc. London B70, 896 (1957).
4 J. L. Birman, Phys. Rev. 115, 1493 (1959).
5 See, for example, C. Kittel, *Quantum Theory of Solids* (Wiley, New York, 1963), Chap. 4.
6 M. F. Thorpe and D. Weaire, Phys. Rev. B4, 3518 (1971).
7 E. Frikkee, J. Phys. C 2, 345 (1969).
8 G. S. Joyce, J. Phys. C 4, L53, (1971).

Moments and Correlation Functions of Solutions of Some Stochastic Matrix Differential Equations

J. A. Morrison
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received 20 September 1971)

The (not necessarily linear) vector differential equation

$$\frac{d\mathbf{u}}{dz} = \mathbf{f}(\mathbf{u}(z), M(z), z), \qquad \mathbf{u}(0) = \mathbf{g}[M(0)]$$

is first considered, where $M(z)$ is a finite-state Markov process which has, in general, a nonstationary transition mechanism. The joint process $\{\mathbf{u}(z), M(z)\}$ is a Markov process, and forward and backward Kolmogorov equations are derived for the transition probability density functions. Attention is then turned to the linear matrix differential equation

$$\frac{d\mathbf{W}}{dz} = \mathbf{A}(M(z), z)\mathbf{W}(z), \qquad \mathbf{W}(0) = \boldsymbol{\gamma}[M(0)],$$

where $\mathbf{W}$ and $\boldsymbol{\gamma}$ are $n \times m$ matrices and $\mathbf{A}$ is an $n \times n$ matrix. The forward equations for the corresponding probability density functions are used to obtain two different, but equivalent, formulations for the calculation of the moments of any given order, and of the correlation functions, of the solution. The calculation of the moments and correlation functions is reduced to the solution of systems of linear ordinary differential equations, with prescribed initial conditions. The inhomogeneous matrix equation

$$\frac{d\mathbf{Y}}{dz} = A(M(z), z)\mathbf{Y}(z) + \mathbf{B}(M(z), z), \qquad \mathbf{Y}(0) = \boldsymbol{\gamma}[M(0)]$$

is also considered. Some applications, in particular to the calculation of the average modal powers in randomly coupled transmission lines, will be given elsewhere.

1. INTRODUCTION

This paper is concerned with the calculation of the moments and correlation functions of the solutions of the stochastic matrix differential equation

$$\frac{d\mathbf{W}}{dz} = \mathbf{A}(M(z), z)\mathbf{W}(z), \tag{1.1}$$

satisfying the initial condition

$$\mathbf{W}(0) = \boldsymbol{\gamma}[M(0)], \tag{1.2}$$

where $M(z)$ is a continuous parameter, finite-state Markov chain[1] which has, in general, a nonstationary transition mechanism. Here $\mathbf{W}(z)$ and $\boldsymbol{\gamma}(\cdot)$ are $n \times m$ matrices and $\mathbf{A}(\cdot, \cdot)$ is an $n \times n$ matrix valued function of its arguments. Application of the results to the calculation of the average power in each of two randomly coupled modes, traveling in the same direction in a transmission line, will be made in another paper.

The results of this paper generalize and extend those of a recent paper by McKenna and Morrison,[2] in which equations were obtained for the first- and second-order moments of the solutions of the two pairs of equations

$$\frac{du_m}{dz} = v_m, \qquad \frac{dv_m}{dz} = -\beta_0^2[1 + \eta f(M(z))]u_m, \qquad m = 1, 2, \tag{1.3}$$

satisfying the nonstochastic initial conditions

$$u_1(0) = 1 = v_2(0), \qquad u_2(0) = 0 = v_1(0). \tag{1.4}$$

The system (1.3) and (1.4) may be written in the matrix form (1.1) with

$$\mathbf{W} = \begin{bmatrix} u_1 & u_2 \\ v_1 & v_2 \end{bmatrix}. \tag{1.5}$$

In an earlier paper by McKenna and Morrison[3] the solutions of (1.3) were considered in the special case $f(M(z)) = T(z)$, where $T(z)$ is the random telegraph process,[4] and the correlation functions were calculated, in addition to the moments.

The starting point of the present analysis is the real (not necessarily linear) vector differential equation

$$\frac{d\mathbf{u}}{dz} = \mathbf{f}(\mathbf{u}(z), M(z), z), \tag{1.6}$$

subject to the initial condition

$$\mathbf{u}(0) = \mathbf{g}[M(0)]. \tag{1.7}$$

Here $\mathbf{u}(z)$ is a column vector with components $u_i(z)$, $i = 1, \ldots, L$, and $\mathbf{g}(\cdot)$ and $\mathbf{f}(\cdot, \cdot, \cdot)$ are L-vector valued functions of their arguments. The sample functions of the Markov process $M(z)$ can take on only the values a_p, $p = 1, \ldots, N$, and the paths are assumed to

be right-continuous. Applications in control theory involving equations of the form (1.6) have been considered by Wonham,[5] and earlier by Krasovskii and Lidskii,[6,7] in the case that $M(z)$ has a stationary transition mechanism, and the initial vector $\mathbf{u}(0)$ does not depend on $M(0)$. However, our interest in system (1.6) is somewhat different from theirs.

We assume that sufficient conditions are imposed on $\mathbf{f}(\cdot,\cdot,\cdot)$ to ensure, for each sample function $M(z)$ (except a set of measure zero), the existence and uniqueness of the solution $\mathbf{u}(z)$ of (1.6) and (1.7) on some interval of the half-line $0 \le z < \infty$, which contains the origin. It is easy to see that the joint process $\{\mathbf{u}(z), M(z)\}$ is a Markov process on this interval. The probability density functions $\sigma_p(\mathbf{u}, z)$, $p = 1, \ldots, N$, are defined by

$$\sigma_p(\mathbf{u}, z)\,d^L\mathbf{u} = \text{Prob}\{\mathbf{u} \le \mathbf{u}(z) \le \mathbf{u} + d\mathbf{u},\ M(z) = a_p\}, \tag{1.8}$$

where $d^L\mathbf{u}$ is the volume element

$$d^L\mathbf{u} = \prod_{i=1}^{L} du_i. \tag{1.9}$$

Here $\mathbf{v} \le \mathbf{w}$ means that the inequality holds component by component. Stochastic averages of functions of the form $F(\mathbf{u}(z), M(z), z)$ are given by

$$\langle F(\mathbf{u}(z), M(z), z)\rangle = \sum_{p=1}^{N} \int_{R^L} F(\mathbf{u}, a_p, z)\sigma_p(\mathbf{u}, z)d^L\mathbf{u}, \tag{1.10}$$

the integration being over the entire Euclidean L space, R^L.

In Sec. 2 we summarize those properties of the finite state Markov chain $M(z)$ which are needed. A formal derivation of the partial differential equations satisfied by the probability density functions $\sigma_p(\mathbf{u}, z), p = 1, \ldots, N$, is given in Appendix A.

We are also interested in the transition probability density functions $\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)$, $p, q = 1, \ldots, N$, defined for $0 \le \zeta \le z$ by

$$\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)\,d^L\mathbf{u} = \text{Prob}\{\mathbf{u} \le \mathbf{u}(z) \le \mathbf{u} + d\mathbf{u}, \quad M(z) = a_p \,|\, \mathbf{u}(\zeta) = \mathbf{v},\ M(\zeta) = a_q\}. \tag{1.11}$$

Stochastic averages of functions of the form $G(\mathbf{u}(z), M(z), z; \mathbf{u}(\zeta), M(\zeta), \zeta)$ are given by

$$\langle G(\mathbf{u}(z), M(z), z; \mathbf{u}(\zeta), M(\zeta), \zeta)\rangle = \sum_{p=1}^{N}\sum_{q=1}^{N} \int_{R^L}\int_{R^L} G(\mathbf{u}, a_p, z; \mathbf{v}, a_q, \zeta) \times \rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)\sigma_q(\mathbf{v}, \zeta)d^L\mathbf{u}\,d^L\mathbf{v}. \tag{1.12}$$

Averages of functions involving three or more points can also be calculated with the aid of the transition probability density functions, since the joint process $\{\mathbf{u}(z), M(z)\}$ is a Markov process.

The forward and backward Kolmogorov equations for the transition probability density functions are given in Sec. 2. The forward equations for $\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)$, with $q, \mathbf{v}$ and ζ fixed, are the same as those for $\sigma_p(\mathbf{u}, z)$, but the boundary conditions are, of course, different. The backward equations for $\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)$, with $p, \mathbf{u}$, and z fixed, are derived in Appendix B, and give the adjoint formulation. From the backward equations, an alternate formulation for calculating the stochastic average of $F(\mathbf{u}(z), M(z), z)$ is obtained.

In Sec. 3 the linear matrix system (1.1) and (1.2), which may be written in vector form, is considered. With the help of the equations for the corresponding probability and transition probability density functions, expressions are derived for the moments of any given order, and the correlation functions, of the solution of (1.1) and (1.2). Some of the details are given in Appendix C. The calculation of the moments and correlation functions involves the solution of systems of linear ordinary differential equations with prescribed initial conditions. In the case in which the process $M(z)$ has a stationary transition mechanism and $\mathbf{A}$ is a function of $M(z)$ alone, these equations have constant coefficients.

In Sec. 4 we give alternate, but equivalent, formulations for the calculation of the moments and correlation functions. The alternate formulations for the first- and second-order moments correspond to those obtained by McKenna and Morrison[2] for the particular system (1.3) and (1.4).

In Sec. 5 the inhomogeneous stochastic matrix differential equation

$$\frac{d\mathbf{Y}}{dz} = \mathbf{A}(M(z), z)\mathbf{Y}(z) + \mathbf{B}(M(z), z), \quad \mathbf{Y}(0) = \boldsymbol{\gamma}[M(0)], \tag{1.13}$$

is considered. Here $\mathbf{Y}$, $\mathbf{B}$ and $\boldsymbol{\gamma}$ are $n \times m$ matrices, and $\mathbf{A}$ is as before. The system (1.13) may be rewritten as an augmented homogeneous system, so that the previous results are applicable. The moments of order s of the solution of (1.13) are coupled to those of order $(s-1)$, for $s = 1, 2, \cdots$. Thus the moments may be calculated successively for increasing order.

2. EQUATIONS FOR THE PROBABILITY DENSITY FUNCTIONS

We first summarize those properties of the finite state Markov chain $M(z)$ which we will need.[1] The sample functions $M(z)$ are defined on the half-line $0 \le z < \infty$, can take on only a finite number N of distinct values $a_p, p = 1, \ldots, N$, and have right-continuous paths. An initial probability distribution is given,

$$\alpha_p = \text{Prob}\{M(0) = a_p\}, \quad p = 1, \ldots, N, \tag{2.1}$$

where $\alpha_p > 0$ and

$$\sum_{p=1}^{N} \alpha_p = 1. \tag{2.2}$$

We consider only those processes which can be defined by means of continuous, bounded infinitesimal generators. Thus we assume given an $N \times N$ matrix function

$$\tau(z) = (\tau_{pq}(z)), \tag{2.3}$$

satisfying the conditions

$$\tau_{pq}(z) \ge 0, \quad p \ne q, \quad \tau_{pp}(z) \le 0, \quad p, q = 1, \ldots, N, \tag{2.4}$$

and

$$\sum_{q=1}^{N} \tau_{pq}(z) = 0, \quad p = 1, \ldots, N. \tag{2.5}$$

Definition 2.1. $E_p^{(n)}(x, y)$ is the event that $M(y) = a_p$ and $M(z)$ changes value n times in the interval (x, y).

The events $E_p^{(n)}(x,y)$ and $E_p^{(m)}(x,y)$ are clearly mutually exclusive for $m \neq n$ and $\cup_{m=0}^{\infty} E_p^{(m)}(x,y)$ is just the event $M(y) = a_p$. Define

$$P_{qp}^{(n)}(x,y) = \text{Prob}\{E_p^{(n)}(x,y) \mid M(x) = a_q\}. \tag{2.6}$$

Then, for $\delta z \to 0+$, $p, q = 1, \ldots, N$,

$$P_{pp}^{(0)}(z, z+\delta z) = 1 + \tau_{pp}(z)\delta z + o(\delta z), \tag{2.7}$$

$$P_{qp}^{(1)}(z, z+\delta z) = \tau_{qp}(z)\delta z + o(\delta z), \quad q \neq p, \tag{2.8}$$

$$\sum_{n=2}^{\infty} P_{qp}^{(n)}(z, z+\delta z) = o(\delta z). \tag{2.9}$$

If the matrix τ is constant, then the process $M(z)$ is said to have a stationary transition mechanism.

In Appendix A we derive the equations satisfied by the probability density functions $\sigma_p(\mathbf{u}, z)$, $p = 1, \ldots, N$, defined by (1.8), where $\mathbf{u}(z)$ is the solution of (1.6) and (1.7). It is found that

$$\frac{\partial \sigma_p}{\partial z} + \sum_{i=1}^{L} \frac{\partial}{\partial u_i}[f_i(\mathbf{u}, a_p, z)\sigma_p(\mathbf{u}, z)] - \sum_{r=1}^{N} \tau_{rp}(z)\sigma_r(\mathbf{u}, z) = 0. \tag{2.10}$$

From (1.7), (1.8), and (2.11), the initial conditions are

$$\sigma_p(\mathbf{u}, 0) = \alpha_p\, \delta[\mathbf{u} - \mathbf{g}(a_p)], \quad p = 1, \ldots, N, \tag{2.11}$$

where δ denotes the delta function.

Now consider the transition probability density functions $\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)$, $p, q = 1, \ldots, N$, defined by (1.11). For $z > \zeta$ and fixed q, $\mathbf{v}$ and ζ, we may derive equations for $\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)$ in a manner analogous to that used in Appendix A in obtaining the equations for $\rho_p(\mathbf{u}, z)$. Since the joint process $\{\mathbf{u}(z), M(z)\}$ is a Markov process, it follows that the probabilities of events at $z + \delta z$, conditioned on $(\mathbf{u}(z), M(z))$ and $(\mathbf{u}(\zeta), M(\zeta))$, are just the probabilities conditioned on $(\mathbf{u}(z), M(z))$ only.

Consequently, for fixed q, $\mathbf{v}$, and ζ, it is found that $\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)$ satisfies Eq. (2.10), that is,

$$\frac{\partial \rho_{pq}}{\partial z} + \sum_{i=1}^{L} \frac{\partial}{\partial u_i}[f_i(\mathbf{u}, a_p, z)\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)] - \sum_{r=1}^{N} \tau_{rp}(z)\rho_{rq}(\mathbf{u}, z; \mathbf{v}, \zeta) = 0, \quad z > \zeta \geq 0, \tag{2.12}$$

$p, q = 1, \ldots, N$. However, from (1.11), the initial conditions are

$$\rho_{pq}(\mathbf{u}, \zeta; \mathbf{v}, \zeta) = \delta_{pq}\, \delta(\mathbf{u} - \mathbf{v}), \quad p, q = 1, \ldots, N, \tag{2.13}$$

where δ_{pq} is a Kronecker delta. It follows from (2.10)–(2.13), or from (1.7), (1.8) and (1.11), that

$$\sigma_p(\mathbf{u}, z) = \sum_{q=1}^{N} \alpha_q \rho_{pq}(\mathbf{u}, z; \mathbf{g}(a_q), 0). \tag{2.14}$$

Equations (2.12) are the forward equations for the transition probability density functions $\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)$. In Appendix B we derive the backward equations, in which p, $\mathbf{u}$, and z are held fixed. It is shown that

$$\frac{\partial \rho_{pq}}{\partial \zeta} + \sum_{i=1}^{L} f_i(\mathbf{v}, a_q, \zeta) \frac{\partial \rho_{pq}}{\partial v_i}(\mathbf{u}, z; \mathbf{v}, \zeta) + \sum_{r=1}^{N} \tau_{qr}(\zeta)\rho_{pr}(\mathbf{u}, z; \mathbf{v}, \zeta) = 0, \quad 0 \leq \zeta < z, \tag{2.15}$$

$p, q = 1, \ldots, N$. From (1.11), the boundary condition is

$$\rho_{pq}(\mathbf{u}, z; \mathbf{v}, z) = \delta_{pq}\, \delta(\mathbf{u} - \mathbf{v}), \quad p, q = 1, \ldots, N. \tag{2.16}$$

The formulation in (2.15) and (2.16) is the adjoint of that in (2.12) and (2.13). We remark that, in the case in which the process $M(z)$ has a stationary transition mechanism, Wonham[5] gives the infinitesimal generator for the joint process $\{\mathbf{u}(z), M(z)\}$, which was derived earlier by Krasovskii and Lidskii.[6,7] The backward equation (2.15) may be derived by means of this generator.

Now, from (1.10) and (2.14),

$$\langle F(\mathbf{u}(z), M(z), z)\rangle = \sum_{p=1}^{N} \sum_{q=1}^{N} \alpha_q \times \int_{R^L} F(\mathbf{u}, a_p, z)\rho_{pq}(\mathbf{u}, z; \mathbf{g}(a_q), 0)\, d^L\mathbf{u}. \tag{2.17}$$

Let

$$\mathcal{F}_q(z; \mathbf{v}, \zeta) = \sum_{p=1}^{N} \int_{R^L} F(\mathbf{u}, a_p, z)\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \zeta)\, d^L\mathbf{u}. \tag{2.18}$$

Then, from (2.15) and (2.16),

$$\frac{\partial \mathcal{F}_q}{\partial \zeta} + \sum_{i=1}^{L} f_i(\mathbf{v}, a_q, \zeta) \frac{\partial \mathcal{F}_q}{\partial v_i} + \sum_{r=1}^{N} \tau_{qr}(\zeta)\mathcal{F}_r = 0, \quad 0 \leq \zeta < z, \tag{2.19}$$

with boundary condition

$$\mathcal{F}_q(z; \mathbf{v}, z) = F(\mathbf{v}, a_q, z), \quad q = 1, \ldots, N. \tag{2.20}$$

Also, from (2.17) and (2.18)

$$\langle F(\mathbf{u}(z), M(z), z)\rangle = \sum_{q=1}^{N} \alpha_q \mathcal{F}_q(z; \mathbf{g}(a_q), 0). \tag{2.21}$$

In the next section we use the forward equations to calculate the moments and correlation functions of the solutions of the system (1.1) and (1.2). The moments may also be calculated from the backward equations, but we omit the details. We have given the backward formulation (2.19)–(2.21) as a matter of completeness, since it is often more convenient for calculating stochastic averages other than moments.

3. MOMENTS AND CORRELATION FUNCTIONS

We now consider the system (1.1) and (1.2), and use the results of the previous section to calculate the moments, and correlation functions, of the solution. Let the $n \times m$ matrices $\mathbf{W}$ and $\boldsymbol{\gamma}$, and the $n \times n$ matrix $\mathbf{A}$, have components

$$\mathbf{W} = (w_{ik}), \quad \boldsymbol{\gamma} = (\gamma_{ik}), \quad \mathbf{A} = (A_{ij}). \tag{3.1}$$

Then

$$\frac{dw_{ik}}{dz} = \sum_{j=1}^{n} A_{ij}(M(z), z)w_{jk}(z), \quad w_{ik}(0) = \gamma_{ik}[M(0)], \tag{3.2}$$

$i = 1, \ldots, n$, $k = 1, \ldots, m$. The system (3.2) may be written in vector form, by forming a column vector with nm components from the m successive column vectors $(w_{i1}), \ldots, (w_{im})$, with n components each. This leads to a system of the form (1.6) and (1.7), with $L = nm$.

The probability density functions $\sigma_p(\mathbf{W}, z)$, $p = 1, \ldots, N$, corresponding to the system (3.2), satisfy, from (2.10), the equations

$$\frac{\partial \sigma_p}{\partial z} + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} A_{ij}(a_p, z)\frac{\partial (w_{jk}\sigma_p)}{\partial w_{ik}} - \sum_{r=1}^{N} \tau_{rp}(z)\sigma_r = 0, \tag{3.3}$$

with initial conditions, from (2.11),

$$\sigma_p(\mathbf{W}, 0) = \alpha_p \delta[\mathbf{W} - \boldsymbol{\gamma}(a_p)]. \tag{3.4}$$

Since equation (3.3) is homogeneous in the elements of $\mathbf{W}$, it is possible to obtain equations for the moments of any given order. Some of the details are given in Appendix C, and we summarize the results here. We denote by $\mathbf{W}^{(s)}$ the s-fold Kronecker product of the matrix $\mathbf{W}$ with itself.[8] Then the sth-order moments of σ_p are given by

$$\langle \mathbf{W}(z)^{(s)} \rangle_p = \int_{R^{nm}} \mathbf{W}^{(s)} \sigma_p(\mathbf{W}, z) d^{nm}\mathbf{W}, \tag{3.5}$$

where

$$d^{nm}\mathbf{W} = \prod_{i=1}^{n}\prod_{k=1}^{m} dw_{ik}.$$

From (1.10), the expected value of $\mathbf{W}^{(s)}$ is given by

$$\langle \mathbf{W}(z)^{(s)} \rangle = \sum_{p=1}^{N} \langle \mathbf{W}(z)^{(s)} \rangle_p. \tag{3.6}$$

For the first-order moments it is found that

$$\frac{d}{dz}\langle \mathbf{W}(z) \rangle_p = \mathbf{A}(a_p, z)\langle \mathbf{W}(z) \rangle_p + \sum_{r=1}^{N} \tau_{rp}(z)\langle \mathbf{W}(z) \rangle_r, \tag{3.7}$$

with initial conditions, from (3.4),

$$\langle \mathbf{W}(0) \rangle_p = \alpha_p \boldsymbol{\gamma}(a_p), \quad p = 1, \ldots, N. \tag{3.8}$$

For the second-order moments it is found that

$$\frac{d}{dz}\langle \mathbf{W}(z) \times \mathbf{W}(z) \rangle_p = \sum_{r=1}^{N} \tau_{rp}(z)\langle \mathbf{W}(z) \times \mathbf{W}(z) \rangle_r + \{[\mathbf{A}(a_p, z) \times \mathbf{I}_n] + [\mathbf{I}_n \times \mathbf{A}(a_p, z)]\}\langle \mathbf{W}(z) \times \mathbf{W}(z) \rangle_p, \tag{3.9}$$

with initial conditions, from (3.4) and (3.5),

$$\langle \mathbf{W}(0) \times \mathbf{W}(0) \rangle_p = \alpha_p[\boldsymbol{\gamma}(a_p) \times \boldsymbol{\gamma}(a_p)], \quad p = 1, \ldots, N. \tag{3.10}$$

Here $\times$ denotes Kronecker product, $[\mathbf{B} \times \mathbf{D} = (b_{ij}) \times \mathbf{D} = (b_{ij}\mathbf{D})]$, and $\mathbf{I}_n$ denotes the unit matrix of order n.

Actually, we could have obtained (3.9) and (3.10) by first deriving the equation satisfied by $\mathbf{W}(z) \times \mathbf{W}(z)$ and then applying the general result for first-order moments. Thus, from (1.1),

$$\frac{d}{dz}[\mathbf{W}(z) \times \mathbf{W}(z)] = \{[\mathbf{A}(M(z), z) \times \mathbf{I}_n] + [\mathbf{I}_n \times \mathbf{A}(M(z), z)]\}[\mathbf{W}(z) \times \mathbf{W}(z)] \tag{3.11}$$

with initial condition, from (1.2),

$$\mathbf{W}(0) \times \mathbf{W}(0) = \boldsymbol{\gamma}[M(0)] \times \boldsymbol{\gamma}[M(0)]. \tag{3.12}$$

In this manner, or via the method of Appendix C, it is found that, generally,

$$\frac{d}{dz}\langle \mathbf{W}(z)^{(s)} \rangle_p = \sum_{r=1}^{N} \tau_{rp}(z)\langle \mathbf{W}(z)^{(s)} \rangle_r + \sum_{l=1}^{s} [\mathbf{I}_n^{(l-1)} \times \mathbf{A}(a_p, z) \times \mathbf{I}_n^{(s-l)}]\langle \mathbf{W}(z)^{(s)} \rangle_p, \tag{3.13}$$

where $\mathbf{I}_n^{(0)} = \mathbf{I}_1$, with initial conditions

$$\langle \mathbf{W}(0)^{(s)} \rangle_p = \alpha_p[\boldsymbol{\gamma}(a_p)]^{(s)}, \quad p = 1, \ldots, N. \tag{3.14}$$

The expectation of $\mathbf{W}^{(s)}$ may thus be calculated by solving the system of linear matrix differential equations (3.13), subject to the initial conditions (3.14), and then using (3.6).

Next, let us consider the calculation of the correlation functions. Corresponding to (3.3), the forward equations satisfied by the transition probability density functions $\rho_{pq}(\mathbf{W}, z; \mathbf{V}, \zeta)$, $p, q = 1, \ldots, N$, are, from (2.12), for $z > \zeta \geq 0$,

$$\frac{\partial \rho_{pq}}{\partial z} + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} A_{ij}(a_p, z)\frac{\partial (w_{jk}\rho_{pq})}{\partial w_{ik}} - \sum_{r=1}^{N} \tau_{rp}(z)\rho_{rq} = 0, \tag{3.15}$$

with initial conditions, from (2.13),

$$\rho_{pq}(\mathbf{W}, \zeta; \mathbf{V}, \zeta) = \delta_{pq}\delta(\mathbf{W} - \mathbf{V}). \tag{3.16}$$

Expressions are derived in Appendix C for the correlation functions. Let the matrices $\boldsymbol{\Psi}_{pq}(z, \zeta)$ satisfy the equations

$$\frac{\partial \boldsymbol{\Psi}_{pq}}{\partial z} = \mathbf{A}(a_p, z)\boldsymbol{\Psi}_{pq} + \sum_{r=1}^{N} \tau_{rp}(z)\boldsymbol{\Psi}_{rq}, \tag{3.17}$$

with initial conditions

$$\boldsymbol{\Psi}_{pq}(\zeta, \zeta) = \delta_{pq}\mathbf{I}_n, \quad p, q = 1, \ldots, N. \tag{3.18}$$

Note, from (3.7) and (3.8), that

$$\langle \mathbf{W}(z) \rangle_p = \sum_{q=1}^{N} \alpha_q \boldsymbol{\Psi}_{pq}(z, 0)\boldsymbol{\gamma}(a_q). \tag{3.19}$$

It is found that, for $0 \leq \zeta \leq z$,

$$\langle \mathbf{W}(z) \times \mathbf{W}(\zeta) \rangle = \sum_{q=1}^{N}\left(\sum_{p=1}^{N} \boldsymbol{\Psi}_{pq}(z, \zeta) \times \mathbf{I}_n\right) \times \langle \mathbf{W}(\zeta) \times \mathbf{W}(\zeta) \rangle_q. \tag{3.20}$$

This result may be expressed in augmented matrix form. Thus, let $\mathbf{K}(\zeta)$ be the column of matrices

$$\mathbf{K}(\zeta) = \mathrm{col}(\langle \mathbf{W}(\zeta) \times \mathbf{W}(\zeta) \rangle_1, \ldots, \langle \mathbf{W}(\zeta) \times \mathbf{W}(\zeta) \rangle_N), \tag{3.21}$$

and also let the matrix $\boldsymbol{\Phi}(z, \zeta)$ have matrix elements $\boldsymbol{\Psi}_{pq}(z, \zeta)$, with $\boldsymbol{\Phi} = (\boldsymbol{\Psi}_{pq})$. Note that matrix equations may be written down for $\mathbf{K}(z)$ and $\boldsymbol{\Phi}(z, \zeta)$, from (3.9) and (3.17). Also, from (3.18), $\boldsymbol{\Phi}(\zeta, \zeta) = (\mathbf{I}_N \times \mathbf{I}_n)$, so that $\boldsymbol{\Phi}$ is a fundamental matrix. From (3.20) and (3.21),

$$\langle \mathbf{W}(z) \times \mathbf{W}(\zeta) \rangle = \{[\mathbf{E}_N \times \mathbf{I}_n)\boldsymbol{\Phi}(z, \zeta)] \times \mathbf{I}_n\}\mathbf{K}(\zeta), \quad 0 \leq \zeta \leq z, \tag{3.22}$$

where $\mathbf{E}_N$ is the row vector with all N elements equal to 1.

It is remarked that if the process $M(z)$ has a stationary transition mechanism, so that τ is constant, and if $\mathbf{A}$ is a function of $M(z)$ alone, then (3.17) is a system of linear equations with constant coefficients and, from (3.18),

$$\boldsymbol{\Psi}_{pq}(z, \zeta) = \boldsymbol{\Psi}_{pq}(z - \zeta, 0), \quad p, q = 1, \ldots, N. \tag{3.23}$$

A result equivalent to (3.20) was obtained by McKenna and Morrison[3] for the special case of the system (1.3) and (1.4), with $f(M(z)), = T(z)$, where $T(z)$ is the random telegraph process, which has a stationary transition mechanism and only two possible states.

In the next section we give alternate, but equivalent, formulations for the calculation of the moments and correlation functions. Which of the two formulations is preferable depends, in particular, on the relative sizes of n and N and the sparseness of the matrices involved.

4. ALTERNATE FORMULATION

We begin by giving the alternate formulation for the first-order moments. Let the vector $\omega_h(z)$ denote the hth row of the matrix $\mathbf{W}(z)$, that is,

$$\omega_h(z) = (w_{h1}(z), \ldots, w_{hm}(z)), \qquad h = 1, \ldots, n, \quad (4.1)$$

and let $\Omega_h(z)$ denote the column of matrices

$$\Omega_h(z) = \mathrm{col}(\langle\omega_h(z)\rangle_1, \ldots, \langle\omega_h(z)\rangle_N). \quad (4.2)$$

Then, from (3.6), the expected value of ω_h is

$$\langle\omega_h(z)\rangle = \mathbf{E}_N\Omega_h(z). \quad (4.3)$$

We define the $N \times N$ diagonal matrices $\mathbf{D}_{ij}(z)$ and ν_{jk}, $i, j = 1, \ldots, n$, $k = 1, \ldots, m$, by

$$\mathbf{D}_{ij}(z) = \mathrm{diag}[A_{ij}(a_p, z)], \quad (4.4)$$

where p ranges from 1 to N along the diagonal and

$$\nu_{jk} = \mathrm{diag}[\gamma_{jk}(a_p)]. \quad (4.5)$$

Then it follows from (C1) that

$$\frac{d\Omega_h}{dz} = \sum_{j=1}^{n} \mathbf{D}_{hj}(z)\Omega_j(z) + \tau^t(z)\Omega_h(z), \quad (4.6)$$

where t denotes transpose. Also, from (3.8),

$$\Omega_h(0) = (\nu_{h1}\alpha^t, \ldots, \nu_{hm}\alpha^t), \qquad h = 1, \ldots, n, \quad (4.7)$$

where

$$\alpha = (\alpha_1, \ldots, \alpha_N) \quad (4.8)$$

is the row vector of initial probabilities given by (2.1). Note, from (2.2), that $\mathbf{E}_N\alpha^t = 1$.

McKenna and Morrison[2] obtained equations for the first-order moments of the solutions of the system (1.3) and (1.4) in a form corresponding to (4.3), (4.6), and (4.7). Thus, for this system, corresponding to (1.5), $n = 2, m = 2$, and

$$\mathbf{D}_{11} = 0 = \mathbf{D}_{22}, \qquad \mathbf{D}_{12} = \mathbf{I}_N, \qquad \mathbf{D}_{21} = -\mathbf{B}, \quad (4.9)$$

where

$$\mathbf{B} = \beta_0^2\mathbf{I}_N + \eta\beta_0^2\,\mathrm{diag}[f(a_p)]. \quad (4.10)$$

Also, from (1.4),

$$\nu_{jk} = \delta_{jk}\mathbf{I}_2. \quad (4.11)$$

Making the identifications

$$\Omega_1 = (\mathbf{U}_1, \mathbf{U}_2), \qquad \Omega_2 = (\mathbf{V}_1, \mathbf{V}_2), \quad (4.12)$$

we obtain the previous results.[2]

We consider next the second-order moments, and define the column of matrices

$$\Omega_{jh}(z) = \mathrm{col}(\langle\omega_j(z) \times \omega_h(z)\rangle_1, \ldots, \langle\omega_j(z) \times \omega_h(z)\rangle_N). \quad (4.13)$$

Then, from (4.1), (4.4), and (C3),

$$\frac{d\Omega_{fh}}{dz} = \tau^t(z)\Omega_{fh}(z) + \sum_{j=1}^{n} [\mathbf{D}_{fj}(z)\Omega_{jh}(z) + \mathbf{D}_{hj}(z)\Omega_{fj}(z)]. \quad (4.14)$$

We define the $N \times N$ diagonal matrices $\xi_{fghl}, f, h = 1, \ldots, n$, $g, l = 1, \ldots, m$, by

$$\xi_{fghl} = \mathrm{diag}[\gamma_{fg}(a_p)\gamma_{hl}(a_p)]. \quad (4.15)$$

Then, from (3.10), the initial conditions are

$$\Omega_{fh}(0) = (\xi_{f1h1}\alpha^t, \ldots, \xi_{f1hm}\alpha^t, \ldots, \xi_{fmh1}\alpha^t, \ldots, \xi_{fmhm}\alpha^t). \quad (4.16)$$

Finally, from (3.6), the expected value of $(\omega_f \times \omega_h)$ is

$$\langle\omega_f(z) \times \omega_h(z)\rangle = \mathbf{E}_N\Omega_{fh}(z). \quad (4.17)$$

For the system (1.3) and (1.4), we obtain the results of McKenna and Morrison[2] for the second-order moments by making the identifications

$$\Omega_{11} = (\mathbf{X}_1, \mathbf{X}_0, \mathbf{X}_0, \mathbf{X}_2), \qquad \Omega_{22} = (\mathbf{Z}_1, \mathbf{Z}_0, \mathbf{Z}_0, \mathbf{Z}_2),$$
$$\tfrac{1}{2}(\Omega_{12} + \Omega_{21}) = (\mathbf{Y}_1, \mathbf{Y}_0, \mathbf{Y}_0, \mathbf{Y}_2). \quad (4.18)$$

From (4.9) and (4.14) it also follows that

$$\frac{d}{dz}(\Omega_{12} - \Omega_{21}) = \tau^t(z)[\Omega_{12}(z) - \Omega_{21}(z)] \quad (4.19)$$

Since $\mathbf{E}_N\tau^t(z) \equiv 0$, from (2.5), this implies that

$$\mathbf{E}_N[\Omega_{12}(z) - \Omega_{21}(z)] = (0, 1, -1, 0), \quad (4.20)$$

using the initial values in (1.4). This result is a consequence of the identity

$$u_1(z)v_2(z) - u_2(z)v_1(z) \equiv 1. \quad (4.21)$$

Returning to the general case, we give the alternate form of the equations from which the sth-order moments may be calculated. Thus, define the column of matrices

$$\Omega_{h_1\ldots h_s}(z) = \mathrm{col}(\langle\omega_{h_1}(z) \times \cdots \times \omega_{h_s}(z)\rangle_1, \ldots, \langle\omega_{h1}(z) \times \cdots \times \omega_{h_s}(z)\rangle_N). \quad (4.22)$$

It may be shown that

$$\frac{d}{dz}\Omega_{h_1\ldots h_s} = \tau^t(z)\Omega_{h_1\ldots h_s}(z) + \sum_{j=1}^{n} [\mathbf{D}_{h_1 j}(z)\Omega_{j\ldots h_s}(z) + \cdots + \mathbf{D}_{h_s j}(z)\Omega_{h_1\ldots j}(z)], \quad (4.23)$$

with appropriate initial conditions obtained from (3.14). Also from (3.6),

$$\langle\omega_{h_1}(z) \times \cdots \times \omega_{h_s}(z)\rangle = \mathbf{E}_N\Omega_{h_1\ldots h_s}(z). \quad (4.24)$$

Lastly, we consider the alternate formulation for the calculation of the correlation functions. Let the matrices $\theta_{hk}(z, \zeta)$ satisfy the equations

$$\frac{\partial\theta_{hk}}{\partial z} = \sum_{j=1}^{n} \mathbf{D}_{hj}(z)\theta_{jk} + \tau^t(z)\theta_{hk}, \quad (4.25)$$

with initial conditions

$$\theta_{hk}(\xi, \zeta) = \delta_{hk}\mathbf{I}_N, \quad h, k = 1, \ldots, n. \tag{4.26}$$

Note, from (4.6), that

$$\Omega_h(z) = \sum_{k=1}^{n} \theta_{hk}(z, 0)\Omega_k(0). \tag{4.27}$$

It is shown in Appendix D that

$$\langle \omega_f(\xi) \times \omega_h(z) \rangle = \sum_{k=1}^{n} \mathbf{E}_N \theta_{hk}(z, \xi)\Omega_{fk}(\xi), \quad 0 \le \xi \le z. \tag{4.28}$$

This result may be expressed in augmented matrix form. Thus, let $\mathbf{H}(\xi)$ be the column of matrices

$$\mathbf{H}(\xi) = \mathrm{col}(\Omega_{11}(\xi), \ldots, \Omega_{1n}(\xi), \ldots, \Omega_{n1}(\xi), \ldots, \Omega_{nn}(\xi)), \tag{4.29}$$

and let the matrix $\boldsymbol{\Theta}(z, \xi)$ have the matrix elements $\theta_{hk}(z, \xi)$, with $\boldsymbol{\Theta} = (\theta_{hk})$. Matrix equations may be written down for $\mathbf{H}(z)$ and $\boldsymbol{\Theta}(z, \xi)$, from (4.14) and (4.25). From (4.26), $\boldsymbol{\Theta}(\xi, \xi) = (\mathbf{I}_n \times \mathbf{I}_N)$, so that $\boldsymbol{\Theta}$ is a fundamental matrix. Also, from (4.16),

$$\mathbf{H}(0) = \Xi(\mathbf{I}_m \times \mathbf{I}_m \times \alpha^t), \tag{4.30}$$

where Ξ is obtained by replacing each element $\gamma_{fg}\gamma_{hl}$ of $\boldsymbol{\gamma} \times \boldsymbol{\gamma}$ by the matrix ξ_{fghl}, given by (4.15). From (4.28) and (4.29),

$$\langle \mathbf{W}(\xi) \times \mathbf{W}(z) \rangle = \{\mathbf{I}_n \times [(\mathbf{I}_n \times \mathbf{E}_N)\boldsymbol{\Theta}(z, \xi)]\}\mathbf{H}(\xi), \quad 0 \le \xi \le z. \tag{4.31}$$

Note the interchange of z and ξ between the left-hand sides of (3.22) and (4.31).

5. INHOMOGENEOUS LINEAR EQUATIONS

In this final section we consider the inhomogeneous linear stochastic matrix differential equation

$$\frac{d\mathbf{Y}}{dz} = \mathbf{A}(M(z), z)\mathbf{Y}(z) + \mathbf{B}(M(z), z), \tag{5.1}$$

with initial condition

$$\mathbf{Y}(0) = \boldsymbol{\gamma}[M(0)]. \tag{5.2}$$

Here $\mathbf{Y}$, $\mathbf{B}$, and $\boldsymbol{\gamma}$ are $n \times m$ matrices, and $\mathbf{A}$ is an $n \times n$ matrix. The system (5.1) and (5.2) may be written in the equivalent homogeneous form

$$\frac{d}{dz}\begin{bmatrix}\mathbf{Y}\\ \mathbf{Z}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{Y}\\ \mathbf{Z}\end{bmatrix}, \quad \begin{bmatrix}\mathbf{Y}(0)\\ \mathbf{Z}(0)\end{bmatrix} = \begin{bmatrix}\boldsymbol{\gamma}[M(0)]\\ \mathbf{I}_m\end{bmatrix}, \tag{5.3}$$

which implies, in particular, that

$$\mathbf{Z}(z) \equiv \mathbf{I}_m. \tag{5.4}$$

Consequently, the results of the previous sections are immediately applicable.

Let us consider the first-order moments. Thus, from (3.7) and (3.8),

$$\frac{d}{dz}\langle \mathbf{Y}(z) \rangle_p = \sum_{r=1}^{N} \tau_{rp}(z)\langle \mathbf{Y}(z) \rangle_r + \mathbf{A}(a_p, z)\langle \mathbf{Y}(z) \rangle_p + \mathbf{B}(a_p, z)\langle \mathbf{Z}(z) \rangle_p, \tag{5.5}$$

with initial condition

$$\langle \mathbf{Y}(0) \rangle_p = \alpha_p \boldsymbol{\gamma}(a_p), \quad p = 1, \ldots, N, \tag{5.6}$$

$$\frac{d}{dz}\langle \mathbf{Z}(z) \rangle_p = \sum_{r=1}^{N} \tau_{rp}(z)\langle \mathbf{Z}(z) \rangle_r, \tag{5.7}$$

with initial condition

$$\langle \mathbf{Z}(0) \rangle_p = \alpha_p \mathbf{I}_m, \quad p = 1, \ldots, N. \tag{5.8}$$

It is evident, from (1.8), (3.5), and (5.4), that

$$\langle \mathbf{Z}(z) \rangle_p = \mathrm{Prob}\{M(z) = a_p\}\mathbf{I}_m. \tag{5.9}$$

These results may also, of course, be obtained by considering the probability density functions $\sigma_p(\mathbf{Y}, z)$, $p = 1, \ldots, N$, corresponding to the system (5.1) and (5.2). Thus, corresponding to (2.10) and (2.11), with $\mathbf{Y} = (y_{ik})$,

$$\frac{\partial \sigma_p}{\partial z} + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} A_{ij}(a_p, z)\frac{\partial (y_{jk}\sigma_p)}{\partial y_{ik}} + \sum_{i=1}^{n}\sum_{k=1}^{m} B_{ik}(a_p, z)\frac{\partial \sigma_p}{\partial y_{ik}} - \sum_{r=1}^{N} \tau_{rp}(z)\sigma_r = 0, \tag{5.10}$$

with initial conditions

$$\sigma_p(\mathbf{Y}, 0) = \alpha_p \delta[\mathbf{Y} - \boldsymbol{\gamma}(a_p)], \quad p = 1, \ldots, N. \tag{5.11}$$

It is evident that the moments of order s are coupled to those of order $(s - 1)$, for $s = 1, 2, \cdots$. The zero-order moments correspond to $\mathrm{Prob}\{M(z) = a_p\}$. Hence the moments may be calculated successively for increasing order.

ACKNOWLEDGMENTS

The author is grateful to his colleagues J. McKenna, L. A. Shepp, and H. S. Witsenhausen for some helpful discussions.

APPENDIX A

We here derive the equations satisfied by the probability density functions $\sigma_p(\mathbf{u}, z), p = 1, \ldots, N$, defined by (1.8), where $\mathbf{u}(z)$ is the solution of (1.6) and (1.7). We first remark that $\sigma_p(\mathbf{u}, z + \delta z)\delta V$ is, approximately, the probability that $M(z + \delta z) = a_p$ and $\mathbf{u}(z + \delta z)$ lies in a small volume element δV around the point $\mathbf{u}$ in R^L. We will calculate $\sigma_p(\mathbf{u}, z + \delta z)\delta V$ in terms of the probabilities of the events at z which can lead to the desired event at $z + \delta z$. From Definition 2.1, the event $M(z + \delta z) = a_p$ is the union of the three mutually exclusive events $E_p^{(0)}(z, z + \delta z)$, $E_p^{(1)}(z, z + \delta z)$, and $\cup_{n=2}^{\infty} E_p^{(n)}(z, z + \delta z)$.

Since the event $E_p^{(0)}(z, z + \delta z)$ implies that $M(x) = a_p$, for $z \le x \le z + \delta z$, then, from (1.1), it also implies that

$$\mathbf{u}(z) = \mathbf{u}(z + \delta z) - \mathbf{f}(\mathbf{u}(z + \delta z), a_p, z)\delta z + o(\delta z). \tag{A1}$$

Under this transformation, the small volume element δV around $\mathbf{u}(z + \delta z) \equiv \mathbf{u}$ is transformed into a small volume element δV_p around $\mathbf{u}(z) = [\mathbf{u} - \mathbf{f}(\mathbf{u}, a_p, z)\delta z + o(\delta z)]$, and

$$\delta V_p = \frac{\partial(u_1(z), \ldots, u_L(z))}{\partial(u_1(z + \delta z), \ldots, u_L(z + \delta z))}\,\delta V = \left(1 - \sum_{i=1}^{L}\frac{\partial f_i}{\partial u_i}(\mathbf{u}, a_p, z)\delta z + o(\delta z)\right)\delta V, \tag{A2}$$

from (A1). Next, the event $E_p^{(1)}(z, z + \delta z)$ implies that $M(z + \delta z) = a_p$ and $M(z) = a_q, q \ne p$. In this case

$\mathbf{u}(z)$ lies in a small volume element δV_q around $\mathbf{u}(z) = [\mathbf{u} + o(1)]$, and $\delta V_q = [1 + o(1)]\delta V$.

We now combine the possible events at z leading to the desired event at $z + \delta z$ and, from (2.6)–(2.9), obtain

$$\sigma_p(\mathbf{u}, z + \delta z)\delta V = [1 + \tau_{pp}(z)\delta z]\sigma_p(\mathbf{u} - \mathbf{f}(\mathbf{u}, a_p, z)\delta z, z)\delta V_p + \left(\sum_{q \neq p} \tau_{qp}(z)\delta z\, \sigma_q(\mathbf{u}, z) + o(\delta z)\right)\delta V. \quad \text{(A3)}$$

Dividing Eq. (A3) by δV and using (A2), then subtracting $\sigma_p(\mathbf{u}, z)$ from each side and dividing by δz, we obtain, in the limit $\delta z \to 0+$,

$$\frac{\partial \sigma_p}{\partial z} + \sum_{i=1}^{L} \frac{\partial}{\partial u_i}[f_i(\mathbf{u}, a_p, z)\sigma_p(\mathbf{u}, z)] - \sum_{q=1}^{N} \tau_{qp}(z)\sigma_q(\mathbf{u}, z) = 0, \quad p = 1, \ldots, N. \quad \text{(A4)}$$

APPENDIX B

We here derive the backward equations for the transition probability density functions, defined by (1.11). For $0 \leq \xi < z$, we consider a small increment $\delta\xi$ in ξ, with $\xi + \delta\xi < z$, and determine the probability of events at z, given that $\mathbf{u}(\xi) = \mathbf{v}$ and $M(\xi) = a_q$, in terms of the events that may occur in the intervals $(\xi, \xi + \delta\xi)$ and $(\xi + \delta\xi, z)$. Firstly, from (2.6) and Definition 2.1,

$$\text{Prob}\{M(\xi + \delta\xi) = a_r \,|\, M(\xi) = a_q\} = \sum_{n=0}^{\infty} P_{qr}^{(n)}(\xi, \xi + \delta\xi). \quad \text{(B1)}$$

The event $E_q^{(0)}(\xi, \xi + \delta\xi)$ implies that $M(x) = a_q$, $\xi \leq x \leq \xi + \delta\xi$, and hence, from (1.6), that

$$\mathbf{u}(\xi + \delta\xi) = \mathbf{u}(\xi) + \mathbf{f}(\mathbf{u}(\xi), a_q, \xi)\delta\xi + o(\delta\xi). \quad \text{(B2)}$$

Secondly, given $M(\xi) = a_q$, the event $E_r^{(1)}(\xi, \xi + \delta\xi)$, $q \neq r$, implies that $\mathbf{u}(\xi + \delta\xi) = \mathbf{u}(\xi) + o(1)$.

Since the joint process $\{\mathbf{u}(z), M(z)\}$ is a Markov process, the probabilities of events at z conditioned on $(\mathbf{u}(x), M(x))$, $\xi \leq x \leq \xi + \delta\xi$, are just the probabilities conditioned on $(\mathbf{u}(\xi + \delta\xi), M(\xi + \delta\xi))$. Hence, from (1.11), (2.7)–(2.9), (B1), and (B2), it follows that

$$\rho_{pq}(\mathbf{u}, z; \mathbf{v}, \xi)d^L\mathbf{u} = \left(\sum_{r \neq q} \tau_{qr}(\xi)\delta\xi\rho_{pr}(\mathbf{u}, z; \mathbf{v}, \xi) + o(\delta\xi)\right)d^L\mathbf{u} + [1 + \tau_{qq}(\xi)\delta\xi]\rho_{pq}(\mathbf{u}, z; \mathbf{v} + \mathbf{f}(\mathbf{v}, a_q, \xi)\delta\xi, \xi + \delta\xi)d^L\mathbf{u}. \quad \text{(B3)}$$

Subtracting the left-hand side of (B3) from the right-hand side, dividing by $d^L\mathbf{u}\delta\xi$, and letting $\delta\xi \to 0+$, we obtain the backward equations (2.15).

APPENDIX C

We begin by considering the moments of $\sigma_p(\mathbf{W}, z)$, defined by (3.5). For the first-order moments we multiply Eq. (3.3) by w_{hl}, and integrate with respect to the elements of $\mathbf{W}$ over R^{nm}. After an integration by parts we obtain

$$\frac{d}{dz}\langle w_{hl}\rangle_p = \sum_{j=1}^{n} A_{hj}(a_p, z)\langle w_{jl}\rangle_p + \sum_{r=1}^{N} \tau_{rp}(z)\langle w_{hl}\rangle_r. \quad \text{(C1)}$$

This may be written in the matrix form (3.7).

Next, for the second-order moments, we multiply Eq. (3.3) by $w_{fg}w_{hl}$, and integrate over R^{nm}. But,

$$\sum_{i=1}^{n}\sum_{k=1}^{m} A_{ij}(a_p, z)w_{jk}\frac{\partial}{\partial w_{ik}}(w_{fg}w_{hl}) = A_{fj}(a_p, z)w_{jg}w_{hl} + A_{hj}(a_p, z)w_{jl}w_{fg}. \quad \text{(C2)}$$

Thus,

$$\frac{d}{dz}\langle w_{fg}w_{hl}\rangle_p = \sum_{r=1}^{N} \tau_{rp}(z)\langle w_{fg}w_{hl}\rangle_r + \sum_{j=1}^{n} [A_{fj}(a_p, z)\langle w_{jg}w_{hl}\rangle_p + A_{hj}(a_p, z)\langle w_{fg}w_{jl}\rangle_p]. \quad \text{(C3)}$$

This may be written in the matrix form (3.9), since

$$(\mathbf{AW}) \times \mathbf{W} = (\mathbf{A} \times \mathbf{I}_n)(\mathbf{W} \times \mathbf{W}), \quad \mathbf{W} \times (\mathbf{AW}) = (\mathbf{I}_n \times \mathbf{A})(\mathbf{W} \times \mathbf{W}). \quad \text{(C4)}$$

The higher-order moments may be calculated in a similar way.

We now consider the correlation functions, and define the matrix functions $\mathbf{X}_{pq}(z; \mathbf{V}, \xi)$, $p, q = 1, \ldots, N$, by

$$\mathbf{X}_{pq}(z; \mathbf{V}, \xi) = \int_{R^{nm}} \mathbf{W}\rho_{pq}(\mathbf{W}, z; \mathbf{V}, \xi)d^{nm}\mathbf{W}. \quad \text{(C5)}$$

It follows from (3.15) that

$$\frac{\partial \mathbf{X}_{pq}}{\partial z} = \mathbf{A}(a_p, z)\mathbf{X}_{pq} + \sum_{r=1}^{N} \tau_{rp}(z)\mathbf{X}_{rq}, \quad \text{(C6)}$$

with initial conditions, from (3.16),

$$\mathbf{X}_{pq}(\xi; \mathbf{V}, \xi) = \delta_{pq}\mathbf{V}. \quad \text{(C7)}$$

It then follows that

$$\mathbf{X}_{pq}(z; \mathbf{V}, \xi) = \Psi_{pq}(z, \xi)\mathbf{V}, \quad \text{(C8)}$$

where the matrices Ψ_{pq}, $p, q = 1, \ldots, N$, satisfy (3.17) and (3.18).

But, from (1.12), for $0 \leq \xi \leq z$,

$$\langle \mathbf{W}(z) \times \mathbf{W}(\xi)\rangle = \sum_{p=1}^{N}\sum_{q=1}^{N} \int_{R^{nm}}\int_{R^{nm}} (\mathbf{W} \times \mathbf{V}) \times \rho_{pq}(\mathbf{W}, z; \mathbf{V}, \xi)\sigma_q(\mathbf{V}, \xi)d^{nm}\mathbf{W}d^{nm}\mathbf{V}. \quad \text{(C9)}$$

Hence, from (C5) and (C8),

$$\langle \mathbf{W}(z) \times \mathbf{W}(\xi)\rangle = \sum_{p=1}^{N}\sum_{q=1}^{N} [\Psi_{pq}(z, \xi) \times \mathbf{I}_n] \int_{R^{nm}} (\mathbf{V} \times \mathbf{V})\sigma_q(\mathbf{V}, \xi)d^{nm}\mathbf{V}. \quad \text{(C10)}$$

Thus, from (3.5) and (C10), we obtain (3.20).

APPENDIX D

We here derive the alternate formulation for the calculation of the correlation functions. We define the $N \times N$ matrix $\boldsymbol{\rho}$ and the column vector $\boldsymbol{\sigma}$ by

$$\boldsymbol{\rho} = (\rho_{pq}), \quad \boldsymbol{\sigma} = \text{col}(\sigma_1, \ldots, \sigma_N). \quad \text{(D1)}$$

Then, from (4.4), Eq. (3.15) may be written in the form

$$\frac{\partial \boldsymbol{\rho}}{\partial z} + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} \mathbf{D}_{ij}(z)\frac{\partial(w_{jk}\boldsymbol{\rho})}{\partial w_{ik}} - \boldsymbol{\tau}^t(z)\boldsymbol{\rho} = 0, \quad \text{(D2)}$$

with initial condition, from (3.16),

$$\boldsymbol{\rho}(\mathbf{W}, \xi; \mathbf{V}, \xi) = \delta(\mathbf{W} - \mathbf{V})\mathbf{I}_N. \quad \text{(D3)}$$

Also, from (1.12),

$$\langle w_{fg}(\xi)w_{hl}(z)\rangle = \mathbf{E}_N \int_{R^{nm}} \int_{R^{nm}} v_{fg}w_{hl}\rho(\mathbf{W}, z; \mathbf{V}, \xi)\sigma(\mathbf{V}, \xi)d^{nm}\mathbf{W}\,d^{nm}\mathbf{V}. \tag{D4}$$

We now define the matrix functions $\mathbf{\Lambda}_{hl}(z; \mathbf{V}, \xi)$, $h = 1, \ldots, n$, $l = 1, \ldots, m$, by

$$\mathbf{\Lambda}_{hl}(z; \mathbf{V}, \xi) = \int_{R^{nm}} w_{hl}\rho(\mathbf{W}, z; \mathbf{V}, \xi)d^{nm}\mathbf{W}. \tag{D5}$$

Then it follows from (D2) that

$$\frac{\partial \mathbf{\Lambda}_{hl}}{\partial z} = \sum_{j=1}^{n} \mathbf{D}_{hj}(z)\mathbf{\Lambda}_{jl} + \tau^t(z)\mathbf{\Lambda}_{hl}, \tag{D6}$$

with initial condition, from (D3),

$$\mathbf{\Lambda}_{hl}(\xi; \mathbf{V}, \xi) = v_{hl}\mathbf{I}_N. \tag{D7}$$

Hence, from (4.25) and (4.26),

$$\mathbf{\Lambda}_{hl}(z; \mathbf{V}, \xi) = \sum_{k=1}^{n} \theta_{hk}(z, \xi)v_{kl}. \tag{D8}$$

Thus, from (D4), (D5), and (D8),

$$\langle w_{fg}(\xi)w_{hl}(z)\rangle = \sum_{k=1}^{n} \mathbf{E}_N \int_{R^{nm}} \theta_{hk}(z, \xi)v_{fg}v_{kl}\sigma(\mathbf{V}, \xi)d^{nm}\mathbf{V}. \tag{D9}$$

Equation (4.28) follows from (3.5), (4.1), (4.13), and (D1).

1 J. L. Doob, *Stochastic Processes* (Wiley, New York, 1953), pp. 235–55.
2 J. McKenna and J. A. Morrison, J. Math. Phys. **12**, 2126 (1971).
3 J. McKenna and J. A. Morrison, J. Math. Phys. **11**, 2348 (1970).
4 A. Blanc-Lapierre and R. Fortet, *Theory of Random Functions* (Gordon and Breach, New York, 1965), Vol. I, p. 161.
5 W. M. Wonham in *Probabilistic Methods in Applied Mathematics*, Vol. II, edited by A. T. Bharucha-Reid (Academic, New York, 1970), p. 131.
6 N. N. Krasovskii and E. A. Lidskii, Appl. Math. Mech. **25**, 627 (1961).
7 N. N. Krasovskii and E. A. Lidskii, Automat. Remote Control **22**, 1021 (1961); **22**, 1141, 1289 (1961).
8 R. Bellman, *Introduction to Matrix Analysis* (McGraw-Hill, New York, 1960), p. 227.

Projective Representations and the Relation of Internal to Space–Time Symmetries*

Russell E. Warren and William H. Klink
Department of Physics and Astronomy, University of Iowa, Iowa City, Iowa 52240

(Received 19 August 1971)

The relation of space–time to internal symmetries in relativistic quantum mechanics is investigated using the Mackey theory of induced projective representations of group extensions. Representation multipliers are found for semidirect products of the Poincaré group and arbitrary internal symmetry groups. Upon investigating a typical example, it is found that when relations such as $U_\pi U_C = (-1)^{2J}U_C U_\pi$, obtained from field theory, are assumed, a unique choice of representation multiplier follows; in particular, this multiplier requires $U^2_{\mathcal{C}\pi t} = 1$ for all J. Representations relative to this multiplier are computed.

1. INTRODUCTION

There has been a recurrent interest in the problem of relating physical symmetries associated with space and time to other symmetries. These symmetries have their origin in the observed invariances of dynamical systems under transformations which mix the charges associated with elementary particles such as baryon number and electric charge. As pointed out by Michel,[1] it is improper to treat the quantum numbers associated with the latter symmetries, the so-called internal symmetries, as though they were independent of space–time quantum numbers since there are empirical relations between them which appear to hold universally such as the observation that strongly interacting fermions possess odd baryon number while bosons have even baryon number. Also, as emphasized by Michel and Kamber and Straumann,[1,2] the necessarily anti-unitary nature of the time inversion operator requires that time reversal as a group element act nontrivially on any internal symmetry group with which time reversal symmetry is combined.

From considerations such as these, one is led to the general question of the possible relations between space–time and internal symmetries. The purpose of this work is to examine this question group theoretically by constructing overall symmetry groups which combine space–time and internal symmetries nontrivially and by finding representations of these groups by unitary or anti-unitary operators acting on the Hilbert space of states of the system which is invariant under the over-all symmetry group. The basis for this investigation is the generalization made by Lee and Wick[3] of the relation between superselected variables and geometric transformations to include arbitrary internal symmetries. They reason that since the Hilbert space of a system is broken into "noncoherent subspaces", each labeled by the charges of the system such that any symmetry operator is defined only up to an arbitrary phase in each sector, every geometrical symmetry operator is represented by a coset of the gauge group generated by the superselected charges. In this way, one is led to a larger symmetry group of the system that contains the gauge group associated with the charges invariantly and in which the geometrical symmetry group need only appear as the quotient of the total symmetry group by the gauge group.

This generalization is made by supposing the system to be described by some model Hamiltonian H, say one in which the electromagnetic and weak interactions are neglected. An internal symmetry operator $\mathcal{I}$ is defined as a nongeometrical unitary operator that commutes with H. The group formed by the set of internal symmetry operators is the internal symmetry group of the system in Lee and Wick's terminology. If, now, one considers an arbitrary space–time transformation represented by an oper-

Also, from (1.12),

$$\langle w_{fg}(\xi)w_{hl}(z)\rangle = \mathbf{E}_N \int_{R^{nm}} \int_{R^{nm}} v_{fg}w_{hl}\rho(\mathbf{W}, z; \mathbf{V}, \xi)\sigma(\mathbf{V}, \xi)d^{nm}\mathbf{W}\,d^{nm}\mathbf{V}. \tag{D4}$$

We now define the matrix functions $\mathbf{\Lambda}_{hl}(z; \mathbf{V}, \xi)$, $h = 1, \ldots, n$, $l = 1, \ldots, m$, by

$$\mathbf{\Lambda}_{hl}(z; \mathbf{V}, \xi) = \int_{R^{nm}} w_{hl}\rho(\mathbf{W}, z; \mathbf{V}, \xi)d^{nm}\mathbf{W}. \tag{D5}$$

Then it follows from (D2) that

$$\frac{\partial \mathbf{\Lambda}_{hl}}{\partial z} = \sum_{j=1}^{n} \mathbf{D}_{hj}(z)\mathbf{\Lambda}_{jl} + \tau^t(z)\mathbf{\Lambda}_{hl}, \tag{D6}$$

with initial condition, from (D3),

$$\mathbf{\Lambda}_{hl}(\xi; \mathbf{V}, \xi) = v_{hl}\mathbf{I}_N. \tag{D7}$$

Hence, from (4.25) and (4.26),

$$\mathbf{\Lambda}_{hl}(z; \mathbf{V}, \xi) = \sum_{k=1}^{n} \theta_{hk}(z, \xi)v_{kl}. \tag{D8}$$

Thus, from (D4), (D5), and (D8),

$$\langle w_{fg}(\xi)w_{hl}(z)\rangle = \sum_{k=1}^{n} \mathbf{E}_N \int_{R^{nm}} \theta_{hk}(z, \xi)v_{fg}v_{kl}\sigma(\mathbf{V}, \xi)d^{nm}\mathbf{V}. \tag{D9}$$

Equation (4.28) follows from (3.5), (4.1), (4.13), and (D1).

1 J. L. Doob, *Stochastic Processes* (Wiley, New York, 1953), pp. 235–55.
2 J. McKenna and J. A. Morrison, J. Math. Phys. **12**, 2126 (1971).
3 J. McKenna and J. A. Morrison, J. Math. Phys. **11**, 2348 (1970).
4 A. Blanc-Lapierre and R. Fortet, *Theory of Random Functions* (Gordon and Breach, New York, 1965), Vol. I, p. 161.
5 W. M. Wonham in *Probabilistic Methods in Applied Mathematics*, Vol. II, edited by A. T. Bharucha-Reid (Academic, New York, 1970), p. 131.
6 N. N. Krasovskii and E. A. Lidskii, Appl. Math. Mech. **25**, 627 (1961).
7 N. N. Krasovskii and E. A. Lidskii, Automat. Remote Control **22**, 1021 (1961); **22**, 1141, 1289 (1961).
8 R. Bellman, *Introduction to Matrix Analysis* (McGraw-Hill, New York, 1960), p. 227.

Projective Representations and the Relation of Internal to Space–Time Symmetries*

Russell E. Warren and William H. Klink
Department of Physics and Astronomy, University of Iowa, Iowa City, Iowa 52240

(Received 19 August 1971)

The relation of space–time to internal symmetries in relativistic quantum mechanics is investigated using the Mackey theory of induced projective representations of group extensions. Representation multipliers are found for semidirect products of the Poincaré group and arbitrary internal symmetry groups. Upon investigating a typical example, it is found that when relations such as $U_\pi U_C = (-1)^{2J}U_C U_\pi$, obtained from field theory, are assumed, a unique choice of representation multiplier follows; in particular, this multiplier requires $U^2_{\mathcal{C}\pi t} = 1$ for all J. Representations relative to this multiplier are computed.

1. INTRODUCTION

There has been a recurrent interest in the problem of relating physical symmetries associated with space and time to other symmetries. These symmetries have their origin in the observed invariances of dynamical systems under transformations which mix the charges associated with elementary particles such as baryon number and electric charge. As pointed out by Michel,[1] it is improper to treat the quantum numbers associated with the latter symmetries, the so-called internal symmetries, as though they were independent of space–time quantum numbers since there are empirical relations between them which appear to hold universally such as the observation that strongly interacting fermions possess odd baryon number while bosons have even baryon number. Also, as emphasized by Michel and Kamber and Straumann,[1,2] the necessarily anti-unitary nature of the time inversion operator requires that time reversal as a group element act nontrivially on any internal symmetry group with which time reversal symmetry is combined.

From considerations such as these, one is led to the general question of the possible relations between space–time and internal symmetries. The purpose of this work is to examine this question group theoretically by constructing overall symmetry groups which combine space–time and internal symmetries nontrivially and by finding representations of these groups by unitary or anti-unitary operators acting on the Hilbert space of states of the system which is invariant under the over-all symmetry group. The basis for this investigation is the generalization made by Lee and Wick[3] of the relation between superselected variables and geometric transformations to include arbitrary internal symmetries. They reason that since the Hilbert space of a system is broken into "noncoherent subspaces", each labeled by the charges of the system such that any symmetry operator is defined only up to an arbitrary phase in each sector, every geometrical symmetry operator is represented by a coset of the gauge group generated by the superselected charges. In this way, one is led to a larger symmetry group of the system that contains the gauge group associated with the charges invariantly and in which the geometrical symmetry group need only appear as the quotient of the total symmetry group by the gauge group.

This generalization is made by supposing the system to be described by some model Hamiltonian H, say one in which the electromagnetic and weak interactions are neglected. An internal symmetry operator $\mathcal{I}$ is defined as a nongeometrical unitary operator that commutes with H. The group formed by the set of internal symmetry operators is the internal symmetry group of the system in Lee and Wick's terminology. If, now, one considers an arbitrary space–time transformation represented by an oper-

ator U, say, it is clear that both U and $\mathcal{G}U$ are physically equivalent; in fact any element of the set $\{\mathcal{G}U\}$ for a fixed U will equally well represent the transformation U. Thus there is a one-to-one correspondence between geometrical transformations and cosets of the group of internal symmetry operators. The totality of cosets generated this way is taken to be the over-all symmetry group G of the system. By definition, the internal symmetry elements appear in G invariantly (this must be so in order that coset multiplication be isomorphic to physical space–time transformations). Also, these geometrical transformations only appear as the quotient of G by the set of internal symmetries; in particular, the geometrical transformations need not be contained in G. It is seen that this approach makes a distinction between internal and space–time symmetries in that a space–time transformation is physically realizable since one can meaningfully speak of translating, rotating, or Lorentz boosting a physical system while for internal symmetry transformations such as phase transformations or rotations in isotopic spin space it is not so obvious what these operations mean; this is reflected in the structure of the over-all symmetry group whereby only the cosets have operational meaning, so that the representatives of a given geometrical transformation U, and $\mathcal{G}U$ are physically indistinguishable.

The analysis to be presented here differs in several respects from that of Lee and Wick. One difference lies in the nature of the space–time symmetry group considered. Lee and Wick restrict their attention to space and time inversions generating in either case what they call "minimal extensions" consisting of the identity coset and the inversion coset of the internal symmetry group. In this work the geometrical symmetry group will ultimately be the full Poincaré group including space and time inversions. In addition to this, the internal symmetries treated will be those common to all strongly interacting particles such as isotopic invariance, hypercharge, and charge conjugation invariance. In contrast to Lee and Wick, symmetries of the super-strong interactions such as $SU(3)$ will not be considered. Finally, throughout their work Lee and Wick deal with groups composed of symmetry transformation operators. This is, they work within the representation spaces of the underlying transformation groups. Due to the complexity of the groups considered here, it is desirable to distinguish the group from the representational aspects of the problem.

The problem of combining space–time with internal symmetries on a group level has been studied by Michel[1] and Kamber and Straumann[2] using the Eilenberg and MacLane[4] formulation of group extension cohomology theory. Michel's analysis is concerned primarily with the construction of over-all symmetry groups of the full Poincaré group (with space and time inversions) P by the set of Abelian gauge groups generated by the conserved charges baryon number, electric charge, and lepton number with an indication of how one proceeds in the case of non-Abelian internal symmetry groups. Michel considers nontrivial actions on the internal symmetry groups by time reversal as indicated above and a combination $C\pi$ of charge conjugation and space inversion elements. This was presumably motivated by Wigner's[5] contemporary redefinition of space inversion to include charge conjugation, since while parity invariance was known to be broken at that time, $C\pi$ invariance appeared to hold universally. However, in light of the since discovered $C\pi$ violation and the Lee and Wick emphasis on distinguishing space–time from internal symmetries, it is desirable to return to the earlier definition of π as the physical operation of space inversion and treat C as an internal symmetry transformation.

Kamber and Straumann[2] have made a detailed analysis of the possible action of time reversal on a class of internal symmetry groups I consisting of direct products of gauge groups with an additional semisimple Lie group such as $SU(2)$. Their results indicate that time reversal essentially can only cause a complex conjugation in the gauge group part of the internal symmetry group. Their work also presents a detailed description of the methods used in calculating the total symmetry groups as well as an indication of the structure of their vector representations.

Michel and Kamber and Straumann conclude that the only physically meaningful automorphism on I that P can induce is generated by time reversal. As mentioned earlier, it is essential that time reversal act nontrivially on I in order to preserve both the antiunitary nature of the operator representative of time reversal and the positive definiteness of the charges under time inversion. Furthermore, the Lorentz invariant character of the charges requires that all other elements of P have no effect on the charge groups.

The calculation of the set of nonequivalent extensions of P by I corresponding to the aforementioned automorphism produces the result that multipliers defined on $P \times P$ having values in I can be nontrivial only for the following elements in $P \times P$: (t, t), $(\pi t, \pi t)$, and $(2\pi, 2\pi)$ where t is time reversal, π is space inversion and 2π is the element of the Lorentz group representing a rotation about an arbitrary axis by 2π. Also, the multipliers can take on values only in the center of I. It is found that multipliers evaluated over the translations or any other Lorentz group elements including space inversion have only the identity in I as an image. Furthermore, it should be noted that there can be no multipliers defined over P and I having values in I. For example, multipliers such as $\sigma(\pi, C)$, where C is the charge conjugation transformation, taking values in I are not allowed.

The group multiplier possibilities for a given I generate nonequivalent extensions which differ most significantly in the relation between spin and the internal symmetry quantum numbers which their structures dictate. In particular, an internal symmetry group containing phase transformations whose generator is baryon number can be coupled to the Poincaré group in such a fashion that the relation $(-1)^{2J+B} = 1$ is satisfied where J is the spin and B is the baryon number representation label.

Only relations of this sort appear to be possible by working simply within the context of the groups themselves and their vector representations. While relations such as $(-1)^{2J+B} = 1$ are interesting, they certainly do not exhaust the connections observed between space–time and internal symmetry variables.

As an example, consider the relation $U_\pi U_C = \omega U_C U_\pi$ where $\omega = \pm 1$ resulting from field theory. One finds that this cannot be obtained straightforwardly on a group level, but it will be shown in Sec. 3 how it can be constructed on a representational level by considering a broader class of representations, namely those which are defined only up to phase factors, the so-called ray or projective representations. The purpose of this work is to explore the additional relations possible between space–time and internal symmetries by making full use of the phase ambiguity inherent in quantum mechanical states to construct projective representations of over-all symmetry groups.

The mathematics needed for constructing projective representations of groups of interest to physics has been developed by Mackey.[6] The results of the Mackey theory of induced projective representations needed here are summarized in Sec. 2, together with some of the modifications necessary for constructing representations involving anti-unitary operators which arise from time reversal. In Sec. 3, these results are applied to the problem of constructing ray representations of groups that combine space–time with strong interaction symmetries as semidirect products since the mathematics needed for constructing representation multipliers for semidirect products has been provided by Mackey. The determination of representation multipliers for arbitrary extensions is a much more difficult problem and will not be undertaken here. It is found that representation multipliers in general do not lead to relations between space–time symmetries P and the continuous transformations belonging to the internal symmetry group I; only discrete elements of I such as charge conjugation are connected to P via representation multipliers. Also, it will be shown that of the multiplier possibilities which do arise, relations such as $U_\pi U_C = (-1)^{2J} U_C U_\pi$ may be used to define a unique multiplier set consistent with quantum field theory.

One of the consequences of fixing the multiplier in this way will be that $U^2_{C\pi t} = 1$ for all values of the spin; such a result disagrees with the relation $U^2_{C\pi t} = (-1)^{2J}$ given by Lee and Wick.[3]

2. INDUCED PROJECTIVE REPRESENTATIONS

The purpose of this section is to provide those familiar with Mackey's[7] work on induced representations of semidirect product groups with Abelian invariant subgroups, with a summary of his generalization to include non-Abelian invariant subgroups and group extensions with nontrivial group multipliers. All questions of an analytical nature such as the existence of quasi-invariant measures or the necessity of introducing Radon–Nikodym derivatives into the formalism will be ignored. Furthermore, it will be assumed here and in the calculations that the orbits are transitive, thereby justifying the use of Mackey's main theorems on the construction of induced projective representations.

In the calculation of induced representations of groups G with Abelian invariant subgroups, one considers functions F over the elements $g \in G$ with the property $F(hg) = L(h)F(g)$, where L is a representation of a closed subgroup H of G. The functions F form a Hilbert space with norm

$$\|F\|^2 = \int_{G/H} d\mu(g) \|F(g)\|^2_{\mathfrak{K}(L)} < \infty,$$

where μ is an (quasi) invariant measure over G/H and $\mathfrak{K}(L)$ is the vector space on which L is defined. A unitary representation U^L of G induced by L is then defined by

$$U^L(g_0)F(g) = F(gg_0).$$

In order to construct induced projective representations, it is necessary to specify initially a multiplier σ for the representation. One now considers functions F over the group manifold with the property

$$F(hg) = [1/\sigma(h,g)]L(h)F(g), \tag{1}$$

where $L(h)$ is a σ representation of a subgroup H of G. As before, the F form a Hilbert space with the norm above. The unitary σ representation U^L of G induced by L is now

$$U^L(g_0)F(g) = \sigma(g,g_0)F(gg_0). \tag{2}$$

Irreducible representations are obtained in the Abelian invariant subgroup case by choosing H to be

$$H = \{g \mid g \in G, M(gkg^{-1}) \cong M(k) \ \forall\ k \in K\},$$

where M is a representation of the Abelian normal subgroup K. The analogous H for constructing irreducible σ representations is given by

$$H = \left\{g \mid g \in G, \frac{\sigma(gk,g^{-1})\sigma(g,k)}{\sigma(g^{-1},g)} M(gkg^{-1}) \cong M(k) \ \forall\ k \in K\right\}, \tag{3}$$

where M is a σ representation of K.

In terms of the groups to be considered here, the Mackey theorem for constructing irreducible induced projective representations may be stated. Given G, a closed normal subgroup K, a multiplier σ, and an irreducible σ representation L of K, Mackey shows that there is a one-to-one correspondence between the set of all possible ω representations N of H/K (H is the inducing subgroup above) and the σ representations of H which induce σ representations of G. The σ representation of H for a given ω representation N is $M \otimes N'$, where M is a $\sigma/\omega\circ f$ representation of H such that $M(k) = L(k)$ for all $k \in K$, f is the canonical homomorphism, $f : H \to H/K$, and N' is the $\omega\circ f$ representation of H constructed by composing N with f. Mackey further shows $U^{M\otimes N'}$ is irreducible if and only if N is irreducible. Thus the problem of finding all irreducible representations of a group G corresponding to a given multiplier σ is reduced to finding all possible irreducible multiplier representations of a subgroup of G, namely H/K.

It should be emphasized that the above theory is needed for non-Abelian K whether or not σ is nontrivial; in particular, the construction of vector representations of G starting with non-Abelian K requires one to consider all possible multiplier representations of H/K.

So far nothing has been said about finding the representation multiplier σ. In the case where G is a semidirect product of a normal subgroup K by another group R, Mackey gives a prescription for finding

every multiplier ν for G. Specifically, he shows that

$$\nu(x_1y_1, x_2y_2) = \sigma(x_1, y_1(x_2))\omega(y_1, y_2)g(x_2, y_1), \quad (4)$$

where x_1 and x_2 belong to K, y_1 and y_2 belong to R, $y_1(x_2)$ is the automorphism generated by y_1 acting on x_2, σ is a multiplier for K, ω is a multiplier for R, and g is a function from K and R into the complex numbers of modulus one such that

$$\begin{aligned}&\text{(a)}\ \ g(x,e) = 1 \ \forall\ x \in K,\\ &\text{(b)}\ \ \sigma(y(x_1), y(x_2)) = \sigma(x_1, x_2)g(x_1x_2, y)/g(x_1, y)g(x_2, y),\\ &\text{(c)}\ \ g(x, y_1y_2) = g(y_2(x), y_1)g(x, y_2).\end{aligned} \quad (5)$$

Also, Mackey proves the converse; given σ, ω, and g as above, the ν generated is a multiplier.

For applications to physics, the theory just sketched must be extended to include groups some of whose elements are represented by anti-unitary operators. Parthasarathy[8] has undertaken this extension in the context of representations defined on functions over homogeneous spaces. He distinguishes two cases depending on whether the including subgroup H contains elements represented by anti-unitary operators or not. The analysis to be made here differs somewhat from Parthasarathy's in that representations are constructed on functions defined over elements of G and that the inducing subgroup is computed only with respect to elements G^+ represented by unitary operators. This means that H will always be contained in G^+ and representations of G will look like Parthasarathy's second case.[8] The reasons for doing this are the desire to remain as close to the Mackey theory as possible (that is, imprimitivity systems based on representations of a normal subgroup K) and the similarity this approach has to Wigner's treatment of time reversal.[9] The price one pays for using this procedure is that the representations are not necessarily irreducible with respect to the anti-unitary elements. Practically, this means time reversal causes a doubling of the representation space in some, but not all cases. The representations found this way must be examined to see if this doubling actually occurs.

The multiplier theorem for semidirect product groups determined by Mackey holds only for groups whose elements are represented by unitary operators. In including anti-unitary operators it is found that Eq. (4) is correct in general, but that Eq. (5) must be replaced by

$$\begin{aligned}&\text{(a)}\ \ g(x,e) = 1 \ \forall\ x \in K,\\ &\text{(b)}\ \ \sigma(y(x_1), y(x_2)) = \bar{\sigma}(x_1, x_2)g(x_1x_2, y)/g(x_1, y)g(x_2, y)\\ &\text{(c)}\ \ g(x, y_1y_2) = g(y_2(x), y_1)\bar{g}(x, y_2),\end{aligned} \quad (6)$$

in case y in (b) and y_1 in (c) are represented by anti-unitary operators. Note that it is assumed that elements belonging to K are always represented by unitary operators since this is the situation in the work undertaken here. The next section will apply the Mackey theory to the calculation of multiplier representations of semidirect products of the Poincaré group and the internal symmetry groups of strong interactions.

3. REPRESENTATIONS OF $I \circledcirc P$

The first goal in this section is the determination of representation multipliers for semidirect products of the full Poincaré group P and a general class of internal symmetry groups I formed by an arbitrary compact Lie group I_0 and the two element group generated by charge conjugation Z_2^C. It will be seen that the Mackey multiplier theorem extended in the last section to include groups with anti-unitary elements provides an almost unique characterization of the multiplier possibilities for $I \circledcirc P$.

Consider, then, an I of the form $I_0 \circledcirc Z_2^C$, where I_0 is any compact Lie group. P acts on I only through time reversal in a manner determined by Kamber and Straumann.[2] The most general multiplier for $G = I \circledcirc P$ is

$$\sigma(i_1p_1, i_2p_2) = \nu(i_1, p_1(i_2))\tau(p_1, p_2)g(i_2, p_1), \quad (7)$$

where i_1, i_2 belong to I, p_1, p_2 belong to P, ν is a multiplier for I and τ is a multiplier for P, and g is the function satisfying Eqs. (5) and (6). The most general form of I_0 is $I_0 = U_1 \times \cdots \times U_1 \times A$, where $U_1 \times \cdots \times U_1$ is an m-dimensional torus consisting here of gauge groups generated by the conserved charges of the system and A is a semisimple Lie group.[2] ν, in general, has the form

$$\nu(i_1, i_2) = \rho(i_1^0, r_1(i_2^0))\gamma(r_1, r_2)k(i_2^0, r_1), \quad (8)$$

where i_1^0, i_2^0 belong to I_0, r_1, r_2 belong to Z_2^C and k satisfies Eq. (5). Continued application of the Mackey formula allows one to construct all possible ρ for a given I_0. It is found that there are nontrivial solutions involving the gauge groups U_1, but that these seem to have no physical significance; consequently ρ will be taken to be 1. Z_2^C has only trivial multipliers, and so $\gamma = 1$. Using the defining relations for k, Eq. (5), noting that A is semisimple, and demanding that k be a single-valued function restricts the possible choices of k to

$$\begin{aligned}&k((\theta_1, \ldots, \theta_m, a), e) = 1, \quad k((\theta_1, \ldots, \theta_m, a), C)\\ &\qquad = e^{i(n_1\theta_1 + \cdots + n_m\theta_m)}\end{aligned} \quad (9)$$

for all $a \in A$, where $e^{i\theta_1}, \ldots, e^{i\theta_m}$ belongs to $U_1 \times \cdots \times U_1$ and $n_1, \ldots, n_m$ are arbitrary integers. Thus, there is an m-fold countable infinity of nonequivalent multipliers for I whose effect is to include more than one set of charges in an irreducible representation of I as will be seen later.

It is well known[5] that multipliers for P are nontrivial only for the discrete elements of P, that is $\tau(p_1, p_2) = \omega(x_1, x_2)$ where ω takes on values

x_1 \ x_2	π	t	πt
π	1	1	1
t	$\alpha\beta$	α	β
πt	$\alpha\beta$	α	β

(10)

for $\alpha, \beta = \pm 1$.

It is found that g is necessarily a one-dimensional

representation of I for all $p \in P$. This implies that g can at most be defined over elements of I belonging to the center of I and Z_2^C. Similarly, since g must be a one-dimensional representation of P_0 for all i, g is nontrivial only for elements of the four group V. Using the relations

$$\begin{aligned} g(i, x_1x_2) &= g(x_2(i), x_1)g(i, x_2), \quad x_1 \in V^+, \\ &= g(x_2(i), x_1)\bar{g}(i, x_2), \quad x_1 \in V^-, \end{aligned}$$

and the fact that time reversal acting on elements of the center of I takes them into their inverses,[2] one finds $g(i, x) = \pm 1$. In the special case where I has a trivial center, $g(i, p)$ is a function only over the discrete elements C, π, t; there are four possible g given by $g_i(e, x) = 1$ for all $x \in V, i = 1, \ldots, 4$, and we have the following.

$g_i((\theta, C), x)$ \ x	e	π	t	πt
g_1	1	1	1	1
g_2	1	1	−1	−1
g_3	1	−1	1	−1
g_4	1	−1	−1	1

(11)

Thus when I has no center, the space-time transformations only "interact" with internal symmetry groups via representation multipliers through the discrete transformations.

The question of which of the multiplier possibilities above are realized physically will be considered in the context of the representations of a specific example. To this end consider the semidirect product G of the cover of the full Poincaré group P by $U_1 \otimes Z_2^C$, where U_1 is the gauge group of baryon number and Z_2^C is the two element group generated by charge conjugation. The automorphism on $U_1 \otimes Z_2^C$ is taken to be complex conjugation by time reversal. A multiplier for G will have the form

$$\begin{aligned} &\sigma((\theta_1, r_1, p_1^0, x_1), (\theta_2, r_2, p_2^0, x_2)) \\ &\quad = k_n(\theta_2, r_1)\omega_{\alpha\beta}(x_1, x_2)g_i(r_2, x_1), \end{aligned} \tag{12}$$

where k, ω, and g are defined above.

In order to find the σ representations of G, take $K = U_1 \otimes Z_2^C$ to be the normal subgroup of G. σ representations of K are found to be

$${}^BL_{\theta,e} = \begin{pmatrix} e^{iB\theta} & 0 \\ 0 & e^{i(n-B)\theta} \end{pmatrix}, \quad {}^BL_{\theta,C} = \begin{pmatrix} 0 & e^{iB\theta} \\ e^{i(n-B)\theta} & 0 \end{pmatrix}, \tag{13}$$

where B is an integer. As mentioned in Sec. 2, the inducing subgroup H will be computed only relative to G^+, that is, the elements of G represented by unitary operators. In this case $G^+ = [U_1 \otimes Z_2^C] \times [P_0 \otimes Z_2^{\mathcal{P}}]$ where $Z_2^{\mathcal{P}}$ is the two element group generated by space inversion. Thus

$$\begin{aligned} H = \Big\{ g \mid g \in G^+, \frac{\sigma(g(\theta, r), g^{-1})\sigma(g, (\theta, r))}{\sigma(g^{-1}, g)} \quad {}^BL_{g(\theta,r)g^{-1}} \\ \cong {}^BL_{\theta,r} \ \forall \ (\theta, r) \in K \Big\}. \end{aligned} \tag{14}$$

H is found to be all of G^+ for any of the four possible choices of g_i. One must now find all σ representations of $H = G^+$; the Mackey prescription is first to construct all possible multiplier representations of $H/K = P_0 \otimes Z_2^{\mathcal{P}}$. There are only vector representations which for mass $m \neq 0$ and spin J are defined on functions over right coset elements,[7] $\Lambda_c \in SL(2, C)$, as

$$N(a, \Lambda, s)F_\lambda(\Lambda_c) = \eta_i(s)e^{i\hat{p}\cdot\Lambda_c(a)}\sum_{\lambda'} D^J_{\lambda\lambda'}(R)F_{\lambda'}(s\Lambda_c' s), \tag{15}$$

where a is a translation, Λ is an element of $SL(2, C)$, $s \in Z_2^{\mathcal{P}}$, $\Lambda_c\Lambda = R\Lambda_c'$ for $R \in SU(2)$, $\hat{p}$ is the standard vector $(m, 0)$, $D^J(R)$ is an irreducible representation of $SU(2)$, and $\eta_i(s)$ is an irreducible representation of $Z_2^{\mathcal{P}}$ given by $\eta_1(e) = \eta_2(e) = 1$, $\eta_1(\pi) = -\eta_2(\pi) = 1$.

To complete the determination of irreducible σ representations of H, one must find a σ representation of H such that $M_{\theta,r} = L_{\theta,r}$. It is found that

$$M(\theta, r, a, \Lambda, s)\xi(r_0) = k_n(\theta, r_0)g_i(r_0 r, s)e^{iBr_0(\theta)}\xi(r_0 r) \tag{16}$$

for all a, Λ is such a representation. Then σ representations V of H are given by $V(\theta, r, a, \Lambda, s) = M(\theta, r, a, \Lambda, s) \otimes N(a, \Lambda, s)$ acting on basis functions of the form $\xi(r)F_\lambda(\Lambda_c)$.

In order to obtain representations defined over more familiar basis functions set $F_\lambda(\Lambda_c) = \varphi_\lambda(p)$, where $p = \Lambda_c^{-1}\hat{p}$ is a momentum variable. In a helicity basis one can then show $F_\lambda(\pi\Lambda_c\pi) = (-1)^J\varphi_{-\lambda}(p')$ where $p' = (p_0, -\mathbf{p})$. Similarly, one may set $\xi(r) = \psi(b)$ where $\xi(e) = \psi(B)$ and $\xi(C) = \psi(n - B)$. Finally, the representation V may be expressed in terms of (non-normalizable) state vectors by writing

$$|\chi\rangle = \sum_{b,\lambda} \int \frac{d^3p}{E}\psi(b)\varphi_\lambda(p)|[m, J, |B|]\mathbf{p}, \lambda, \eta_b, b\rangle, \tag{17}$$

where the state vector under the integral sign is labeled by the irreducible representation labels m, J, $|B|$ corresponding to mass, spin, and absolute value of the baryon number, and the diagonal quantum numbers $\mathbf{p}, \lambda, \eta_b, b$ corresponding to momentum, helicity, and the intrinsic parity associated with the charge state b within the baryon number doublet labeled by $|B|$. The transformation properties of basis states are readily computed by knowing how the wave functions transform. For example,

$$\begin{aligned} V(C)|\chi\rangle &= \sum_{b,\lambda} \int \frac{d^3p}{E}V(C)\psi(b)\varphi_\lambda(\mathbf{p})|[m, J, |B|]\mathbf{p}, \lambda, \eta_b, b\rangle \\ &= \sum_{b,\lambda} \int \frac{d^3p}{E}\psi(n - b)\varphi_\lambda(\mathbf{p})|[m, J, |B|]\mathbf{p}, \lambda, \eta_b, b\rangle \\ &= \sum_{b',\lambda} \int \frac{d^3p}{E}\psi(b')\varphi_\lambda(\mathbf{p})|[m, J, |B|]\mathbf{p}, \lambda, \eta_{n-b'}, n - b'\rangle \end{aligned} \tag{18}$$

implies $V(C)|[m, J, |B|]\mathbf{p}, \lambda, \eta_b, b\rangle = |[m, J, |B|]\mathbf{p}, \lambda, \eta_{n-b}, n - b\rangle$. The relation of η_b to η_{n-b} may be determined by evaluating

$$\begin{aligned} V(\pi)|\chi\rangle &= \sum_{b,\lambda} \int \frac{d^3p}{E}V(\pi)\psi(b)\varphi_\lambda(\mathbf{p})|[m, J, |B|]\mathbf{p}, \lambda, \eta_b, b\rangle \\ &= \sum_{\lambda} \int \frac{d^3p}{E}(-1)^J[\eta\psi(B)\varphi_{-\lambda}(-\mathbf{p})|[m, J, |B|] \\ &\quad \times \mathbf{p}, \lambda, \eta_B, B\rangle + g_i(C, \pi)\eta\psi(n - B)\varphi_{-\lambda}(-\mathbf{p}) \\ &\quad \times |[m, J, |B|]\mathbf{p}, \lambda, \eta_{n-B}, n - B\rangle] = \sum_{b,\lambda} \int \frac{d^3p}{E} \\ &\quad \times \psi(b)\varphi_\lambda(\mathbf{p})V(\pi)|[m, J, |B|]\mathbf{p}, \lambda, \eta_b, b\rangle. \end{aligned} \tag{19}$$

This requires $V(\pi)|[m, J, |B|]\mathbf{p}, \lambda, \eta_b, b\rangle = \eta_b(-1)^J |[m, J, |B|] - \mathbf{p}, -\lambda, \eta_b, b\rangle$, where $\eta_B = \eta$, $\eta_{n-B} = g_i(C, \pi)\eta$. Thus representations corresponding to g_1 and g_2 (g_3 and g_4) require states of baryon number B and $n - B$ to have the same (opposite) intrinsic parity. Since it is known physically that $V(\pi)V(C) = (-1)^{2J} V(C)V(\pi)$, one must choose g_1 or g_2 for boson and g_3 or g_4 for fermion particle representations. Clearly, charge conjugation relates charge states B and $n - B$ in this construction; the usual charge conjugation transformation corresponds to the special case when n is taken to be zero. By the same procedure used above, one finds $V(\theta, a, \Lambda)|[m, J, |B|]\mathbf{p}, \lambda, \eta_b, b\rangle = e^{ib\theta} e^{i\Lambda p \cdot a} \sum_{\lambda'} D^J_{\lambda'\lambda}(p, \Lambda)|[m, J, |B|]\Lambda\mathbf{p}, \lambda', \eta_b, b\rangle$, where $(p, \Lambda) \equiv \Lambda_c(\Lambda p)\Lambda\Lambda_c^{-1}(p)$ for $\Lambda_c^{-1}(p)\hat{p} = p$ is a Wigner rotation.

There remains only the problem of adding time reversal. To do this, Theorem 4.2 of Parthasarathy[8] is used; one finds

$$U(\theta, r, a, \Lambda, s) = \begin{pmatrix} V(\theta, r, a, \Lambda, s) & 0 \\ 0 & \gamma(g)KV(-\theta, r, t(a), t\Lambda t, s)K \end{pmatrix},$$
$$U(t) = \begin{pmatrix} 0 & K \\ \alpha K & 0 \end{pmatrix}, \tag{20}$$

where V_g for $g \in G^+$ is as above, K is the complex conjugation operator, and $\gamma(g) \equiv \bar{\sigma}(t,g)\bar{\sigma}(tg, t)/\sigma(t, t)$ has values over r, s.

r \ s	e	π
e	1	$\alpha\beta$
C	$g_i(C, t)$	$g_i(C, t)\alpha\beta$

(21)

U acts on basis functions of the form

$$\begin{pmatrix} \xi(r)F_\lambda(\Lambda_c) \\ \xi(r)F_\lambda(t\Lambda_c t) \end{pmatrix} \tag{22}$$

and is, in general, reducible. In order to determine under what conditions U is reducible, consider $F_\lambda(t\Lambda_c t) = \sum_{\lambda'} A_{\lambda\lambda'} F_{\lambda'}(\Lambda_c)$. If A exists, U is clearly reducible since either manifold labeled by Λ_c or $t\Lambda_c t$ is invariant under all transformations belonging to G. Suppose A exists; it is easily seen that A must satisfy $\bar{A}^2 = \alpha, AV_gA^{-1} = \gamma(g)\overline{V}_{t(g)}$ for all $g \in G^+$. This implies $AD^J(R)A^{-1} = \gamma(r, s)\bar{D}^J(R)$ for all r, s, R which requires $\gamma(r, s) = 1$ and $A_{\lambda\lambda'} = (-1)^{J+\lambda}\delta_{\lambda,-\lambda'}$.[9] Then $\bar{A}^2 = (-1)^{2J} = \alpha = \beta$ and $g_i(C, t) = 1$.

Therefore, in order that the representation space not double under time reversal, g_i must be taken to be g_1 or g_3. This means g_1 is the proper choice for integer spin systems and g_3 for half-integer spin systems due to the $U_\pi U_C = (-1)^{2J} U_C U_\pi$ relation. Notice that setting $\alpha = \beta = (-1)^{2J}$ is the usual choice made for $\omega(x_1, x_2)$.[5] With this choice of multiplier it follows that $U_{\pi t}^2 = 1$ for all values of spin. It is unclear what physical significance an additional degeneracy due to time reversal would have.

As before, wavefunctions may be defined as $\xi(r)F_\lambda(\Lambda_c) = \psi(b)\varphi_\lambda(\mathbf{p})$; $F_\lambda(t\Lambda_c t)$ is found to be $(-1)^\lambda\varphi_\lambda(-\mathbf{p})$ in a helicity basis and one finds $U(t)|[m, J, |B|]\mathbf{p}, \lambda, \eta_b, b\rangle = (-1)^\lambda|[m, J, |B|] - \mathbf{p}, \lambda, \eta_b, b\rangle$.

4. CONCLUSION

This work has been concerned with studying the possible group theoretic means of relating internal to space–time symmetries using the Lee and Wick[3] ansatz as a working basis for constructing extensions and their projective representations. It was found that the group extension approach leading to vector representations has been quite thoroughly investigated by Michel[1] and Kamber and Straumann,[2] for the case of the usual internal symmetry groups one encounters in strong interaction physics. Although relations such as $(-1)^{2J+B} = 1$ emerge from such a treatment, it is found that the very limited ability of the space–time elements to act on internal symmetry elements due to charge invariance requirements severely restricts the number and complexity of nontrivial over-all symmetry groups possible. In particular, there seems to be no physically meaningful way of introducing actions either through automorphisms or group multipliers of the connected Poincaré group elements on the internal symmetry groups. Also, the actions that do occur, generated by time reversal elements, only lead to automorphisms acting on or multipliers having values in the centers of the internal symmetry groups. This means, for example, that non-Abelian internal symmetry groups such as $SU(2)$ cannot be influenced by space–time elements except for their "trivial" centers. Thus, it would appear that the group extension approach to the problem has been pushed about as far as it can be without the introduction of a radically new basis for understanding the underlying connection between space–time and internal symmetries.

On the other hand, the investigation of the projective representations of group extensions using the Mackey theory of induced projective representations[6] leads to definite physical relations and is an area that has not yet been thoroughly worked through. It has been shown for a particular class of extensions of P by I, namely the semidirect product, that representation multipliers only connect the discrete space–time transformations with the center of I and the discrete internal symmetries such as charge conjugation. The additional possibilities of finding representation multipliers for nontrivial extensions of P by I remain to be examined.

As a specific example, projective representations of $[U_1 \circledcirc Z_2^C] \circledcirc P$ were calculated relative to a representation multiplier which generates irreducible representations in which baryon number states B and $n - B$ for n an integer are both contained. The physical implications of choosing n to be nonzero have yet to be investigated. It was found that of the sixteen mathematically possible multipliers (for a given n), a unique multiplier choice could be made by demanding the representations to be consistent with the quantum field theory and requiring the representation space not to double when time reversal invariance is imposed. Consistency with field theory is attained by requiring the relation $U_\pi U_C = (-1)^{2J} U_C U_\pi$ to hold between the operator representatives of space inversion and charge conjugation. Time reversal doubling

is prevented by specifying $U_I^2 = U_{\pi t}^2 = (-1)^{2J}$. With the multiplier thus fixed, it is found that $U_{C\pi t}^2 = 1$ in contrast to the relation $U_{C\pi t}^2 = (-1)^{2J}$ given by Lee and Wick.[3] The physical significance of a possible representation space doubling under time reversal is not clear.

* This work was supported in part by the National Science Foundation.

1 L. Michel, in *Group Theoretical Concepts and Methods in Elementary Particle Physics*, edited by F. Gürsey (Gordon and Breach, New York, 1964), Vol. 1.
2 F. Kamber and N. Straumann, Helv. Phys. Acta **37**, 563 (1964).
3 T. D. Lee and G. Wick, Phys. Rev. **148**, 1385 (1966).
4 S. Eilenberg and S. MacLane, Ann. Math. **48**, 57, 326 (1947).
5 E. Wigner, "Unitary Representations of the Inhomogeneous Lorentz Group Including Reflections," in Ref. 1.
6 G. Mackey, Acta Math. **99**, 265 (1958).
7 W. Klink, in *Lectures in Theoretical Physics*, edited by W. Brittin and K. Mahanthappa (Gordon and Breach, New York, 1969), Vol. XI-D.
8 K. Parthasarathy, Commun. Math. Phys. **15**, 305 (1969).
9 E. Wigner, *Group Theory and Its Applications to the Quantum Mechanics of Atomic Spectra* (Academic, New York, 1959).

Position Operators as "Internal" Symmetries

Russell E. Warren and William H. Klink
Department of Physics and Astronomy, University of Iowa, Iowa City, Iowa 52240
(Received 8 September 1971)

Space–time variables are generated as representation labels of an underlying group, the group itself being combined with the Poincaré group in a manner reminiscent of the way in which internal symmetries are combined with the Poincaré group. After representations of the group are found, a transform is introduced which allows one to pass from spinor to Wigner wavefunctions in a boost independent manner, exhibiting clearly the spin dependence of the wavefunctions.

I. INTRODUCTION

The construction of relativistic covariant position operators is a well-known problem. As noted by Fleming and others,[1] one of the main difficulties associated with position operators has been the treatment of the time part of the position operator as a c-number. Recently, several papers have appeared[2,3] which attempt to define covariant position operators in a Lie-algebraic sense in which no c-number distinction is made between the space and time components. In these papers, one postulates a Lie algebra having certain desirable features, usually justified on nonrelativistic or classical relativistic grounds. In this paper we wish to justify the construction of position operators on a somewhat different basis, and then show that our general assumptions lead, in the simple case discussed here, to a Lie algebra of the same sort and having the same mass spectrum difficulties as those previously mentioned.

We would like to start with the following postulates for generating a position operator group theoretically:

1. The group elements (labeled by α_μ), whose infinitesimal operators are the position operators, form an invariant subgroup of the over-all symmetry group G. This requirement guarantees that α transformations, acting on wavefunctions $\varphi(x)$ will only generate phases. The requirement of an invariant subgroup is reminiscent of internal symmetries,[4] with the important difference that whereas internal symmetry labels seem to be superselected,[5] position variables are not; the result of this is when, in Sec. III, we go from a position to a momentum representation, the global α transformations will not merely generate a phase but will act as momentum translations. Here the point of requiring the α to form an invariant subgroup is motivated by the distinction between global space–time transformations, such as Poincaré transformations in which both the generators and the transformations themselves are physically meaningful, and "internal" symmetry transformations in which only the generators are meaningful and the global transformations seem to make no sense physically.

2. We also demand that the eigenvalues of the position operator X_μ transform as $x' = \Lambda x + a$, so that one can associate them with Minkowski space–time points. The manner in which this requirement is satisfied arises from the orbits generated from representations of the "α" subgroup.

3. Finally, we demand that only positive M^2 representations of the Poincaré group appear. In the simple model we investigate, this third requirement will be violated, as it is violated in the other models. Basically, the source of difficulty here, as pointed out by Noga,[6] is the commutator $[X_\mu, P_\nu]$; taking the simplest choice for this commutator leads to the mass difficulties.

In the usual investigation of space–time symmetry, one begins with the Minkowski space–time manifold M on which a bilinear form $\tau_{xy} = (x-y)^2$ is defined, and asks for the group of transformations leaving τ invariant. This leads to a consideration of the Poincaré group. Here a different viewpoint is suggested in which a symmetry group G containing the Poincaré group is made the starting point. G will be taken to be the semidirect product of the Poincaré group P with an abelian group $\mathfrak{A}$ of four vectors α_μ. The elements of $\mathfrak{A}$ will be seen to play the role of translations in momentum space. It will be shown that a representation multiplier for G can be so chosen that the space–time manifold M appears as the representation labels of a subgroup of G in the same way that momentum arises as representation labels of the translation subgroup of the Poincaré group.

The advantage of this viewpoint lies in the symmetrical relationship between the generators correspond-

is prevented by specifying $U_I^2 = U_{\pi t}^2 = (-1)^{2J}$. With the multiplier thus fixed, it is found that $U_{C\pi t}^2 = 1$ in contrast to the relation $U_{C\pi t}^2 = (-1)^{2J}$ given by Lee and Wick.[3] The physical significance of a possible representation space doubling under time reversal is not clear.

* This work was supported in part by the National Science Foundation.

1 L. Michel, in *Group Theoretical Concepts and Methods in Elementary Particle Physics*, edited by F. Gürsey (Gordon and Breach, New York, 1964), Vol. 1.
2 F. Kamber and N. Straumann, Helv. Phys. Acta **37**, 563 (1964).
3 T. D. Lee and G. Wick, Phys. Rev. **148**, 1385 (1966).
4 S. Eilenberg and S. MacLane, Ann. Math. **48**, 57, 326 (1947).
5 E. Wigner, "Unitary Representations of the Inhomogeneous Lorentz Group Including Reflections," in Ref. 1.
6 G. Mackey, Acta Math. **99**, 265 (1958).
7 W. Klink, in *Lectures in Theoretical Physics*, edited by W. Brittin and K. Mahanthappa (Gordon and Breach, New York, 1969), Vol. XI-D.
8 K. Parthasarathy, Commun. Math. Phys. **15**, 305 (1969).
9 E. Wigner, *Group Theory and Its Applications to the Quantum Mechanics of Atomic Spectra* (Academic, New York, 1959).

Position Operators as "Internal" Symmetries

Russell E. Warren and William H. Klink
Department of Physics and Astronomy, University of Iowa, Iowa City, Iowa 52240
(Received 8 September 1971)

Space–time variables are generated as representation labels of an underlying group, the group itself being combined with the Poincaré group in a manner reminiscent of the way in which internal symmetries are combined with the Poincaré group. After representations of the group are found, a transform is introduced which allows one to pass from spinor to Wigner wavefunctions in a boost independent manner, exhibiting clearly the spin dependence of the wavefunctions.

I. INTRODUCTION

The construction of relativistic covariant position operators is a well-known problem. As noted by Fleming and others,[1] one of the main difficulties associated with position operators has been the treatment of the time part of the position operator as a c-number. Recently, several papers have appeared[2,3] which attempt to define covariant position operators in a Lie-algebraic sense in which no c-number distinction is made between the space and time components. In these papers, one postulates a Lie algebra having certain desirable features, usually justified on nonrelativistic or classical relativistic grounds. In this paper we wish to justify the construction of position operators on a somewhat different basis, and then show that our general assumptions lead, in the simple case discussed here, to a Lie algebra of the same sort and having the same mass spectrum difficulties as those previously mentioned.

We would like to start with the following postulates for generating a position operator group theoretically:

1. The group elements (labeled by α_μ), whose infinitesimal operators are the position operators, form an invariant subgroup of the over-all symmetry group G. This requirement guarantees that α transformations, acting on wavefunctions $\varphi(x)$ will only generate phases. The requirement of an invariant subgroup is reminiscent of internal symmetries,[4] with the important difference that whereas internal symmetry labels seem to be superselected,[5] position variables are not; the result of this is when, in Sec. III, we go from a position to a momentum representation, the global α transformations will not merely generate a phase but will act as momentum translations. Here the point of requiring the α to form an invariant subgroup is motivated by the distinction between global space–time transformations, such as Poincaré transformations in which both the generators and the transformations themselves are physically meaningful, and "internal" symmetry transformations in which only the generators are meaningful and the global transformations seem to make no sense physically.

2. We also demand that the eigenvalues of the position operator X_μ transform as $x' = \Lambda x + a$, so that one can associate them with Minkowski space–time points. The manner in which this requirement is satisfied arises from the orbits generated from representations of the "α" subgroup.

3. Finally, we demand that only positive M^2 representations of the Poincaré group appear. In the simple model we investigate, this third requirement will be violated, as it is violated in the other models. Basically, the source of difficulty here, as pointed out by Noga,[6] is the commutator $[X_\mu, P_\nu]$; taking the simplest choice for this commutator leads to the mass difficulties.

In the usual investigation of space–time symmetry, one begins with the Minkowski space–time manifold M on which a bilinear form $\tau_{xy} = (x-y)^2$ is defined, and asks for the group of transformations leaving τ invariant. This leads to a consideration of the Poincaré group. Here a different viewpoint is suggested in which a symmetry group G containing the Poincaré group is made the starting point. G will be taken to be the semidirect product of the Poincaré group P with an abelian group $\mathfrak{A}$ of four vectors α_μ. The elements of $\mathfrak{A}$ will be seen to play the role of translations in momentum space. It will be shown that a representation multiplier for G can be so chosen that the space–time manifold M appears as the representation labels of a subgroup of G in the same way that momentum arises as representation labels of the translation subgroup of the Poincaré group.

The advantage of this viewpoint lies in the symmetrical relationship between the generators correspond-

ing to position and momentum which arises; it will be seen that the representation multiplier, besides leading to a physically reasonable appearance of the space–time manifold as representation labels of $\mathfrak{A}$, is also responsible for generating the commutation relations between the position and momentum generators. On the one hand, this approach is closely related to the work of Johnson,[2] who defined position operators by forming an extension of the Poincaré algebra by the position elements X_μ and found representations of this extension by unitary operators acting on linear vector spaces.

On the other hand, the approach used here is quite similar to the procedure used in a previous investigation[4] for relating space–time to internal symmetries by finding projective representations of group extensions of the Poincaré group by an internal symmetry group. In this case the "internal" symmetry group is $\mathfrak{A}$ and $G/\mathfrak{A} = P$. As in this previous study, the work here depends heavily on the Mackey[7] theory of induced projective representations of group extensions. A brief summary of some of the results of this theory used here can be found in Ref. 4; however, the original Mackey paper should be consulted for anything more than a superficial discussion of the mathematical basis of this analysis.

In Sec. II projective representations of G are found using the Mackey theory. It is found that these representations are a global extension of those determined by Johnson by splitting his extended algebra into the Heisenberg and Lorentz algebras; the work here proceeds directly in terms of the groups themselves and not their associated algebras. The representations are defined on configuration space basis functions and state vectors and constitute a generalization to arbitrary Lorentz group representation labels of the single spin covariant representations found by Weinberg.[8]

In order to make contact with the usual noncovariant Poincaré group representations of Wigner,[9] the Poincaré representational content of an irreducible representation of G is determined in Sec. III using Mackey's subgroup theorem. It is discovered that irreducible representations of G have a continuous mass content (as they must because of the O'Raifeartaigh[10] theorem) and that representations in which space–time variables are diagonal are necessarily indeterminant in mass so that physically meaningful states must have a finite extension in configuration space.

Therefore, in suggesting a new way in which space and time arise, namely from a more fundamental underlying symmetry, it is possible to derive group theoretically the connection between spinor wavefunctions and Wigner wavefunctions even for infinite component fields. This is done within the context of projective representations of a group which combines in a simple manner the usual Poincaré transformations with momentum space translations generated by position operators. The relation between the position operators and the Poincaré algebra implied by our choice of G, while appearing to be the simplest choice consistent with the Heisenberg commutation relations, leads to the appearance of unphysical mass states and thus indicates the need for external constraints or a more complicated choice of G.

II. PROJECTIVE REPRESENTATIONS OF G

Consider the group $G = \mathfrak{A} \circledcirc P$ with elements $g = (\alpha, a, \Lambda, \pi)$, where $\alpha \in \mathfrak{A}$, $(a, \Lambda) \in P_0$, the connected Poincaré group, and $\pi \in Z_2^\pi$, the space inversion group. P is taken to include space inversion in order to obtain spinor representations; for example, the Dirac spinors for spin $\frac{1}{2}$ particles are necessarily four component objects due to the requirement of space inversion invariance. G has the combination law

$$g_1 g_2 = (\alpha_1 + \Lambda_1 \pi_1(\alpha_2), a_1 + \Lambda_1 \pi_1(a_2), \Lambda_1 \pi_1 \Lambda_2 \pi_1, \pi_1 \pi_2). \tag{1}$$

Elements $\alpha \in \mathfrak{A}$ are four vectors which behave under automorphisms induced by P exactly as the translations a do.

The goal here is the construction of multiplier representations of G; that is we want a unitary representation of G such that

$$U_{g_1} U_{g_2} = \sigma(g_1, g_2) U_{g_1 g_2} \tag{2}$$

for all g_1, g_2 belonging to G. A multiplier σ for G is given by

$$\sigma(g_1, g_2) = e^{-i\Lambda_1 \pi_1(\alpha_2)\cdot a_1}. \tag{3}$$

By evaluating

$$U_{g_1} U_{g_2} U_{g_3} \text{ as } [U_{g_1} U_{g_2}] U_{g_3} \text{ and } U_{g_1} [U_{g_2} U_{g_3}],$$

it is easily checked that σ is, in fact, a multiplier of G. Multiplier representations for G are determined using Mackey's[7] theory of induced projective representations of group extensions. Representations of G are found by first finding representations of a normal subgroup of G. Take this to be $\mathfrak{A}$. $\mathfrak{A}$ has irreducible representations labeled by a four vector, x_0,

$$^{x_0}L_\alpha = e^{ix_0\cdot\alpha}.$$

The orbit of $^{x_0}L_{\alpha'}$ is found by finding all nonequivalent representations $^{x}L_{\alpha'}$ as[4]

$$\begin{aligned} {}^{x}L_{\alpha'} &= \frac{\sigma(g\alpha', g^{-1})\sigma(g, \alpha')}{\sigma(g^{-1}, g)}\, {}^{x_0}L_{g\alpha' g^{-1}} \\ &= e^{-i\Lambda\pi\alpha'\cdot a}\, {}^{x_0}L_{\Lambda\pi\alpha'} \\ &= e^{i\pi\Lambda^{-1}(x_0 - a)\cdot\alpha'}, \end{aligned} \tag{4}$$

where $g = (\alpha, a, \Lambda, \pi)$. Thus $x = \pi\Lambda^{-1}(x_0 - a)$ and there is a one-to-one correspondence between translations a and the representation labels x. Thus two space–time points x and x' are related by $x' = \Lambda x + a$ in accordance with the second postulate; this is the justification for calling x a space–time point. If the standard vector x_0 is taken to be the null vector, the inducing subgroup H is found to be

$$H = \mathfrak{A} \circledcirc [SL(2, C) \circledcirc Z_2^\pi]. \tag{5}$$

σ representations of H are found in the Mackey theory by first finding all multiplier representations of $H/\mathfrak{A} = SL(2, C) \circledcirc Z_2^\pi$. There are only vector representations of $H/\mathfrak{A}$ and these are given in terms of irreducible representations $D^{s_1 s_2}(\Lambda)$ of $SL(2, C)$[11] by

$$N(\Lambda, e) = \begin{pmatrix} D^{s_1 s_2}(\Lambda) & 0 \\ 0 & D^{s_1 s_2}(\pi\Lambda\pi) \end{pmatrix}, \quad N(e, \pi) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{6}$$

This representation is reducible if and only if $s_1 = s_2$. σ representations of H are now given by $N \otimes M$, where M is a σ representation of H which reduces to I on $\mathfrak{A}$. Due to the choice of standard vector x_0,

$$M(\alpha, \Lambda, \pi) = e^{ix_0 \cdot \alpha} I_{\Lambda,\pi}, \tag{7}$$

where $I_{\Lambda,\pi}$ is the identity representation of $SL(2, C)$ @ Z_2^{π} will work. The induced representation over right-coset representatives taken to be the translations a is

$$\begin{aligned} &U(\alpha, a, \Lambda, \pi) F_\sigma(a_0) \\ &\quad = \sigma(a_0, (\alpha, a, \Lambda, \pi))\, F_\sigma[(0, a_0, e, e)(\alpha, a, \Lambda, \pi)] \\ &\quad = e^{i\alpha \cdot a_0} F_\sigma[(\alpha, 0, \Lambda, \pi)(0, \pi\Lambda^{-1}(a + a_0), e, e)] \\ &\quad = e^{i\alpha \cdot a_0} \sum_{\sigma'} N^{s_1 s_2}_{\sigma\sigma'}(\Lambda, \pi) F_{\sigma'}[\pi\Lambda^{-1}(a + a_0)]. \end{aligned} \tag{8}$$

Due to the one-to-one correspondence between x and a, one may define representations over functions $\varphi_\sigma(x) = \varphi_\sigma(x_0 + a_0) \equiv F_\sigma(a_0)$ as

$$U(\alpha, a, \Lambda, \pi)\varphi_\sigma(x) = e^{i\alpha \cdot x} \sum_{\sigma'} N^{s_1 s_2}_{\sigma\sigma'}(\Lambda, \pi)\varphi_{\sigma'}(\pi\Lambda^{-1}(x + a)). \tag{9}$$

This is the desired representation over configuration space-wave functions. Representations defined on (nonnormalizable) state vectors may be found by setting

$$|\varphi\rangle = \sum_\sigma \int d^4x \varphi_\sigma(x) |[s_1 s_2]x, \sigma\rangle, \tag{10}$$

where the vector $|[s_1 s_2]x, \sigma\rangle$ is labeled by the irreducible representation labels (s_1, s_2) arising from representations of the Lorentz group and the diagonal quantum numbers x, σ representing the position four vector and the (s_1, s_2) representation component. Then

$$\begin{aligned} &U(\alpha, a, \Lambda, \pi)|\varphi\rangle \\ &\quad = \sum_\sigma \int d^4x U(\alpha, a, \Lambda, \pi)\varphi_\sigma(x) |[s_1 s_2]x, \sigma\rangle \\ &\quad = \sum_\sigma \int d^4x\, e^{i\alpha \cdot x} \sum_{\sigma'} N^{s_1 s_2}_{\sigma\sigma'}(\Lambda, s)\varphi_{\sigma'}[\pi\Lambda^{-1}(x - a)] \\ &\qquad \times |[s_1, s_2]x, \sigma\rangle \\ &\quad = \sum_{\sigma'} \int d^4x'\, e^{i\alpha \cdot (\Lambda\pi x' + a)} \sum_\sigma N_{\sigma\sigma'}(\Lambda, \pi)\varphi_{\sigma'}(x') \\ &\qquad \times |[s_1 s_2]\Lambda\pi x' + a, \sigma\rangle \end{aligned} \tag{11}$$

$$\begin{aligned} &U(\alpha, a, \Lambda, \pi)|[s_1 s_2]x, \sigma\rangle \\ &\quad = e^{i\alpha \cdot (\Lambda\pi x + a)} \sum_{\sigma'} N^{s_1 s_2}_{\sigma'\sigma}(\Lambda, \pi) |[s_1 s_2]\Lambda\pi x + a, \sigma'\rangle. \end{aligned} \tag{12}$$

Define the Hermitian operators P_μ and X_μ by setting

$$U(a) = e^{iP \cdot a}, \quad U(\alpha) = e^{iX \cdot \alpha}. \tag{13}$$

The relation

$$U(\alpha)U(a) = e^{i\alpha \cdot a} U(a)U(\alpha) \tag{14}$$

then implies

$$P_\mu X_\nu - X_\nu P_\mu = i g_{\mu\nu}. \tag{15}$$

These are the commutation relations between P_μ and X_μ, regarded now as a position operator, postulated by Johnson.[2] However, as mentioned in the introduction it has been shown by Johnson[2] and Noga[6] that representations determined relative to these commutation relations result in the appearance of unphysical masses so that the algebra of G cannot be regarded as a physically valid combination of position and the Poincaré observables.

Clearly, the representation above coincides with the transformation law for Dirac spinors by choosing $(s_1, s_2) = (\frac{1}{2}, 0)$, since $D^{1/2,0}(\pi\Lambda\pi) \cong D^{0,1/2}(\Lambda)$ making $N^{1/2,0}(\Lambda, e)$ into the usual $S(\Lambda)$ matrix of the Dirac theory.[12] In this way one obtains group theoretically a representation of a group containing the Poincaré group whose basis states transform covariantly in contrast to the usual Poincaré states which transform in a way that depends on the momentum variable.[9] As noted by Weinberg[8] and others,[13] covariantly transforming Poincaré states are essential to the construction of Poincaré invariant S-matrix elements.

The representation just constructed is a global expression of the algebraic representation found by Johnson[2] by postulating an extension of the Poincaré algebra by X_μ such that $[P_\mu, X_\nu] = ig_{\mu\nu}$ is satisfied. The irreducible Poincaré content (i.e., mass and spin content) of this representation as well as the consequences of taking $\alpha \neq 0$ will be examined in the next section.

III. POINCARÉ CONTENT OF REPRESENTATIONS OF G

We now wish to make the connection between the representations of G and the Poincaré group. For simplicity only the connected part of P, P_0 will be considered. To do this Mackey's subgroup theorem[14] will be used to examine the breakup of representations of G into subspaces invariant with respect to transformations of the Poincaré group. This is done by making a double coset decomposition of G with respect to $H = \mathfrak{A} \,@\, SL(2, C)$, the inducing subgroup of G, and $H' = P_0$,

$$G = \bigcup_D H g_D H', \tag{16}$$

where g_D is a double coset representative. It is easy to see that there is only one double coset so g_D can be taken to be the identity. This indicates that the representation spaces of P_0 and G coincide. The group inducing representations of P_0 within this representation space is found to be $SL(2, C)$. Since $SL(2, C)$ will induce reducible representations of P_0 in general, the problem is to find the irreducible Poincaré content of representations of G. The representation labels of G are just the $SL(2, C)$ labels (s_1, s_2) while P_0 representations are specified by (M, J), the mass and spin. Since irreducible representations of P_0 are induced by $T \,@\, SU(2)$, basis functions over right coset elements Λ_c[14], $f_\sigma(\Lambda_c)$, transforming irreducibly under induced representations of P_0 must have the property $f(h\Lambda_c) = \mathfrak{K}(h) f(\Lambda_c)$, where

$\mathcal{K}(h)$ is a representation of $T \otimes SU(2)$. Basis functions defined over the translations $F(a)$ of representations of P_0 induced by $SL(2,C)$ must likewise have the property $F(\Lambda a) = D^{s_1 s_2}(\Lambda)F(a)$. It can be seen that the transform given by Hermann,[15]

$$f(\Lambda_c) = \int d^4a dR\, \mathcal{K}[(a,R)^{-1}]F[(a,R)\cdot\Lambda_c] \tag{17}$$

relates f and F in a manner consistent with the above requirements. This is done by noting that $d^4a dR$ is defined to be an invariant measure over the group manifold $T \otimes SU(2)$, so that

$$\begin{aligned} f(h'\Lambda_c) &= \int dh \mathcal{K}(h^{-1})F(hh'\Lambda_c) \\ &= \int dh'' \mathcal{K}(h'h''^{-1})F(h''\Lambda_c) \\ &= \mathcal{K}(h')f(\Lambda_c). \end{aligned} \tag{18}$$

The reducible representation $\mathcal{K}$ is taken to be $\mathcal{K}(a,R) = e^{i\hat{p}\cdot a}D^{s_1 s_2}(R)$, where $\hat{p}$ is $\binom{M}{0}$ for $M \neq 0$ representations. Then

$$\begin{aligned} f_{Jm}(\Lambda_c) &= \int d^4a dR\, e^{-i\hat{p}\cdot R^{-1}a} \\ &\quad \sum_{J'm'} D^{s_1 s_2}_{JmJ'm'}(R^{-1})F_{J'm'}((a,R)(0,\Lambda_c)) \\ &= \int d^4a dR\, e^{-i\hat{p}\cdot R^{-1}a} \sum_{J'm'} D^{s_1 s_2}_{JmJ'm'}(R^{-1}) \\ &\quad \times \sum_{J''m''} D^{s_1 s_2}_{J'm'J''m''}(R\Lambda_c)F_{J''m''}((R\Lambda_c)^{-1}a) \\ &= \sum_{J''m''} D^{s_1 s_2}_{JmJ''m''}(\Lambda_c) \int d^4a' e^{-\Lambda_c^{-1}\hat{p}\cdot a'}F_{J''m''}(a') \\ &= \sum_{J'm'} D^{s_1 s_2}_{JmJ'm'}(\Lambda_c)\psi_{J'm'}(\Lambda_c^{-1}\hat{p}). \end{aligned} \tag{19}$$

$\psi_{J'm'}(\Lambda_c^{-1}\hat{p})$ is just a Fourier transform to momentum variables of the basis function over configuration space corresponding to an arbitrary mass M arising from $\hat{p}$. Identifying $f_{Jm}(\Lambda_c)$ as $\varphi^J_m(B(p)\hat{p}) = \varphi^J_m(p)$, a Wigner basis function of spin J, mass M, where $B(p)$ is an arbitrary boost from the rest frame momentum $\hat{p}$ to momentum p such that $B(p)\hat{p} = p$ and is equivalent to the coset representative Λ_c^{-1}, leads to the relation

$$\psi^{s_1 s_2}_{Jm}(p) = \sum_{J'm'} D^{s_1 s_2}_{JmJ'm'}(B(p))\varphi^{J'}_{m'}(p). \tag{20}$$

This demonstrates the Poincaré spin content of an irreducible representation of G labeled by (s_1, s_2) and is a generalization to an arbitrary $SL(2,C)$ representation of the relation obtained by Weinberg[8] for the $(s, 0)$ case.

Using the transform above it is easy to compute the action of $U(\alpha)$ on $\varphi^{M,J}_m(p_M)$, the Wigner[9] basis function of mass M, spin J. The result is

$$\begin{aligned} &U(\alpha)\varphi^{M,J}_m(p_M) \\ &= \sum_{J',m'} D^{s_1 s_2}_{JmJ'm'}[(B(p_M)^{-1}B(p_M + \alpha)]\varphi^{M',J'}_{m'}(p_{M'}), \end{aligned} \tag{21}$$

where $p_M = \Lambda_c^{-1}\hat{p}_M$, $p_{M'} = \Lambda_c^{-1}\hat{p}_M + \alpha = \overline{\Lambda}_c^{-1}\hat{p}_{M'}$. Thus the irreducible representation space of G for a given (s_1, s_2) is a reducible representation space of P_0 having a continuous mass content and a spin content dependent on (s_1, s_2).

The action of $U(\alpha)$ on basis states $|[s_1 s_2]x, \sigma\rangle$ reflects the observation of Johnson[2] and others on the lack of localizability of mass eigenstates;

$$U(\alpha)|[s_1 s_2]x, \sigma\rangle = e^{ix\cdot\alpha}|[s_1 s_2]x, \sigma\rangle \tag{22}$$

implies that a sharply localized state is physically unchanged by a transformation which changes its mass. Therefore, states sharp in x have a complete ambiguity in mass. There is, besides the appearance of a continuum of masses, the more serious problem of the appearance of negative and imaginary masses since M^2 is unrestricted. This undesirable feature limits the physically sensible representation spectrum of G to only a subset of those representations allowed mathematically. From the set of wavefunctions $\psi^{s_1 s_2}_{Jm}(x)$ having continuously varying mass, one can select a particular mass wavefunction transforming according to Eq. (9) with $\alpha = 0$ by applying the Klein–Gordon operator $\Box^2 - M^2$,

$$(\Box^2 - M^2)\psi^{s_1 s_2}_{Jm}(x) = 0. \tag{23}$$

Adding on the Klein–Gordon operator does away with the unphysical masses which arise, but at the cost of breaking the group symmetry. However, the Klein–Gordon constraint allows one to use the transform between spinor states and Wigner wavefunctions for completely arbitrary internal spin and arbitrary boosting operation, in contrast to transforms which are often given only for a specific spin content.

IV. CONCLUSION

The point of view adopted here of postulating an underlying symmetry group G in which space–time appears as representation labels leads to several interesting results. When projective representations are calculated relative to a multiplier chosen so as to lead to all space–time points appearing as representation labels of a subgroup of G, it is found that the commutation relations $[P_\mu, X_\nu] = ig_{\mu\nu}$ between the position and momentum operators in the Lie algebra of G are automatically satisfied. Furthermore, it is found that the representations of G on wavefunctions that are defined over position coordinates have the same transformation properties as the usual covariant spinor representations one associates with the solutions of the Klein–Gordon or Dirac equations. Finally, using Mackey's subgroup theorem, the connection between the spinor representations of G and the usual noncovariant Poincaré group representations was established as a relation independent of the choice of boost made. It is found that the representations of G have a continuous mass spectrum and that within the context of the symmetry group G, sharp localization in space–time requires a complete indeterminacy in mass so that physical states of a well defined mass must have a nonzero extension in space–time. The appearance of unphysical masses in the representation spectrum of G requires that one only consider a subset of the mathematically possible representations. Particular mass states may be selected from the continuum by applying the Klein–Gordon operator as a constraint on the representation space. Finally, the group elements α were interpreted simply as translating the momentum. There has been work done in the last few years[16] in attempting to extend the applicability of the Poincaré group to include symmetries in the presence of electromagnetic fields. It is hoped that the approach indicated here of constructing projective representations of extensions of the Poincaré group using the Mackey theory[7] can be

widened to include extensions by an arbitrary function of space–time $\alpha_\mu(x)$, that has vector transformation properties under the Poincaré group. Interpreting $\alpha_\mu(x)$, then, as the electromagnetic potential would open up possibilities for investigating group theoretically subject areas closely related to the work of Yang and Mills[17] on space–time dependent gauge transformations.

Further, there are other ways of combining the Poincaré group with the "position operator" group, for example, by also bringing in internal symmetries which may have the virtue of generating a more physical mass spectrum; that is, by changing the commutation relations between X_μ and P_ν. These topics will be dealt with in future publications.

[1] G. N. Fleming, Phys. Rev. **137**, B188 (1965) and references cited therein.
[2] J. Johnson, Phys. Rev. **181**, B1755 (1969).
[3] J. J. Aghassi, P. Roman, and R. M. Santilli, Phys. Rev. D **1**, 2753 (1970) and H. E. Moses, Ann. Phys. **52**, 444 (1969); see also L. Castell, Nuovo Cimento **49**, 285 (1967).
[4] R. Warren and W. Klink, J. Math. Phys. 13, 306 (1972).
[5] R. Mirman, Phys. Rev. D **1**, 3349 (1970) and references cited therein.
[6] M. Noga, Phys. Rev. D **2**, 304 (1970).
[7] G. Mackey, Acta Math. **99**, 265 (1958).
[8] S. Weinberg, Phys. Rev. **133**, B1318 (1964).
[9] E. Wigner, Ann. Math. **40**, 149 (1939).
[10] L. O'Raifeartaigh, Phys. Rev. **139**, B1052 (1965).
[11] I. M. Gel'fand, R. A. Minlos, and Z. Yu Shapiro, *Representations of the Rotation and Lorentz Groups and Their Applications* (MacMillan, New York, 1963).
[12] S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1961).
[13] P. Matthews, in *Lectures in Theoretical Physics*, edited by A. O. Barut and W. E. Brittin (Gordon and Breach, New York, 1967), Vol. X-A.
[14] W. Klink, in *Lectures in Theoretical Physics*, edited by W. Brittin and K. Mahanthappa (Gordon and Breach, New York, 1969), Vol. XI-D.
[15] R. Hermann, *Lie Groups for Physicists* (Benjamin, New York, 1966), p. 110.
[16] H. Bacry, P. Combe, and J. L. Richard, Nuovo Cimento **67A**, 267 (1970).
[17] R. L. Mills and C. N. Yang, Phys. Rev. **96**, 191 (1954).

On the Correlations of the Resonance Parameters for the Overlapping Resonances

K. R. Sandhya Devi
Tata Institute of Fundamental Research, Bombay-5, India

and

Nazakat Ullah*
Department of Physics, University of Toronto, Toronto 181, Canada

(Received 23 July 1970)

A statistical study of the correlations of the complex poles of the unitary collision matrix is carried out. It is shown that both for the elastic and the inelastic scattering the correlation coefficient of the two total widths is always very small. A simple relation satisfied by the correlation coefficient of the real parts of the complex poles is given. The distribution of the single width is calculated and compared with the Porter–Thomas distribution and the one obtained by a numerical calculation. Some other interesting results, like the energy correlation function for the purely elastic scattering cross section and a relation satisfied by the resonance parameters for the fluctuation calculation, are also given.

I. INTRODUCTION

In the last couple of years there has been considerable interest in the problem of overlapping resonances. This interest arises due to the study of Ericson's fluctuations,[1] intermediate structure[2,3] and the other low energy nuclear reactions, which pass through the formation of a compound nucleus, in which the compound nucleus resonances are not well separated. The main difficulty which arises when the resonances start interfering is that the parameters of the scattering matrix do not remain independent because of the unitarity constraint. Even though a number of models have been constructed recently which satisfy the unitarity constraint, not much work has been done so far as the statistical study of the resonance parameters of the interfering resonances is concerned. Such work is needed, e.g., when one has either to average over the resonance parameters of the scattering matrix[3] or to justify certain of the statistical assumptions which are used in the evaluation of the expressions for the average of the cross section and its fluctuation around the mean.[1]

For the purely elastic scattering case, a generalized distribution of the poles of the unitary scattering function has been given recently,[4] which can be used to study the important correlations between the resonance parameters. The same is not true when more than one channel is open, since the corresponding multichannel distribution for the poles of the unitary scattering matrix is not at all easy to work out. In this paper, we would like to show that a number of identities can be exploited to study the statistical correlations for the multichannel case, without going into the problem of the multichannel distribution of the poles of the unitary scattering matrix. To check our results and to show what kind of new correlations arise when the resonances are not isolated, we shall compare them with the well-known parameters of the R-matrix theory.[5,6] Some of these results will also be compared with the ones which have been obtained by Moldauer[7] using numerical calculations.

We describe the general formulation in Sec. II. In Sec. III we study the statistical properties of the resonance parameters for the purely elastic scattering case. The multichannel results are presented in Sec. V.

II. GENERAL FORMULATION

It has been shown recently[8] that the unitary scattering matrix S, based on R-matrix theory[9] or Feshbach's unified theory,[10] can always be written in the form

$$S = V\left(1 - i \sum_{\lambda,\mu=1}^{N} (X_\lambda \times X_\mu) A_{\lambda\mu}\right) V, \tag{1a}$$

widened to include extensions by an arbitrary function of space–time $\alpha_\mu(x)$, that has vector transformation properties under the Poincaré group. Interpreting $\alpha_\mu(x)$, then, as the electromagnetic potential would open up possibilities for investigating group theoretically subject areas closely related to the work of Yang and Mills[17] on space–time dependent gauge transformations.

Further, there are other ways of combining the Poincaré group with the "position operator" group, for example, by also bringing in internal symmetries which may have the virtue of generating a more physical mass spectrum; that is, by changing the commutation relations between X_μ and P_ν. These topics will be dealt with in future publications.

[1] G. N. Fleming, Phys. Rev. **137**, B188 (1965) and references cited therein.
[2] J. Johnson, Phys. Rev. **181**, B1755 (1969).
[3] J. J. Aghassi, P. Roman, and R. M. Santilli, Phys. Rev. D **1**, 2753 (1970) and H. E. Moses, Ann. Phys. **52**, 444 (1969); see also L. Castell, Nuovo Cimento **49**, 285 (1967).
[4] R. Warren and W. Klink, J. Math. Phys. 13, 306 (1972).
[5] R. Mirman, Phys. Rev. D **1**, 3349 (1970) and references cited therein.
[6] M. Noga, Phys. Rev. D **2**, 304 (1970).
[7] G. Mackey, Acta Math. **99**, 265 (1958).
[8] S. Weinberg, Phys. Rev. **133**, B1318 (1964).
[9] E. Wigner, Ann. Math. **40**, 149 (1939).
[10] L. O'Raifeartaigh, Phys. Rev. **139**, B1052 (1965).
[11] I. M. Gel'fand, R. A. Minlos, and Z. Yu Shapiro, *Representations of the Rotation and Lorentz Groups and Their Applications* (MacMillan, New York, 1963).
[12] S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1961).
[13] P. Matthews, in *Lectures in Theoretical Physics*, edited by A. O. Barut and W. E. Brittin (Gordon and Breach, New York, 1967), Vol. X-A.
[14] W. Klink, in *Lectures in Theoretical Physics*, edited by W. Brittin and K. Mahanthappa (Gordon and Breach, New York, 1969), Vol. XI-D.
[15] R. Hermann, *Lie Groups for Physicists* (Benjamin, New York, 1966), p. 110.
[16] H. Bacry, P. Combe, and J. L. Richard, Nuovo Cimento **67A**, 267 (1970).
[17] R. L. Mills and C. N. Yang, Phys. Rev. **96**, 191 (1954).

On the Correlations of the Resonance Parameters for the Overlapping Resonances

K. R. Sandhya Devi
Tata Institute of Fundamental Research, Bombay-5, India

and

Nazakat Ullah*
Department of Physics, University of Toronto, Toronto 181, Canada

(Received 23 July 1970)

A statistical study of the correlations of the complex poles of the unitary collision matrix is carried out. It is shown that both for the elastic and the inelastic scattering the correlation coefficient of the two total widths is always very small. A simple relation satisfied by the correlation coefficient of the real parts of the complex poles is given. The distribution of the single width is calculated and compared with the Porter–Thomas distribution and the one obtained by a numerical calculation. Some other interesting results, like the energy correlation function for the purely elastic scattering cross section and a relation satisfied by the resonance parameters for the fluctuation calculation, are also given.

I. INTRODUCTION

In the last couple of years there has been considerable interest in the problem of overlapping resonances. This interest arises due to the study of Ericson's fluctuations,[1] intermediate structure[2,3] and the other low energy nuclear reactions, which pass through the formation of a compound nucleus, in which the compound nucleus resonances are not well separated. The main difficulty which arises when the resonances start interfering is that the parameters of the scattering matrix do not remain independent because of the unitarity constraint. Even though a number of models have been constructed recently which satisfy the unitarity constraint, not much work has been done so far as the statistical study of the resonance parameters of the interfering resonances is concerned. Such work is needed, e.g., when one has either to average over the resonance parameters of the scattering matrix[3] or to justify certain of the statistical assumptions which are used in the evaluation of the expressions for the average of the cross section and its fluctuation around the mean.[1]

For the purely elastic scattering case, a generalized distribution of the poles of the unitary scattering function has been given recently,[4] which can be used to study the important correlations between the resonance parameters. The same is not true when more than one channel is open, since the corresponding multichannel distribution for the poles of the unitary scattering matrix is not at all easy to work out. In this paper, we would like to show that a number of identities can be exploited to study the statistical correlations for the multichannel case, without going into the problem of the multichannel distribution of the poles of the unitary scattering matrix. To check our results and to show what kind of new correlations arise when the resonances are not isolated, we shall compare them with the well-known parameters of the R-matrix theory.[5,6] Some of these results will also be compared with the ones which have been obtained by Moldauer[7] using numerical calculations.

We describe the general formulation in Sec. II. In Sec. III we study the statistical properties of the resonance parameters for the purely elastic scattering case. The multichannel results are presented in Sec. V.

II. GENERAL FORMULATION

It has been shown recently[8] that the unitary scattering matrix S, based on R-matrix theory[9] or Feshbach's unified theory,[10] can always be written in the form

$$S = V\left(1 - i \sum_{\lambda,\mu=1}^{N} (X_\lambda \times X_\mu) A_{\lambda\mu}\right) V, \tag{1a}$$

where V is a unitary and symmetric matrix which gives rise to the background scattering, X_λ is a real vector in channel space, and N is the number of resonances. The matrix elements $A_{\lambda\mu}$ of the level matrix are given by

$$(A^{-1})_{\lambda\mu} = (E - E_\lambda)\delta_{\lambda\mu} + \tfrac{1}{2}(X_\lambda, X_\mu), \tag{1b}$$

where the E_λ are the real eigenvalues of the compound nucleus Hamiltonian and (X_λ, X_μ) denotes the scalar product of the vectors X_λ and X_μ in the channel space.

The form of the unitary scattering matrix S given by expression (1) involves the inversion of the level matrix and is, therefore, suitable if the number of resonances are small. If the number of resonances are large but the number of channels is small, then the channel inversion form is suitable. This form of S can be written in the usual way as

$$S = V(1 - \tfrac{1}{2}iK)\,(1 + \tfrac{1}{2}iK)^{-1}V, \tag{2a}$$

where the real-symmetric matrix K is of the form

$$K = \sum_{\lambda=1}^{N} \frac{X_\lambda \times X_\lambda}{E - E_\lambda}, \tag{2b}$$

which is similar to the one given by Feshbach.[10]

The pole resonance form of the scattering matrix S is given by

$$S = V\left(1 - i\sum_{\lambda=1}^{N} \frac{G_\lambda \times G_\lambda}{E - Z_\lambda}\right)V \tag{3}$$

where the Z_λ are the complex poles, $Z_\lambda = \epsilon_\lambda - \frac{1}{2}i\Gamma_\lambda$, and G_λ is a complex vector in the channel space. Comparing expression (3) with either expression (1a) or (2a), we can write the relations between the complex resonance parameters Z_λ, G_λ of the scattering matrix S and the real parameters E_λ, X_λ of the R-matrix theory. Since the statistical properties of the real R-matrix theory have been very well studied,[6] these relations can be used to study the statistical properties of the new resonance parameters Z_λ, G_λ. We first consider the purely elastic scattering case in the next section.

III. ELASTIC SCATTERING

For the purely elastic scattering case the relations between the complex poles Z_λ and the real parameters X_λ, E_λ are given by the following identity in E:

$$\prod_{\lambda=1}^{N}(E - Z_\lambda) = \prod_{\lambda=1}^{N}(E - E_\lambda) + \tfrac{1}{2}i\sum_{\mu=1}^{N} X_\mu^2 \prod_{\lambda\neq\mu}^{N}(E - E_\lambda). \tag{4a}$$

In order to satisfy unitarity the complex quantities G_λ^2 are given by

$$G_\lambda^2 = \Gamma_\lambda \prod_{\mu\neq\lambda}(Z_\lambda - Z_\mu^*)(Z_\lambda - Z_\mu)^{-1}. \tag{4b}$$

Since we are dealing with the single-channel case here, we have dropped the channel index c from the amplitudes X_λ, G_λ in the above expressions. Expression (4b) has also been given by Mahaux and Weidenmüller.[11] Before we study the statistical properties of the resonance parameters Z_λ, G_λ, we choose a suitable boundary condition[9] $\mathrm{Re}[L^0(1 - R^0L^0)^{-1}] = 0$ in R-matrix theory, which allows us to use the same statistical properties for E_λ, X_λ as have been given earlier.[5,6] The case when a different boundary condition is used will not be discussed here, but can be worked out without much difficulty using the generalized distribution given in Ref. 4. In the following subsection we study the correlation coefficients of the new resonance parameters.

A. Correlation Coefficients

To study the correlation coefficients, we write the following relations, which are easily obtained using the identity (4):

$$\sum_{\mu=1}^{N} \epsilon_\mu = \sum_{\mu=1}^{N} E_\mu, \tag{5a}$$

$$\sum_{\mu=1}^{N} \Gamma_\mu = \sum_{\mu=1}^{N} X_\mu^2, \tag{5b}$$

$$\sum_{\mu<\lambda}^{N}(\epsilon_\mu\epsilon_\lambda - \tfrac{1}{4}\Gamma_\mu\Gamma_\lambda) = \sum_{\mu<\lambda}^{N} E_\mu E_\lambda, \tag{5c}$$

$$\sum_{\mu<\lambda}^{N}(\epsilon_\mu\Gamma_\lambda + \epsilon_\lambda\Gamma_\mu) = \sum_{\mu=1}^{N} X_\mu^2\left(\sum_{\lambda=\mu}^{N} E_\lambda\right). \tag{5d}$$

The correlation coefficient $C_{\Gamma\mu,\Gamma\lambda}$ between two widths $\Gamma_\mu, \Gamma_\lambda$ is defined by

$$C_{\Gamma\mu,\Gamma\lambda} = (\langle\Gamma_\mu\Gamma_\lambda\rangle - \langle\Gamma_\mu\rangle^2)/(\langle\Gamma_\mu^2\rangle - \langle\Gamma_\mu\rangle^2),$$

where the bracket sign $\langle\ \rangle$ denotes an ensemble average.

Taking the ensemble average of Eq. (5b) and its square, we get

$$\langle\Gamma_\mu\rangle = \langle X_\mu^2\rangle, \tag{6a}$$

$$\langle\Gamma_\mu^2\rangle + (N-1)\langle\Gamma_\mu\Gamma_\lambda\rangle_{\mu\neq\lambda} = \langle X_\mu^4\rangle + (N-1)\langle X_\mu^2 X_\lambda^2\rangle_{\mu\neq\lambda}. \tag{6b}$$

The ensemble averages of the real amplitudes X_μ are given by[12]

$$\langle X_\mu^4\rangle = 3N(N+2)^{-1}\langle X_\mu^2\rangle^2, \tag{7a}$$

$$\langle X_\mu^2 X_\lambda^2\rangle_{\mu\neq\lambda} = N(N+2)^{-1}\langle X_\mu^2\rangle^2. \tag{7b}$$

Putting in these ensemble averages in Eqs. (6) and using the definition of the correlation coefficient, we find that the width–width correlation coefficient is given by

$$C_{\Gamma_\mu,\Gamma_\lambda} = -(N-1)^{-1}. \tag{8}$$

Since N is large in practice, we find that the width–width correlation is quite weak even when the resonances start overlapping.

We next consider the correlation coefficient between Γ_μ and ϵ_μ. Multiplying expressions (5a), (5b) together and substracting out (5d) from it, we find the following relation between the ensemble averages of the two sets of resonance parameters:

$$\langle\epsilon_\mu\Gamma_\mu\rangle = \langle E_\mu\rangle\langle X_\mu^2\rangle,$$

This relation together with (5a), (5b) implies that the correlation coefficient of ϵ_μ, Γ_μ is zero. Even though the correlation coefficient of ϵ_μ, Γ_μ turns out to be

zero, the same is not true for the correlations of the higher powers of ϵ_μ, Γ_μ. This is reflected in the fact that the joint distribution of the quantities ϵ_μ, Γ_μ is not an independent distribution with respect to each other.[4] This is a major difference between the distribution of the resonance parameters of the scattering function S and of the real parameters E_μ, X_μ of R-matrix theory, which are always independent with respect to each other.[13]

A similar calculation can also be carried out for the correlation coefficient of $\epsilon_\mu, \epsilon_\lambda$. But it turns out that the relations (5) are not sufficient for this calculation. We need further relations which are provided by identity (4), or alternatively we can use the joint distribution[4] of ϵ_μ, Γ_μ to calculate the ensemble averages of the product quantities $\epsilon_\mu \epsilon_\lambda$. Unfortunately, no simple closed-form expressions can be obtained for these ensemble averages unless some approximations are made. Instead of these approximate expressions, we give here an exact relation satisfied by the correlation coefficient $C_{\epsilon_\mu, \epsilon_\lambda}$. To do this, we take the distribution of the real eigenvalues E_μ to be Wishart distribution[13]

$$P(E_1, E_2, \ldots, E_N) = K\left(\prod_{\mu<\lambda} |E_\mu - E_\lambda|\right) \exp\left(-(4\sigma^2)^{-1} \sum_{\mu=1}^{N} E_\mu^2\right), \quad (9)$$

where K is the normalization constant and σ^2 is the mean-square dispersion of the off-diagonal matrix elements of the compound-nucleus Hamiltonian.

Using expressions (5a), (9), we get

$$C_{\epsilon_\mu, \epsilon_\lambda} = -1/(N-1) + [2\sigma^2/(N-1)\langle \epsilon_\mu^2 \rangle]. \quad (10a)$$

The parameter σ^2 can be eliminated using the relations (5) and expression (9). If this elimination is carried out, we get the following exact relation satisfied by the correlation coefficient $C_{\epsilon_\mu, \epsilon_\lambda}$:

$$C_{\epsilon_\mu, \epsilon_\lambda} = -\frac{1}{N+1} - \frac{1}{2(N-1)(N+1)} \times \frac{\langle \Gamma_\mu^2 \rangle - \langle \Gamma_\mu \rangle^2}{\langle \epsilon_\mu^2 \rangle} + \frac{1}{2(N+1)} \frac{\langle \Gamma_\mu \rangle^2}{\langle \epsilon_\mu^2 \rangle}. \quad (10b)$$

We see from expression (10b) that if the resonances are isolated, then the correlation coefficient $C_{\epsilon_\mu, \epsilon_\lambda}$ becomes

$$C_{\epsilon_\mu, \epsilon_\lambda} = -1/(N+1), \quad (10c)$$

which is just the correlation coefficient of two real eigenvalues E_μ, E_λ, as it should be.

B. Distribution of The Single Width

The joint distribution of the complex poles Z_μ is given by[4]

$$P(\{\epsilon_\mu, \Gamma_\mu\}) \prod_\mu d\epsilon_\mu \, d\Gamma_\mu = K\delta\left(1 - \sum_{\mu=1}^{N} \frac{\Gamma_\mu}{N\langle X_\mu^2 \rangle}\right)\left(\prod_\mu \Gamma_\mu\right)^{-1/2} \times \left\{ \exp\left[-(8\sigma^2)^{-1}\left(\sum_{\mu<\lambda}^{N} \Gamma_\mu \Gamma_\lambda\right)\right]\right\}\left\{\left(\exp\left[-(4\sigma^2)^{-1} \sum_\mu \epsilon_\mu^2\right]\right) \times \prod_{\mu<\lambda} \frac{(\epsilon_\mu - \epsilon_\lambda)^2 + \frac{1}{4}(\Gamma_\mu - \Gamma_\lambda)^2}{[(\epsilon_\mu - \epsilon_\lambda)^2 + \frac{1}{4}(\Gamma_\mu + \Gamma_\lambda)^2]^{1/2}} \prod_\mu d\epsilon_\mu \, d\Gamma_\mu, \quad (11)$$

where K is the normalization constant. In writing expression (11) we have taken the distribution of the real eigenvalues E_μ to be given by expression (9) and the distribution of the real amplitudes X_μ to be the one given in Ref. 12.

One important characteristic of a statistical distribution is the mean-square deviation or dispersion. We shall now show that the dispersion of the width Γ_μ is larger than the dispersion of the quantity X_μ^2, which is the corresponding width in R matrix. To show this, we make the following approximation in expression (11). We write the term

$$\prod_{\mu<\lambda} \frac{(\epsilon_\mu - \epsilon_\lambda)^2 + \frac{1}{4}(\Gamma_\mu - \Gamma_\lambda)^2}{[(\epsilon_\mu - \epsilon_\lambda)^2 + \frac{1}{4}(\Gamma_\mu + \Gamma_\lambda)^2]^{1/2}},$$

as

$$\prod_{\mu<\lambda} (\epsilon_\mu - \epsilon_\lambda)\left[\left(1 - \frac{\Gamma_\mu \Gamma_\lambda}{(\epsilon_\mu - \epsilon_\lambda)^2 + \frac{1}{4}(\Gamma_\mu + \Gamma_\lambda)^2}\right) \times \left(1 + \tfrac{1}{4} \frac{(\Gamma_\mu - \Gamma_\lambda)^2}{(\epsilon_\mu - \epsilon_\lambda)^2}\right)\right]^{1/2},$$

and expand the quantities in square brackets by the binomial theorem. Keeping only the first term in the above expansion, we can write the unnormalized distribution of the widths as

$$P(\Gamma_\mu) \prod_\mu d\Gamma_\mu = \left(1 - \frac{\sum \Gamma_\mu}{N\langle X_\mu^2 \rangle}\right)(\Pi_\mu \Gamma_\mu)^{-1/2} \times \left(\exp\left[-(8\sigma^2)^{-1} \sum_{\mu<\lambda} \Gamma_\mu \Gamma_\lambda\right]\right) \prod_\mu d\Gamma_\mu. \quad (12)$$

For large values of N, the parameter σ^2 can be related to the average spacing d of the poles of the R matrix using Wigner's semi-circle law.[14] It is given by

$$\sigma^2 = \Pi^{-2} N d^2. \quad (13)$$

To calculate the mean-square deviation of the width Γ_μ, we put the value of σ^2 in expression (12), expand the exponential and carry out the integrations over Γ_μ. This gives us

$$\frac{\langle \Gamma_\mu^2 \rangle - \langle \Gamma_\mu \rangle^2}{\langle \Gamma_\mu \rangle^2} = \frac{3N}{N+1} \times \left(\frac{1 + \alpha N(3N+32)/(N+4)(N+6)}{1 + 3\alpha N/(N+2)}\right) - 1, \quad (14a)$$

where the parameter $\alpha = \Pi^2 \langle X_\mu^2 \rangle^2 / 16 d^2$.

The mean-square deviation of the square of the amplitude X_μ using Eq. (17a) is given by

$$(\langle X_\mu^4 \rangle - \langle X_\mu^2 \rangle^2)/\langle X_\mu^2 \rangle^2 = 2[(N-1)/(N+2)]. \quad (14b)$$

A comparison of Eqs. (14a) and (14b) shows that the width Γ_μ has a larger mean-square deviation than the square of the amplitude X_μ. This was also found to be the case for the model in which only two resonances were considered,[15] but no approximations had to be made to carry out the integrations over the variables ϵ_μ, Γ_μ.

Using the same approximations, we can also calculate the distribution of a single width Γ_μ. We would like to compare this distribution with the one obtained numerically by Moldauer.[7] In Moldauer's numerical

calculation the distribution of the amplitudes X_μ is taken to be an independent Gaussian distribution rather than the δ-function distribution which we have used. However, it is known[16] that the δ-function distribution gives the same results as the Gaussian distribution for large N, which is the case discussed here.

After carrying out some simple mathematical manipulations, we find that the distribution of a single dimensionless width $y = \Gamma_\mu / \langle \Gamma_\mu \rangle$ is given by

$$P(y)dy = K[1 - (6\alpha/N)y + (\alpha/N)y^2]y^{-1/2}[\exp(-\tfrac{1}{2}y)]dy, \tag{15}$$

where K is the normalization constant. This distribution is plotted in Fig. 1, which also shows Porter–Thomas (P–T) distribution[17] and Moldauer's numerical histogram.[7] We see from this figure that our curve based on expression (15) lies closer to the numerical plot than the (P–T) distribution.

C. Correlation Function for the Elastic Cross Section

An important quantity in the theory of fluctuations of cross sections[1] is the energy correlation function, which is defined by

$$F(\epsilon) = \langle [\sigma(E + \epsilon) - \langle \sigma \rangle][\sigma(E) - \langle \sigma \rangle] \rangle, \tag{16}$$

where $\sigma(E), \sigma(E + \epsilon)$ are the cross sections at energies E and $E + \epsilon$, respectively. The elastic cross section $\sigma(E)$ is related to the scattering function $S(E)$ by the relation

$$\sigma(E) = (\pi/k^2)\,|\,1 - S(E)\,|^2, \tag{17}$$

where k is the magnitude of the wave vector. We use the following definition of the energy average[2,3] of the function $S(E)$:

$$\langle S(E) \rangle = \int_{-\infty}^{\infty} S(E')\rho(E,E')dE', \tag{18}$$

where $\rho(E,E')$ is the Lorentz weighting function

$$\rho(E,E') = (I/2\pi)[(E - E')^2 + \tfrac{1}{4}I^2]^{-1}, \tag{19}$$

with $I = 2\Delta E/\pi$.

Using the definitions (16)–(18), we can write $F(\epsilon)$ as

$$F(\epsilon) = (\pi^2/k^2)^2 [(\langle S(E + \epsilon)S^*(E) \rangle + \text{c.c.}) - (\langle S(E + \epsilon) \rangle \langle S^*(E) \rangle + \text{c.c.})], \tag{20}$$

where c.c. denotes the complex conjugate. The energy integrals in expression (20) are carried out in the usual way [2,3] by going over to the complex E plane, and are given by

$$\begin{aligned} \langle S(E + \epsilon)S^*(E) \rangle + \text{c.c.} &= -i \sum_{\mu=1}^{N} \frac{G_\mu^2}{E + \epsilon + iI/2 - Z_\mu} \prod_{\lambda=1}^{N} \frac{\epsilon - Z_\mu + Z_\lambda}{\epsilon - Z_\mu + Z_\lambda^*} \\ &\quad - i \sum_{\mu} \frac{G_\mu^2}{E + iI/2 - Z_\mu} \prod_{\lambda} \frac{-\epsilon - Z_\mu + Z_\lambda}{-\epsilon - Z_\mu + Z_\lambda^*} \\ &\quad + \text{c.c.} + 2, \end{aligned} \tag{21}$$

$$\langle S(E + \epsilon) \rangle = [\exp(-2i\phi)]\left(1 - i \sum_{\mu=1}^{N} \frac{G_\mu^2}{E + \epsilon + iI/2 - Z_\mu}\right). \tag{22}$$

Expressions (21) and (22) are exact. As a check, if we put $\epsilon = 0$ in expression (21), then we get $\langle |S(E)|^2 \rangle + \text{c.c.} = 2$, as we should. To simplify expression (21)

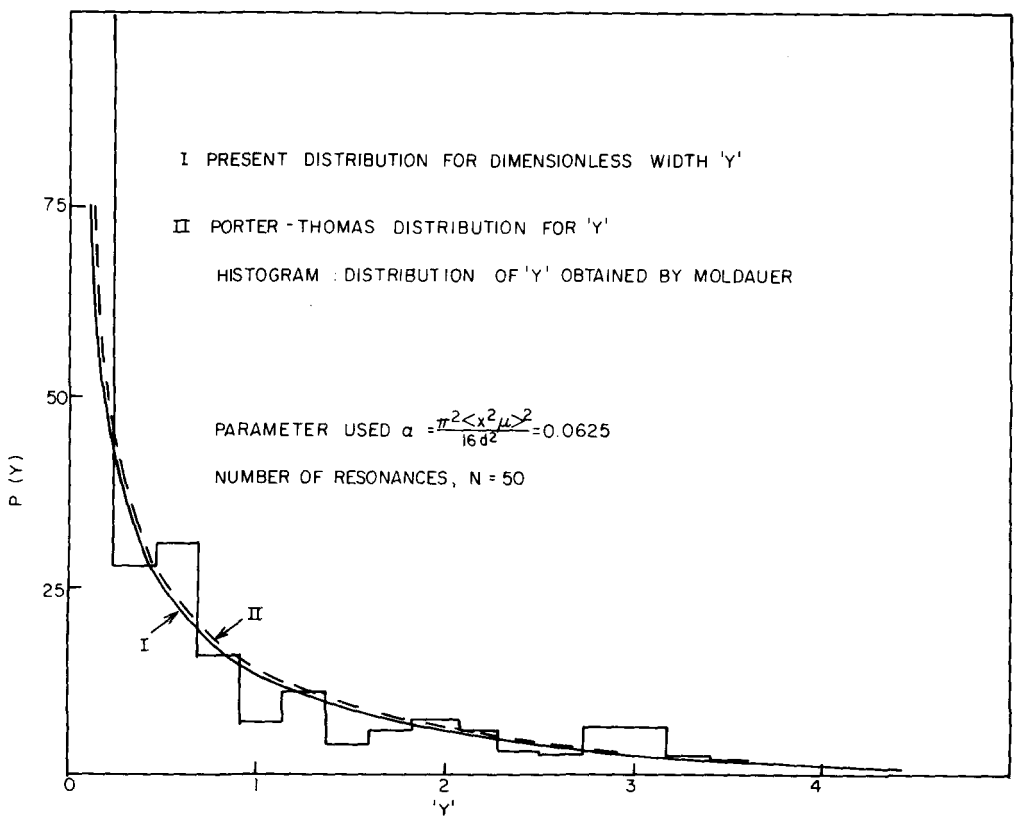


FIG. 1. A plot of the dimensionless width $y = \Gamma_\mu/\langle \Gamma_\mu \rangle$. Also shown are the Porter–Thomas distribution and the numerical histogram.

further, we make the approximation of expanding the product and keeping the first term in in this expansion. The quantity $F(\epsilon)$ in this approximation is given by

$$F(\epsilon) = 2\left(\frac{\pi}{k^2}\right)^2 \frac{2\pi}{D}\left\langle \frac{\Gamma_\mu}{1+\epsilon^2/\Gamma_\mu^2}\right\rangle. \tag{23}$$

Expression (23) is similar to the expression for $F(\epsilon)$ obtained by Ericson[1] for the total cross section.

IV. MULTICHANNEL SCATTERING

We would now like to extend our discussion to m open channels. The previous identity in E which gives the relations between the new resonance parameters Z_λ and the real eigenvalues E_λ and the real width amplitudes $X_{\lambda c}$ now becomes

$$\prod_{\lambda=1}^{N}(E - Z_\lambda) = \det(A^{-1}), \tag{24}$$

where det denotes the determinant of the matrix A^{-1}. As in Sec. III, we write the following relations using the identity (24):

$$\sum_{\mu=1}^{N}\epsilon_\mu = \sum_{\mu=1}^{N} E_\mu, \tag{25a}$$

$$\sum_{\mu=1}^{N}\Gamma_\mu = \sum_{\mu=1,c=1}^{n,m} X_{\mu c}^2, \tag{25b}$$

$$\sum_{\mu<\lambda}^{N}(\epsilon_\mu\epsilon_\lambda - \tfrac{1}{4}\Gamma_\mu\Gamma_\lambda) = \sum_{\mu<\lambda}^{N} E_\mu E_\lambda - \tfrac{1}{4}\sum_{\substack{\mu<\lambda\\ c<c'}}(X_{\mu c}X_{\lambda c'} - X_{\mu c'}X_{\lambda c})^2, \tag{25c}$$

$$\sum_{\mu<\lambda}^{N}(\epsilon_\mu\Gamma_\lambda + \epsilon_\lambda\Gamma_\mu) = \sum_{\mu,c} X_{\mu c}^2\Big(\sum_{\lambda\neq\mu} E_\lambda\Big). \tag{25d}$$

As earlier, we choose the boundary condition $\mathrm{Re}[L^0(1 - R^0L^0)^{-1}] = 0$, for each channel, and assume the constant matrix R^0 to be diagonal.

A. Correlation Coefficients

To calculate the correlation coefficient between two widths $\Gamma_\mu, \Gamma_\lambda$, we take the ensemble averages of Eq. (25b) and its square. They are given by

$$\langle\Gamma_\mu\rangle = \sum_{c=1}^{m}\langle X_{\mu c}^2\rangle, \tag{26a}$$

$$\langle\Gamma_\mu^2\rangle + (N-1)\langle\Gamma_\mu\Gamma_\lambda\rangle_{\mu\neq\lambda} = \sum_{c=1}^{m}[\langle X_{\mu c}^4\rangle + (N-1)\langle X_{\mu c}^2 X_{\lambda c}^2\rangle_{\mu\neq\lambda}] + \sum_{c\neq c'}[\langle X_{\mu c}^2 X_{\mu c'}^2\rangle + (N-1)\langle X_{\mu c}^2 X_{\lambda c'}^2\rangle_{\mu\neq\lambda}]. \tag{26b}$$

The ensemble averages of the real amplitudes $X_{\mu c}$ of R-matrix theory are given by[12]

$$\langle X_{\mu c}^4\rangle = [3N/(N+1)]\langle X_{\mu c}^2\rangle^2, \tag{27a}$$

$$\langle X_{\mu c}^2 X_{\lambda c}^2\rangle_{\mu\neq\lambda} = [N/(N+2)]\langle X_{\mu c}^2\rangle^2, \tag{27b}$$

$$\langle X_{\mu c}^2 X_{\mu c'}^2\rangle_{c\neq c'} = [N/(N+2)]\langle X_{\mu c}^2\rangle\langle X_{\mu c'}^2\rangle(1 + 2C_{X_{\mu c}, X_{\mu c'}}^2), \tag{27c}$$

$$\langle X_{\mu c}^2 X_{\lambda c'}^2\rangle_{\substack{\mu\neq\lambda\\ c\neq c'}} = [N(N+1)/(N-1)(N+2)]\langle X_{\mu c}^2\rangle\langle X_{\mu c'}^2\rangle \times (1 - 2C_{X_{\mu c}, X_{\mu c'}}^2), \tag{27d}$$

where $C_{X_{\mu c}, X_{\mu c'}}$ is the correlation coefficient of two real amplitudes $X_{\mu c}, X_{\mu c'}$. Putting in these ensemble averages in expressions (26) and using the definition of the correlation coefficent $C_{\Gamma_\mu,\Gamma_\lambda}$, we find that it is given by

$$C_{\Gamma_\mu,\Gamma_\lambda} = -(N-1)^{-1}. \tag{28}$$

This is the same as the one which we had obtained for the purely elastic scattering case.

In a similar fashion it can be shown, using expression (25a) and the distribution given by expression (9), that

$$C_{\epsilon_\mu,\epsilon_\lambda} = -1/(N+1) + 2\sigma^2/(N-1)\langle\epsilon_\mu^2\rangle.$$

Substituting for the value of σ^2, which can be obtained using the relations (25), (29), and the following ensemble average,[12]

$$\langle X_{\mu c}X_{\mu c'}X_{\lambda c}X_{\lambda c'}\rangle = -[N^2/(N-1)(N+2)] \times\langle X_{\mu c}^2\rangle\langle X_{\mu c'}^2\rangle(N^{-1} - C_{X_{\mu c}, X_{\mu c'}}), \tag{29}$$

we can finally write the correlation coefficient $C_{\epsilon_\mu,\epsilon_\lambda}$ as

$$C_{\epsilon_\mu,\epsilon_\lambda} = -\frac{1}{N+1} - \frac{1}{2(N+1)(N-1)}\frac{\langle\Gamma_\mu^2\rangle - \langle\Gamma_\mu\rangle^2}{\langle\epsilon_\mu^2\rangle} + \frac{1}{2(N+1)}\frac{\langle\Gamma_\mu\rangle^2}{\langle\epsilon_\mu^2\rangle} + \frac{N}{2(N-1)(N+1)}\frac{1}{\langle\epsilon_\mu^2\rangle} \times\sum_{c\neq c'}\langle X_{\mu c}^2\rangle\langle X_{\mu c'}^2\rangle(C_{X_{\mu c}, X_{\mu c'}}^2 - 1). \tag{30}$$

This expression differs from the corresponding expression (10b) for the purely elastic case in the last term. We see that the last term is nothing but the multichannel effect.

As in the purely elastic case, we find, using expressions (25a), (25b) and (25d), that the correlation coefficient of ϵ_μ, Γ_μ is zero.

We have shown that both for the elastic and the inelastic scattering the correlation coefficient of two total widths $\Gamma_\mu, \Gamma_\lambda$ is of the order N^{-1}. We had derived these results choosing a specific boundary condition and assuming the matrix R^0 to be diagonal. However, it can be shown that this result is independent of the particular choice of the boundary condition and the assumption of the diagonal R^0. The same is not true about the correlation coefficients $C_{\epsilon_\mu,\Gamma_\mu}$ and $C_{\epsilon_\mu,\epsilon_\lambda}$. As a matter of fact for a different choice of the boundary condition the correlation coefficient $C_{\epsilon_\mu,\Gamma_\mu}$ becomes nonzero.

B. A Relation between Resonance Parameters

In the theory of fluctuations of the cross section,[18] one introduces two resonance parameters N_μ, B_c. They are defined by[18]

$$N_\mu = |(VG_\mu)_c|^2/\Gamma_{\mu c}, \tag{31a}$$

$$B_c = |\langle(VG_\mu)_c^2\rangle_\mu/\langle|(VG_\mu)_c|^2\rangle_\mu|^2, \tag{31b}$$

where

$$\sum_{c=1}^{m}\Gamma_{\mu c} = \Gamma_\mu. \tag{31c}$$

TABLE I. A comparison of the values of $\langle B_c \rangle_c$ calculated using the relation (32) with the ones given by numerical calculation

No. of channels	No. of cases	$\langle N_\mu \rangle_\mu$	$\langle B_c \rangle_c$ Moldauer's values	$\langle B_c \rangle_c$ Present calculation
1	5	1.0	0.96	1
20	1	1.1	0.88	0.83
100	1	1.2	0.64	0.694
1	5	1.1	0.70	0.83
20	1	1.2	0.74	0.694
100	1	1.5	0.44	0.44
1	1	1.1	0.55	0.83
20	1	1.8	0.27	0.3086
100	1	2.3	0.19	0.18903
1	5	1.3	0.38	0.5917
20	5	2.1	0.21	0.2267
100	5	3.1	0.12	0.1041
300	5	2.9	0.14	0.12
1	1	1.6	0.05	0.3906
20	1	2.0	0.22	0.25

We would now like to show that these two parameters are related in the following way,

$$\langle B_c \rangle_c = \langle N_\mu \rangle_\mu^{-2}, \tag{32}$$

if V is assumed to be diagonal and the total width Γ_μ is assumed to be almost constant.

This can be shown using the sum rules[8]

$$\sum_{\mu=1}^{N} (G_\mu X G_\mu) = \sum_{\mu=1}^{N} (X_\mu \times X_\mu), \tag{33}$$

and some straightforward mathematical steps using expressions (31) together with the assumptions of diagonal V and constant Γ_μ. The assumption that the total width Γ_μ is almost constant will be a good assumption if the number of channels are large. We use Moldauer's numerical calculation[7] to check relation (32). This is done in Table I where we have shown the values of $\langle B_c \rangle_c$ calculated using relation (32), assuming $\langle N_\mu \rangle_\mu$ to be known, and compared them with the ones obtained by Moldauer using his numerical calculations. We see from this table that relation (32), holds very well particularly when the number of channels are large.

* On leave of absence from Tata Institute of Fundamental Research, Bombay, India.

1 T. Ericson, Ann. Phys. (N.Y.) 23, 390 (1963).

2 H. Feshbach, A. K. Kerman, and R. H. Lemmer, Ann. Phys. (N.Y.) 41, 230 (1967).

3 C. Mahaux and H. A. Weidenmüller, *Shell Model Approach to Nuclear Reactions* (North-Holland, Amsterdam, 1969).

4 Nazakat Ullah, J. Math. Phys. 10, 2099 (1969).

5 T. J. Krieger and C. E. Porter, J. Math. Phys. 4, 1272 (1963).

6 C. E. Porter, *Statistical Theories of Spectra: Fluctuations* (Academic, New York, 1965); N. Rosenzweig, *Brandeis University Summer Institute Lectures in Theoretical Physics 1962 Lectures* (Benjamin, New York 1963), Vol. 3, p. 91.

7 P. A. Moldauer, Phys. Rev. 171, 1164 (1968).

8 Nazakat Ullah and C. S. Warke, Phys. Letters 26B, 556 (1968); Phys. Rev. 164, 1316 (1968).

9 A. M. Lane and R. G. Thomas, Rev. Mod. Phys. 30, 257 (1958).

10 H. Feshbach, Ann. Phys. (N.Y.) 5, 357 (1958); 19, 287 (1962); 43, 410 (1967).

11 C. Mahaux and H. A. Weidenmüller, Nucl. Phys. A91, 241 (1967).

12 Nazakat Ullah, J. Math. Phys. 8, 1095 (1967).

13 C. E. Porter and N. Rosenzweig, Ann. Acad. Sci. Fennicae, Ser. A VI, No. 44 (1960).

14 E. P. Wigner, *Statistical Properties of Real Symmetric Matrices with Many Dimensions*, Canadian Mathematical Congress Proceedings (University of Toronto Press, Toronto, 1957), p. 174.

15 Nazakat Ullah, Phys. Rev. 173, 971 (1968).

16 N. Rosenzweig, Phys. Letters 6, 123 (1963).

17 C. E. Porter and R. G. Thomas, Phys. Rev. 104, 483 (1956).

18 P. A. Moldauer, Phys. Rev. 135, B642 (1964).

Solution of the Difference Equations of Generalized Lucas Polynomials

I. V. V. Raghavacharyulu and A. R. Tekumalla

MATSCIENCE, The Institute of Mathematical Sciences, Madras-20, India

(Received 11 September 1970)

Barakat and Baumann have introduced polynomials $U^{(N)}(a_1, a_2, \ldots, a_N)$ termed the generalized Lucas polynomials satisfying a difference equation with a set of initial conditions. We show that these polynomials can be obtained directly from the symmetric functions h_n, which are of basic importance in combinatorial analysis. Moreover, we extend the definition of $V(a_1, a_2)$ to $V^{(N)}(a_1, a_2, \ldots, a_N)$ and establish that these polynomials too can be obtained from the symmetric functions S_n. Further, closed expressions for the U and V are obtained.

1. INTRODUCTION

In a recent paper, Barakat and Baumann[1] indicated the importance of generalized Lucas polynomials in a variety of physical problems[2,3] and suggested that it is desirable to obtain them in *a closed form*. In Barakat's notation, these polynomials are conveniently defined through a set of difference equations given by

$$U_{n+N}^{(N)}(a_1, a_2, \ldots, a_N) = \sum_{i=1}^{N} (-1)^{i-1} a_i U_{n+N-i}^{(N)} \quad (A_N)$$

together with the N initial conditions

$$U_i^{(N)} = \delta_{N-1,i}, \quad i = 0, 1, \ldots, N-1, \quad (B_N)$$

where δ_{ij} is the Kronecker symbol. Here, in fact, we solve the problem even when the difference equation satisfies a set of arbitrary initial conditions

$$U_i^{(N)} = b_i, \quad i = 0, 1, \ldots, N-1, \quad (C_N)$$

The usual method of solving Eq. (A_N) is by the method of generating functions, making use of the roots of the characteristic equation

$$F(x) = x^N - a_1 x^{N-1} + \cdots \pm a_N = 0. \quad (D_N)$$

Obviously, it is difficult to solve for the roots of Eq. (D_N) in terms of the coefficients $(a_i / i = 1, 2, \ldots, N)$. So, here the solutions of Eqs. (A_N) and (C_N) are obtained in terms of the coefficients $(a_i / i = 1, 2, \ldots, N)$ themselves directly without solving the characteristic equation.

In Sec. 2, we prove that the difference Eq. (A_N) together with initial conditions Eq. (C_N) is solvable in terms of the coefficients of the characteristic Eq. (D_N). In Sec. 3, we summarize conveniently the avail-

TABLE I. A comparison of the values of $\langle B_c \rangle_c$ calculated using the relation (32) with the ones given by numerical calculation

No. of channels	No. of cases	$\langle N_\mu \rangle_\mu$	$\langle B_c \rangle_c$ Moldauer's values	$\langle B_c \rangle_c$ Present calculation
1	5	1.0	0.96	1
20	1	1.1	0.88	0.83
100	1	1.2	0.64	0.694
1	5	1.1	0.70	0.83
20	1	1.2	0.74	0.694
100	1	1.5	0.44	0.44
1	1	1.1	0.55	0.83
20	1	1.8	0.27	0.3086
100	1	2.3	0.19	0.18903
1	5	1.3	0.38	0.5917
20	5	2.1	0.21	0.2267
100	5	3.1	0.12	0.1041
300	5	2.9	0.14	0.12
1	1	1.6	0.05	0.3906
20	1	2.0	0.22	0.25

We would now like to show that these two parameters are related in the following way,

$$\langle B_c \rangle_c = \langle N_\mu \rangle_\mu^{-2}, \tag{32}$$

if V is assumed to be diagonal and the total width Γ_μ is assumed to be almost constant.

This can be shown using the sum rules[8]

$$\sum_{\mu=1}^{N} (G_\mu X G_\mu) = \sum_{\mu=1}^{N} (X_\mu \times X_\mu), \tag{33}$$

and some straightforward mathematical steps using expressions (31) together with the assumptions of diagonal V and constant Γ_μ. The assumption that the total width Γ_μ is almost constant will be a good assumption if the number of channels are large. We use Moldauer's numerical calculation[7] to check relation (32). This is done in Table I where we have shown the values of $\langle B_c \rangle_c$ calculated using relation (32), assuming $\langle N_\mu \rangle_\mu$ to be known, and compared them with the ones obtained by Moldauer using his numerical calculations. We see from this table that relation (32), holds very well particularly when the number of channels are large.

* On leave of absence from Tata Institute of Fundamental Research, Bombay, India.

1 T. Ericson, Ann. Phys. (N.Y.) 23, 390 (1963).

2 H. Feshbach, A. K. Kerman, and R. H. Lemmer, Ann. Phys. (N.Y.) 41, 230 (1967).

3 C. Mahaux and H. A. Weidenmüller, *Shell Model Approach to Nuclear Reactions* (North-Holland, Amsterdam, 1969).

4 Nazakat Ullah, J. Math. Phys. 10, 2099 (1969).

5 T. J. Krieger and C. E. Porter, J. Math. Phys. 4, 1272 (1963).

6 C. E. Porter, *Statistical Theories of Spectra: Fluctuations* (Academic, New York, 1965); N. Rosenzweig, *Brandeis University Summer Institute Lectures in Theoretical Physics 1962 Lectures* (Benjamin, New York 1963), Vol. 3, p. 91.

7 P. A. Moldauer, Phys. Rev. 171, 1164 (1968).

8 Nazakat Ullah and C. S. Warke, Phys. Letters 26B, 556 (1968); Phys. Rev. 164, 1316 (1968).

9 A. M. Lane and R. G. Thomas, Rev. Mod. Phys. 30, 257 (1958).

10 H. Feshbach, Ann. Phys. (N.Y.) 5, 357 (1958); 19, 287 (1962); 43, 410 (1967).

11 C. Mahaux and H. A. Weidenmüller, Nucl. Phys. A91, 241 (1967).

12 Nazakat Ullah, J. Math. Phys. 8, 1095 (1967).

13 C. E. Porter and N. Rosenzweig, Ann. Acad. Sci. Fennicae, Ser. A VI, No. 44 (1960).

14 E. P. Wigner, *Statistical Properties of Real Symmetric Matrices with Many Dimensions*, Canadian Mathematical Congress Proceedings (University of Toronto Press, Toronto, 1957), p. 174.

15 Nazakat Ullah, Phys. Rev. 173, 971 (1968).

16 N. Rosenzweig, Phys. Letters 6, 123 (1963).

17 C. E. Porter and R. G. Thomas, Phys. Rev. 104, 483 (1956).

18 P. A. Moldauer, Phys. Rev. 135, B642 (1964).

Solution of the Difference Equations of Generalized Lucas Polynomials

I. V. V. Raghavacharyulu and A. R. Tekumalla

MATSCIENCE, The Institute of Mathematical Sciences, Madras-20, India

(Received 11 September 1970)

Barakat and Baumann have introduced polynomials $U^{(N)}(a_1, a_2, \ldots, a_N)$ termed the generalized Lucas polynomials satisfying a difference equation with a set of initial conditions. We show that these polynomials can be obtained directly from the symmetric functions h_n, which are of basic importance in combinatorial analysis. Moreover, we extend the definition of $V(a_1, a_2)$ to $V^{(N)}(a_1, a_2, \ldots, a_N)$ and establish that these polynomials too can be obtained from the symmetric functions S_n. Further, closed expressions for the U and V are obtained.

1. INTRODUCTION

In a recent paper, Barakat and Baumann[1] indicated the importance of generalized Lucas polynomials in a variety of physical problems[2,3] and suggested that it is desirable to obtain them in *a closed form*. In Barakat's notation, these polynomials are conveniently defined through a set of difference equations given by

$$U_{n+N}^{(N)}(a_1, a_2, \ldots, a_N) = \sum_{i=1}^{N} (-1)^{i-1} a_i U_{n+N-i}^{(N)} \quad (A_N)$$

together with the N initial conditions

$$U_i^{(N)} = \delta_{N-1,i}, \quad i = 0, 1, \ldots, N-1, \quad (B_N)$$

where δ_{ij} is the Kronecker symbol. Here, in fact, we solve the problem even when the difference equation satisfies a set of arbitrary initial conditions

$$U_i^{(N)} = b_i, \quad i = 0, 1, \ldots, N-1, \quad (C_N)$$

The usual method of solving Eq. (A_N) is by the method of generating functions, making use of the roots of the characteristic equation

$$F(x) = x^N - a_1 x^{N-1} + \cdots \pm a_N = 0. \quad (D_N)$$

Obviously, it is difficult to solve for the roots of Eq. (D_N) in terms of the coefficients $(a_i / i = 1, 2, \ldots, N)$. So, here the solutions of Eqs. (A_N) and (C_N) are obtained in terms of the coefficients $(a_i / i = 1, 2, \ldots, N)$ themselves directly without solving the characteristic equation.

In Sec. 2, we prove that the difference Eq. (A_N) together with initial conditions Eq. (C_N) is solvable in terms of the coefficients of the characteristic Eq. (D_N). In Sec. 3, we summarize conveniently the avail-

able information about the explicit form of $U^{(2)}(a_1, a_2)$ and $U^{(3)}(a_1, a_2, a_3)$, generalize $V^{(2)}(a_1, a_2)$ to $V^{(N)}(a_1, a_2, \ldots, a_N)$ and give explicit expressions in *closed form* for $U_n^{(N)}(a_1, a_2, \ldots, a_N)$ and $V_n^{(N)}(a_1, a_2, \ldots, a_N)$. In Sec. 4 and in the Appendix, we prove that indeed the generalized Lucas U and V polynomials have the form given in Sec. 3.

2. SOLVABILITY OF THE DIFFERENCE EQUATION IN TERMS OF $(a_i/i = 1, 2, \ldots, N)$

The usual method of solving Eqs. $(\mathbf{A}_N)$ and $(\mathbf{C}_N)$ is by the method of generating functions. Let

$$f(x) = \sum_{n=0}^{\infty} U_n x^n$$

and

$$g(x) = \sum_{i=0}^{N} (-1)^i a_i x^i \tag{2.1}$$

be the generating functions of the sequences $\{U_n\}$ and $\{1, -a_1, a_2, \ldots, (-1)^N a_N\}$, respectively. (The superscript N has been dropped in this section for convenience). Then the product

$$f(x)g(x) = \sum_{n=0}^{\infty} W_n x^n \tag{2.2}$$

is the generating function of the sequence $\{W_n\}$, where

$$W_n = U_n - a_1 U_{n-1} + a_2 U_{n-2} - \cdots \pm a_N U_{n-N} \tag{2.3}$$

and $n = 0, 1, 2, \cdots$.

Also from Eqs. $(\mathbf{A}_N)$ and (2.3),

$$W_n = 0 \quad \text{for } n \geq N \tag{2.4}$$

and from the initial conditions and Eq. (2.3),

$$W_n = b_n - a_1 b_{n-1} + a_2 b_{n-2} - \cdots \pm a_n b_0, \quad n = 0, 1, \ldots, N-1. \tag{2.5}$$

Therefore, from Eq. (2.2),

$$f(x) = \frac{W_0 + W_1 x + \cdots + W_{N-1} x^{N-1}}{1 - a_1 x + a_2 x^2 - \cdots \pm a_N x^N} \equiv \frac{n(x)}{p(x)} \quad \text{say,} \tag{2.6}$$

Theorem 1: The generating function $f(x)$ of $\{U_n\}$ can be expanded in terms of the coefficients of the polynomials $p(x)$ and $n(x)$.

Proof: From Eq. (2.6), it is enough to show that $1/p(x) = \sum_n h_n x^n$, where each h_n is a function of $(a_i/i = 1, 2, \ldots, N)$.

It is obvious from the form of $p(x)$ that $1/p(x)$ can be written as

$$\frac{1}{p(x)} = [1 - q(x)]^{-1} = 1 + q(x) + [q(x)]^2 + \cdots,$$

where

$$q(x) = \sum_{i=1}^{N} (-1)^{i-1} a_i x^i \tag{2.7}$$

and the radius of convergence of the series can be taken as $\min(|\alpha_i| \, i = 1, 2, \ldots, N)$, where α_i are the roots of Eq. $(\mathbf{D}_N)$. Hence the theorem.

3. LUCAS POLYNOMIALS FOR $N = 2, 3$

When $N = 2$, the Lucas polynomials[4-6] are obtained as the solution of difference equation $(\mathbf{A}_2)$ satisfying the initial conditions $(\mathbf{B}_2)$. The solution for Eqs. $(\mathbf{A}_2)$ and $(\mathbf{B}_2)$ is

$$U_{n+1}^{(2)}(a_1, a_2) = \sum_{\nu} (-1)^{\nu} \binom{n-\nu}{\nu} a_1^{n-2\nu} a_2^{\nu}, \tag{3.1}$$

the series terminating when the exponent of a_1 or a_2 turns negative.

The general term of Eq. (3.1) can be more conveniently written as

$$(-1)^{n-\Sigma\lambda} \frac{(\lambda_1 + \lambda_2)!}{\lambda_1! \lambda_2!} a_1^{\lambda_1} a_2^{\lambda_2}, \tag{3.2}$$

where $\lambda_1 + 2\lambda_2 = n$.

For $N = 2$ the Lucas $V^{(2)}(a_1, a_2)$ polynomials are defined by the difference equation $(\mathbf{A}_2)$ satisfying the initial conditions $V_0^{(2)} = 2$ and $V_1^{(2)} = a_1$.

For $N = 3$, one has to solve the difference equation $(\mathbf{A}_3)$ satisfying the initial conditions $(\mathbf{B}_3)$. The possible general solution suggested in closed form by Barakat and Baumann[1] for $U^{(3)}$ is

$$U_{n+2}^{(3)}(a_1, a_2, a_3) = U_{n+1}^{(2)}(a_1, a_2) + \sum_{K=1} \sum_{K=l} (-1)^{l-K} \binom{l}{K} \binom{n-l-K}{l} a_1^{n-2l-K} a_2^{l-K} a_3^{K} \tag{3.3}$$

again with the understanding that all the exponents are ≥ 0.

Making use of Eq. (3.1) the general term of Eq. (3.3) can be written as

$$(-1)^{n-\Sigma\lambda} [(\lambda_1 + \lambda_2 + \lambda_3)!/\lambda_1! \lambda_2! \lambda_3!] a_1^{\lambda_1} a_2^{\lambda_2} a_3^{\lambda_3}, \tag{3.4}$$

where $\lambda_1 + 2\lambda_2 + 3\lambda_3 = n$.

Assuming the correctness of Eqs. (3.1) and (3.3), it is easy to guess the general solution of the difference equations of generalized Lucas polynomials. Indeed, we have the following:

Theorem 2: The general solution to the difference equation

$$U_{n+N}^{(N)}(a_1, a_2, \ldots, a_N) = a_1 U_{n+N-1}^{(N)} - a_2 U_{n+N-2} + \cdots \pm a_N U_n^{(N)}$$

satisfying the initial conditions

$$U_0^{(N)} = U_1^{(N)} = \cdots = U_{N-2}^{(N)} = 0, \quad U_{N-1}^{(N)} = 1$$

is

$$U_{n+N-1}^{(N)}(a_1, a_2, \ldots, a_N) = h_n(a_1, a_2, \ldots, a_N, \ldots) \tag{3.5}$$

with $a_{N+1} = a_{N+2} = \cdots = 0$, where h_n the symmetric functions called homogeneous[7] product sums of weight n (see Appendix).

Note that Eq. (2.6) gives the solution to Eq. $(\mathbf{A}_N)$ with any initial conditions $(\mathbf{C}_N)$. In particular, for $N = 2$ and $b_0 = 2, b_1 = a_1$ we get Lucas $V^{(2)}(a_1, a_2)$ polynomials which can be identified as the one part symmetric functions s_n with $a_3 = a_4 = \cdots = 0$. We now characterize them in the following:

Theorem 3: The general solution of the difference equation

$$V_{n+N}^{(N)}(a_1, a_2, \ldots, a_N) = a_1 V_{n+N-1}^{(N)} - a_2 V_{n+N-2}^{(N)} + \cdots a_N V_n^{(N)}$$

satisfying the initial conditions

$$V_0 = N, \quad V_i = s_i, \quad i = 1,2,\ldots,N-1 \tag{3.6}$$

is

$$V_n^{(N)}(a_1,a_2,\ldots,a_N) = s_n(a_1,a_2,\ldots,a_N,\ldots)$$

with

$$a_{N+1} = a_{N+2} = \cdots = 0, \tag{3.7}$$

where s_n is the one-part symmetric function. (See Appendix)

The expressions for $U_n^{(N)}$ and $V_n^{(N)}$ given by Eqs. (3.5) and (3.7), respectively, enable us to obtain, simply, the Lucas polynomials. The first few h_n and s_n ($n = 1$–6) are given in the Appendix from which all the results that have been given by Barakat and Baumann for $N(n) = 3(8), 4(9)$, and $5(10)$ are easily obtained. Hence, in tabulating $U_n^{(N)}$ and $V_n^{(N)}$ it is sufficient to give h_n and s_n only.

The theorems are proved in the next section.

4. SOLUTIONS OF THE DIFFERENCE EQUATIONS

Proof of Theorem 2: We know from the well-known connection between the symmetric functions[7] a_n and h_n that

$$\frac{1}{1 - a_1x + a_2x^2 - \cdots \pm a_rx^r \mp \cdots} = 1 + h_1x + h_2x^2 + \cdots, \tag{4.1}$$

where, in the denominator on the left-hand side, we can assume without loss of generality an infinite number of terms. Expanding the left-hand side by the multinomial theorem,[7] we obtain

$$h_n = \sum_{\lambda_1,\lambda_2,\cdots} (-1)^{n-\Sigma\lambda} \frac{(\Sigma\lambda!)}{\lambda_1!\lambda_2!\cdots\lambda_n!} a_1^{\lambda_1}a_2^{\lambda_2}\cdots a_n^{\lambda_n}, \tag{4.2}$$

where $\Sigma i\lambda_i = n$. Note that $(-1)^{n-\Sigma\lambda} = (-1)^{\lambda_2+\lambda_4+\cdots}$. Hence, the general solution of the difference equation (A_N) with the initial conditions (B_N) is obtained by comparing the coefficients of $f(x) = x^{N-1}/p(x)$.

In particular, as the solution for the Lucas polynomials, we obtain

$$U_{n+N-1}^{(N)} = \sum_{\lambda_1,\lambda_2,\ldots,\lambda_N} (-1)^{n-\Sigma\lambda} \frac{(\Sigma\lambda)!}{\lambda_1!\lambda_2!\ldots\lambda_N!} \times a_1^{\lambda_1}a_2^{\lambda_2}\cdots a_N^{\lambda_N}, \tag{4.3}$$

where $\Sigma i\lambda_i = n$ which proves Theorem 2.

Proof of Theorem 3: Let $F(x) = x^N - a_1x^{N-1} + \cdots \pm a_N$. Then $p(x) = F(1/x)x^N$.

Solving for W_n by making use of the initial conditions (3.6), we obtain $n(x) = F^1(1/x)x^{N-1}$, where $F^1(x)$ is the derivative of $F(x)$. Further, we know that

$$\frac{F^1(x)}{F(x)} = \frac{1}{x-\alpha_1} + \frac{1}{x-\alpha_2} + \cdots + \frac{1}{x-\alpha_N},$$

where $(\alpha_i/i = 1,2,\ldots,N)$ are the roots of $F(x) = 0$. Hence, we obtain

$$f(x) = \frac{1}{x}\frac{F^1(1/x)}{F(1/x)} = \frac{1}{1-\alpha_1x} + \frac{1}{1-\alpha_1x} + \cdots + \frac{1}{1-\alpha_Nx} = 1 + s_1x + s_2x^2 + \cdots \tag{4.4}$$

giving s_n as the solution of the difference equation (A_N) satisfying the initial conditions (3.6). It is not difficult to see that the general expression for the generalized Lucas $V_n^{(N)}$ polynomials in closed form is

$$V_n^{(N)} = \sum_{\lambda_1,\lambda_2,\ldots,\lambda_N} (-1)^{n-\Sigma\lambda} \frac{(-1+\Sigma\lambda)!n}{\lambda_1!\lambda_2!\cdots\lambda_N!} a_1^{\lambda_1}a_2^{\lambda_2}\cdots a_N^{\lambda_N} \tag{4.5}$$

with $\Sigma i\lambda_i = n$.

Theorem 4: If $\alpha_1,\alpha_2,\ldots,\alpha_N$ are the roots of the equation $x^N - a_1x^{N-1} + \cdots \pm a_N = 0$, then

$$U_n^{(N)} = \begin{vmatrix} \alpha_1^n & \alpha_2^n & \cdots & \alpha_N^n \\ \alpha_1^{N-2} & \alpha_2^{N-2} & \cdots & \alpha_N^{N-2} \\ \vdots & \vdots & & \vdots \\ \alpha_1 & \alpha_2 & \cdots & \alpha_N \\ 1 & 1 & \cdots & 1 \end{vmatrix} \div \begin{vmatrix} \alpha_1^{N-1} & \alpha_2^{N-1} & \cdots & \alpha_N^{N-1} \\ \alpha_1^{N-2} & \alpha_2^{N-2} & \cdots & \alpha_N^{N-2} \\ \vdots & \vdots & & \vdots \\ \alpha_1 & \alpha_2 & \cdots & \alpha_N \\ 1 & 1 & \cdots & 1 \end{vmatrix} \tag{4.6}$$

and $V_n^{(N)} = \alpha_1^n + \alpha_2^n + \cdots + \alpha_N^n$.

$U_n^{(N)}$ and $V_n^{(N)}$ obviously satisfy the initial conditions (B_N) and (3.6), respectively. When repeated roots occur, we apply d'Hospitals rule to simplify these expressions.

The first part of the theorem follows immediately from the expression

$$\frac{1}{p(x)} = \sum_{i=1}^{N} \frac{\alpha_i^{N-1}}{\Pi_{j=1}'^{N}(\alpha_i - \alpha_j)} \frac{1}{1-\alpha_ix}$$

by expanding the terms on the right-hand side in power series and collecting the coefficients of x^{n-N+1}. The second part is obvious. When $N = 2$, the above expressions reduce to[4]

$$U_n^{(2)} = \frac{\alpha_1^n - \alpha_2^n}{\alpha_1 - \alpha_2} \text{ and } V_n^{(2)} = \alpha_1^n + \alpha_2^n.$$

A further extension of the Lucas V polynomials is possible by writing

$$V_{n,m}^{(N)} = S_{n+m,m} \tag{4.7}$$

with $a_{N+1} = a_{N+2} = \cdots = 0$, where $S_{w,m}$ are the m part symmetric functions of weight w, given by

$$S_{w,m} = \sum (-1)^{w+m-\Sigma\lambda-1} \times \frac{(\Sigma\lambda-1)!\left[\lambda_m + \binom{m+1}{1}\lambda_{m+1} + \cdots\right]}{\lambda_1!\lambda_2!\cdots\lambda_w!} \times a_1^{\lambda_1}a_2^{\lambda_2}\cdots a_w^{\lambda_w},$$

where $\Sigma i\lambda_i = w$ and $w \geq m$.

$V_{n,m}^{(N)}$ satisfies the difference equation

$$V_{n,m}^{(N)}(a_1,a_2,\ldots,a_N) = a_1V_{n-1,m}^{(N)} - a_2V_{n-2,m}^{(N)} + \cdots \pm a_NV_{n-N,m}^{(N)}$$

with the initial conditions $V_{i,m}^{(N)} = S_{i+m,m}$, $i = 1,2,\ldots,N$. For $m = 1$, we get the generalized Lucas $V^{(N)}$ polynomials with the initial conditions $V_i^{(N)} = S_i$, $i = 1,2,\ldots,N$. Further, it is very well known that *h_n and s_n could be extended to what are called Schur functions.*[8] Similar considerations as are developed in this paper can be generalized to apply to these functions too. This extension together with the applications will be considered elsewhere.

ACKNOWLEDGMENT

It is a pleasure to thank Professor Alladi Ramakrishnan for giving us an opportunity to work at Matscience, where the above research was carried out.

APPENDIX

1. The symmetric functions h_n are given by

$$h_1 = \Sigma\alpha_1,$$
$$h_2 = \Sigma\alpha_1^2 + \Sigma\alpha_1\alpha_2,$$
$$h_3 = \Sigma\alpha_1^3 + \Sigma\alpha_1^2\alpha_2 + \Sigma\alpha_1\alpha_2\alpha_3,$$

where $\alpha_1, \alpha_2, \cdots$ are the roots of the equation

$$x^n - a_1x^{n-1} + \cdots = 0.$$

In terms of a_i, they are given by

$$h_1 = a_1, \quad h_2 = a_1^2 - a_2, \quad h_3 = a_1^3 - 2a_1a_2 + a_3,$$
$$h_4 = a_1^4 - 3a_1^2a_2 + a_2^2 + 2a_1a_3 - a_4,$$
$$h_5 = a_1^5 - 4a_1^3a_2 + 3a_1a_2^2 + 3a_1^2a_3 - 2a_2a_3 - 2a_1a_4 + a_5,$$
$$h_6 = a_1^6 - 5a_1^4a_2 + 6a_1^2a_2^2 - a_2^3 + 4a_1^3a_3 - 6a_1a_2a_3 + a_3^2 - 3a_1^2a_4 + 2a_2a_4 + 2a_1a_5 - a_6.$$

2. The symmetric functions s_n are the successive sums of powers of $\alpha_1, \alpha_2, \cdots$,

$$s_1 = \Sigma\alpha_1,$$
$$s_2 = \Sigma\alpha_1^2,$$
$$s_3 = \Sigma\alpha_1^3.$$

In terms of a_i they are given by

$$s_1 = a_1, \quad s_2 = a_1^2 - 2a_2, \quad s_3 = a_1^3 - 3a_1a_2 + 3a_3,$$
$$s_4 = a_1^4 - 4a_1^2a_2 + 2a_2^2 + 4a_1a_3 - 4a_4,$$
$$s_5 = a_1^5 - 5a_1^3a_2 + 5a_1a_2^2 + 5a_1^2a_3 - 5a_2a_3 - 5a_1a_4 + 5a_5.$$

[1] R. Barakat and E. Baumann, J. Math. Phys. **10**, 1474 (1969).
[2] V. Moweny, IRE Trans. Circuit Theory **11**, 232 (1964).
[3] S. Basin, Math. Mag. **37**, 83 (1964).
[4] E. Lucas, *Théorie des nombres* (Gauthier-Villars, Paris, 1891).
[5] A. Herpin, Compt. Rend. **225**, 17 (1947).
[6] R. Barakat, J. Math. Phys. **43**, 325 (1964).
[7] P. A. MacMahon, *Combinatory Analysis* (Chelsea, New York, 1960)
[8] D. E. Littlewood, *The Theory of Group Characters and Matrix Representations of Groups* (Clarendon, Oxford, 1950), 2nd ed.

Normalization of Certain Higher-Order Phase-Integral Approximations for Wavefunctions of Bound States in a Potential Well. II*

Staffan Yngve
Department of Theoretical Physics, University of Uppsala, Uppsala, Sweden
(Received 21 May 1971; Revised Manuscript Received 15 October 1971)

Using a formula derived from equations given by Furry for the normalization integral of the wavefunction corresponding to a bound state, we derive the normalization factor for the higher-order phase-integral approximations introduced by N. Fröman. The present treatment is based on the method developed by N. Fröman and P. O. Fröman, in which one uses exact formulas in the calculations and makes the approximations in the final stage. We particularize the resulting general formula to the case of a single-well potential previously discussed by the present author and to the case of a double-well potential, which has been treated in a series of papers from this institute.

1. INTRODUCTION

In a previous paper by the present author,[1] which is part I of the present paper, the normalization factor for the higher-order phase-integral approximations introduced by N. Fröman[2] was obtained for the wavefunction of a bound state in a single-well potential. The difference between these higher-order phase-integral approximations and the higher-order JWKB approximations has recently been discussed in some detail by N. Fröman.[3] Both approximations fail near the classical turning points, i.e., near the points where the particle, according to classical mechanics, would stop and begin to move in the opposite direction. One cannot, therefore, directly insert the approximate wavefunction into the normalization condition for the exact wavefunction ψ,

$$\int_{-\infty}^{+\infty} |\psi(x)|^2 dx = 1.$$

From an intuitive discussion for the first-order JWKB approximation, Pauli[4] has derived a formula for the normalization factor of the approximate wavefunction of a bound state in a single-well potential. This formula, which is met with also in textbooks (see, e.g., Ref. 5) has been rigorously derived by Furry.[6] In his treatment Furry[6] was able to avoid the vicinity of the turning points and thus used the JWKB approximation only in regions where it is known to be good.

In Ref. 1 the present author has generalized Furry's discussion[6] to the case of the higher-order phase-integral approximations introduced by N. Fröman.[2] The intuitive approach of Pauli[4] would fail for approximations of higher order than the first.

In order to obtain the normalization factor for more complicated potential wells, we shall in the present

with the initial conditions $V_{i,m}^{(N)} = S_{i+m,m}$, $i = 1,2,\ldots,N$. For $m = 1$, we get the generalized Lucas $V^{(N)}$ polynomials with the initial conditions $V_i^{(N)} = S_i$, $i = 1,2,\ldots,N$. Further, it is very well known that *h_n and s_n could be extended to what are called Schur functions.*[8] Similar considerations as are developed in this paper can be generalized to apply to these functions too. This extension together with the applications will be considered elsewhere.

ACKNOWLEDGMENT

It is a pleasure to thank Professor Alladi Ramakrishnan for giving us an opportunity to work at Matscience, where the above research was carried out.

APPENDIX

1. The symmetric functions h_n are given by

$$h_1 = \Sigma\alpha_1,$$
$$h_2 = \Sigma\alpha_1^2 + \Sigma\alpha_1\alpha_2,$$
$$h_3 = \Sigma\alpha_1^3 + \Sigma\alpha_1^2\alpha_2 + \Sigma\alpha_1\alpha_2\alpha_3,$$

where $\alpha_1, \alpha_2, \cdots$ are the roots of the equation

$$x^n - a_1x^{n-1} + \cdots = 0.$$

In terms of a_i, they are given by

$$h_1 = a_1, \quad h_2 = a_1^2 - a_2, \quad h_3 = a_1^3 - 2a_1a_2 + a_3,$$
$$h_4 = a_1^4 - 3a_1^2a_2 + a_2^2 + 2a_1a_3 - a_4,$$
$$h_5 = a_1^5 - 4a_1^3a_2 + 3a_1a_2^2 + 3a_1^2a_3 - 2a_2a_3 - 2a_1a_4 + a_5,$$
$$h_6 = a_1^6 - 5a_1^4a_2 + 6a_1^2a_2^2 - a_2^3 + 4a_1^3a_3 - 6a_1a_2a_3 + a_3^2 - 3a_1^2a_4 + 2a_2a_4 + 2a_1a_5 - a_6.$$

2. The symmetric functions s_n are the successive sums of powers of $\alpha_1, \alpha_2, \cdots$,

$$s_1 = \Sigma\alpha_1,$$
$$s_2 = \Sigma\alpha_1^2,$$
$$s_3 = \Sigma\alpha_1^3.$$

In terms of a_i they are given by

$$s_1 = a_1, \quad s_2 = a_1^2 - 2a_2, \quad s_3 = a_1^3 - 3a_1a_2 + 3a_3,$$
$$s_4 = a_1^4 - 4a_1^2a_2 + 2a_2^2 + 4a_1a_3 - 4a_4,$$
$$s_5 = a_1^5 - 5a_1^3a_2 + 5a_1a_2^2 + 5a_1^2a_3 - 5a_2a_3 - 5a_1a_4 + 5a_5.$$

[1] R. Barakat and E. Baumann, J. Math. Phys. **10**, 1474 (1969).
[2] V. Moweny, IRE Trans. Circuit Theory **11**, 232 (1964).
[3] S. Basin, Math. Mag. **37**, 83 (1964).
[4] E. Lucas, *Théorie des nombres* (Gauthier-Villars, Paris, 1891).
[5] A. Herpin, Compt. Rend. **225**, 17 (1947).
[6] R. Barakat, J. Math. Phys. **43**, 325 (1964).
[7] P. A. MacMahon, *Combinatory Analysis* (Chelsea, New York, 1960)
[8] D. E. Littlewood, *The Theory of Group Characters and Matrix Representations of Groups* (Clarendon, Oxford, 1950), 2nd ed.

Normalization of Certain Higher-Order Phase-Integral Approximations for Wavefunctions of Bound States in a Potential Well. II*

Staffan Yngve
Department of Theoretical Physics, University of Uppsala, Uppsala, Sweden
(Received 21 May 1971; Revised Manuscript Received 15 October 1971)

Using a formula derived from equations given by Furry for the normalization integral of the wavefunction corresponding to a bound state, we derive the normalization factor for the higher-order phase-integral approximations introduced by N. Fröman. The present treatment is based on the method developed by N. Fröman and P. O. Fröman, in which one uses exact formulas in the calculations and makes the approximations in the final stage. We particularize the resulting general formula to the case of a single-well potential previously discussed by the present author and to the case of a double-well potential, which has been treated in a series of papers from this institute.

1. INTRODUCTION

In a previous paper by the present author,[1] which is part I of the present paper, the normalization factor for the higher-order phase-integral approximations introduced by N. Fröman[2] was obtained for the wavefunction of a bound state in a single-well potential. The difference between these higher-order phase-integral approximations and the higher-order JWKB approximations has recently been discussed in some detail by N. Fröman.[3] Both approximations fail near the classical turning points, i.e., near the points where the particle, according to classical mechanics, would stop and begin to move in the opposite direction. One cannot, therefore, directly insert the approximate wavefunction into the normalization condition for the exact wavefunction ψ,

$$\int_{-\infty}^{+\infty} |\psi(x)|^2 dx = 1.$$

From an intuitive discussion for the first-order JWKB approximation, Pauli[4] has derived a formula for the normalization factor of the approximate wavefunction of a bound state in a single-well potential. This formula, which is met with also in textbooks (see, e.g., Ref. 5) has been rigorously derived by Furry.[6] In his treatment Furry[6] was able to avoid the vicinity of the turning points and thus used the JWKB approximation only in regions where it is known to be good.

In Ref. 1 the present author has generalized Furry's discussion[6] to the case of the higher-order phase-integral approximations introduced by N. Fröman.[2] The intuitive approach of Pauli[4] would fail for approximations of higher order than the first.

In order to obtain the normalization factor for more complicated potential wells, we shall in the present

paper use the method for handling the phase-integral approximations developed by N. Fröman and P. O. Fröman.[7] An essential feature of this method is that one uses exact formulas in the calculations and makes the approximations in the final stage. The discussion in Sec. 2 of the present paper leads to an exact normalization condition, valid for a potential with an arbitrary number of classically allowed regions. This normalization condition is shown to be simply related to the quantization condition. In Secs. 3 and 4 the results of Sec. 2 are applied to the case of a single-well potential previously discussed by the present author[1] and to the case of a double-well potential. The energy levels and wavefunctions of a double oscillator have been treated in a series of papers from this institute[8-10] For any order of the phase-integral approximations the authors of Ref. 9 have given approximate expressions for the unnormalized wavefunction of a symmetric double oscillator on the real axis, except for small regions where the approximation fails. Numerical results for the energy eigenvalues reported in Ref. 9 and Ref. 10 show the great accuracy of the higher-order phase-integral approximations. In order to make full use of this accuracy, it is important to know also the higher-order corrections to the normalization factor.

2. AN EXACT FORMULA FOR THE NORMALIZATION FACTOR

Consider the one-dimensional differential equation

$$\frac{d^2\psi}{dz^2} + Q^2(z)\psi = 0, \tag{1}$$

where $Q^2(z)$ is real for real values x of z. For the case of the Schrödinger equation, which is of interest to us, we have well-known notations

$$Q^2(z) = (2m/\hbar^2)[E - V(z)], \tag{2}$$

where E is a real parameter. It is emphasized that quantities related to the differential equation (1) with $Q^2(z)$ given by (2) will in general depend on the parameter E. We only consider such values of the parameter E that $Q^2(x)$ is negative when $|x|$ is sufficiently large.

First we shall derive a convenient formula for obtaining the normalization integral. Using (1), we obtain the equation

$$[\psi(x)]^2 \frac{\partial}{\partial E} Q^2(x) = \frac{d}{dx}\left(\psi'(x) \frac{\partial}{\partial E}\psi(x) - \psi(x) \frac{\partial}{\partial E}\psi'(x)\right), \tag{3}$$

where prime denotes differentiation with respect to x. As in Ref. 1, we define $\psi_1(z)$ as a certain solution of (1), which tends to zero as z tends to $-\infty$ through real values. This solution is uniquely determined except for a factor, which is independent of z but may depend on E. It can be shown that the function to the right of the operator d/dx in (3) with $\psi = \psi_1$ tends to zero as $x \to -\infty$. We similarly define $\psi_2(z)$ as a solution of (1), which tends to zero as z tends to $+\infty$ through real values. From (3) with $\psi = \psi_1$ and $\psi = \psi_2$, respectively, we obtain the following formulas:

$$\int_{-\infty}^{x_0} [\psi_1(x)]^2 \frac{\partial}{\partial E} Q^2(x)dx = \left(\psi_1'(x_0) \frac{\partial}{\partial E}\psi_1(x_0) - \psi_1(x_0) \frac{\partial}{\partial E}\psi_1'(x_0)\right), \tag{4a}$$

$$\int_{x_0}^{+\infty} [\psi_2(x)]^2 \frac{\partial}{\partial E} Q^2(x)dx = \left(\psi_2(x_0) \frac{\partial}{\partial E}\psi_2'(x_0) - \psi_2'(x_0) \frac{\partial}{\partial E}\psi_2(x_0)\right), \tag{4b}$$

where the point x_0 can be chosen arbitrarily. Formulas (4) with Q^2 given by (2) were obtained by Furry.[6] As in Ref. 1, we assume that

$$\psi_1(z) = \psi_2(z) = \psi(z) \tag{5}$$

when E is an eigenvalue. Adding (4a) and (4b), and using (5), we obtain, when E is an eigenvalue,

$$\int_{-\infty}^{+\infty} [\psi(x)]^2 \frac{\partial}{\partial E} Q^2(x)dx = \frac{\partial}{\partial E}(\psi_1\psi_2' - \psi_2\psi_1'), \tag{6}$$

where on the right-hand side we shall put $z = x_0$. However, the quantity to the right of the operator $\partial/\partial E$ in the right-hand member of (6) is the Wronskian of the solutions $\psi_1(z)$ and $\psi_2(z)$, and it is well known that this Wronskian does not depend on z but may depend on E. It should be emphasized that Eqs. (1)-(4) are valid for all values of the parameter E considered but that Eqs. (5) and (6) are only valid when E is an eigenvalue.

We shall now give a few basic formulas directly taken from Refs. 2 and 7. For details we refer to Chaps. 3-5 in Ref. 7. We define two linearly independent functions

$$f_1(z) = q^{-1/2}(z) \exp[+ iw(z)], \tag{7a}$$

$$f_2(z) = q^{-1/2}(z) \exp[- iw(z)], \tag{7b}$$

where

$$w'(z) = q(z) \tag{8}$$

and the prime denotes differentiation with respect to z. From (7a), (7b), and (8) it follows that

$$f_1(z)f_2'(z) - f_2(z)f_1'(z) = -2i. \tag{9}$$

The function $q(z)$ is determined such that $f_1(z)$ and $f_2(z)$ are approximate solutions of (1). By using a Riemann surface and defining $w(z)$ as a convenient contour integral of $q(z)$ [cf. (8)] we can make the functions $f_1(z)$ and $f_2(z)$ single-valued in the region of the complex plane which we consider. To begin with, we need not specify the expression for $q(z)$, but in Secs. 3 and 4 we shall let $q(z)$ be the function $q(z)$ appearing in the higher-order phase-integral approximations introduced by N. Fröman.[2] Recursion formulas for obtaining this function up to any order of the phase-integral approximations are given in Ref. 2, and explicit expressions for $q(z)$ up to the ninth-order approximation can be found in Ref. 1 or in Ref. 11.

According to (3.25a), (3.25b) in Ref. 7 any solution ψ of (1) and its derivative ψ' can be written as follows:

$$\psi(z) = \boldsymbol{f}(z)\boldsymbol{a}(z), \tag{10a}$$

$$\psi'(z) = \boldsymbol{f}'(z)\boldsymbol{a}(z). \tag{10b}$$

In these formulae, which involve no approximation, $\boldsymbol{f}(z)$ is the row vector

$$\boldsymbol{f}(z) = (f_1(z), f_2(z)), \tag{11}$$

the components of which are given by (7a), (7b), and

$a(z)$ is a column vector, the components of which are denoted by $a_1(z)$ and $a_2(z)$:

$$a(z) = \begin{pmatrix} a_1(z) \\ a_2(z) \end{pmatrix}. \tag{12}$$

If $a(z_0)$ is known, the column vector $a(z)$ at an arbitrary point z is given by the formula

$$a(z) = F(z, z_0)a(z_0), \tag{13}$$

where $F(z, z_0)$ is a matrix obeying the multiplication rule

$$F(z, z_0) = F(z, z_1)F(z_1, z_0) \tag{14}$$

and the inversion formula (3.20) in Ref. 7, which can be written

$$F_{lm}(z_0, z) = (-1)^{l+m}F_{3-m,3-l}(z, z_0), \tag{15}$$

where l and m can take the values 1 or 2. If $Q^2(x)$ is real and x_1 and x_2 are points on the real axis, there exist certain symmetry relations between the elements of $F(x_1, x_2)$. These relations are given by (5.7), (5.8), and (5.9) in Ref. 7.

Consider now a smooth potential with an arbitrary number of classically allowed regions separated by classically forbidden regions. As in Ref. 7, we use the expressions classically allowed and classically forbidden in a generalized sense for the regions, where $q^2(x)$ is positive and negative, respectively. We only consider such values of the real parameter E that $q^2(x)$ and $Q^2(x)$ are negative when $|x|$ is sufficiently large, but to begin with we shall not restrict ourselves to the case when E is an eigenvalue. Of the two functions $f_1(x)$ and $f_2(x)$ we denote by $f_j(x)$ the one which tends to zero as x tends to $-\infty$, and we denote by $f_k(x)$ the one which tends to zero as x tends to $+\infty$. We thus have the formulas

$$f_j(x) \to 0 \quad \text{as} \quad x \to -\infty, \tag{16a}$$

$$f_k(x) \to 0 \quad \text{as} \quad x \to +\infty, \tag{16b}$$

and similarly as on p. 104 in Ref. 7 we shall assume that the integrals

$$\int_{-\infty} |f_{3-j}(x)|^2 dx, \tag{17a}$$

$$\int^{+\infty} |f_{3-k}(x)|^2 dx \tag{17b}$$

are both divergent. The pair of integers (j, k) can take the values (1, 1), (1, 2), (2, 1), and (2, 2) depending on the number of classically allowed regions and on the phase chosen for $q^{1/2}(z)$.

Proceeding similarly as on pp. 103–05 in Ref. 7, one can show that the a-coefficients at an arbitrary point z for the solution $\psi_1(z)$ of (1), which tends to zero as z tends to $-\infty$ through real values, are given by the formulas

$$a_1(z) = F_{1j}(z, -\infty)a_j(-\infty), \tag{18a}$$

$$a_2(z) = F_{2j}(z, -\infty)a_j(-\infty), \tag{18b}$$

which are analogous to (10.8a), (10.8b) in Ref. 7. Substituting (18a), (18b) into the right-hand members of (10a), (10b), we obtain the following formulas:

$$\psi_1(z) = [f_1(z)F_{1j}(z, -\infty) + f_2(z)F_{2j}(z, -\infty)]a_j(-\infty), \tag{19a}$$

$$\psi_1'(z) = [f_1'(z)F_{1j}(z, -\infty) + f_2'(z)F_{2j}(z, -\infty)]a_j(-\infty), \tag{19b}$$

which are valid for all values of the parameter E considered. Analogously we obtain the formulas

$$\psi_2(z) = [f_1(z)F_{1k}(z, +\infty) + f_2(z)F_{2k}(z, +\infty)]a_k(+\infty), \tag{20a}$$

$$\psi_2'(z) = [f_1'(z)F_{1k}(z, +\infty) + f_2'(z)F_{2k}(z, +\infty)]a_k(+\infty), \tag{20b}$$

which are also valid for all values of the parameter E considered. The quantities $a_j(-\infty)$ in (18a), (18b) and (19a), (19b) and $a_k(+\infty)$ in (20a), (20b) may depend on E.

With the aid of the inversion formula (15) we obtain from the multiplication rule (14) the following identity, containing the F-matrix elements appearing in the right-hand members of (18)–(20):

$$F_{3-j,k}(-\infty, +\infty) = (-1)^{j+1}[F_{1j}(z, -\infty)F_{2k}(z, +\infty) - F_{2j}(z, -\infty)F_{1k}(z, +\infty)], \tag{21}$$

where the point z is arbitrary. Formula (21) is also valid when E is not an eigenvalue.

The Wronskian of the solutions $\psi_1(z)$ and $\psi_2(z)$ is easily obtained from (19a), (19b), (20a), (20b), (9), and (21):

$$\psi_1(z)\psi_2'(z) - \psi_2(z)\psi_1'(z) = 2i(-1)^j \times F_{3-j,k}(-\infty, +\infty)a_j(-\infty)a_k(+\infty). \tag{22}$$

The quantities $a_j(-\infty)$ and $a_k(+\infty)$ must be different from zero since otherwise the functions $\psi_1(z)$ and $\psi_2(z)$ would be identically equal to zero. For values of the parameter E such that $F_{3-j,k}(-\infty, +\infty)$ is different from zero, the Wronskian given by (22) is nonzero, and hence in this case $\psi_1(z)$ and $\psi_2(z)$ are linearly independent solutions of (1). For such values of E there exist no nontrivial solution of (1), tending to zero as z tends to both $-\infty$ and $+\infty$ through real values.

Except for Eqs. (5) and (6) the previous formulas in this section are valid even if the parameter E is not an eigenvalue, but in the rest of this section we shall discuss formulas, which apart from the unnumbered equation below are valid only when E fulfils the equation

$$F_{3-j,k}(-\infty, +\infty) = 0. \tag{23}$$

This is an exact quantization condition, selecting the values of the parameter E for which bound states exist. It is a generalization of Eq. (10.13) in Ref. 7 to the case of more than one classically allowed region.

Using (3.18) and (3.13) in Ref. 7 and Eqs. (8) and (21) in the present paper, we obtain the formula

$$\frac{d}{dz}\frac{F_{lk}(z, +\infty)}{F_{lj}(z, -\infty)} = (-1)^{j+1}\tfrac{1}{2}i\epsilon q(z)\exp[(-1)^l 2iw(z)] \times \frac{F_{3-j,k}(-\infty, +\infty)}{[F_{lj}(z, -\infty)]^2},$$

where l can take the values 1 or 2 and where ϵ is a function defined by (3.5) in Ref. 7. We thus find that the quantities $F_{1k}(z,+\infty)/F_{1j}(z,-\infty)$ and $F_{2k}(z,+\infty)/F_{2j}(z,-\infty)$ do not depend on z when E fulfils Eq. (23).

By imposing on the quantities $a_j(-\infty)$ and $a_k(+\infty)$ the relation

$$a_j(-\infty)/a_k(+\infty) = F_{jk}(z,+\infty)/F_{jj}(z,-\infty), \tag{24}$$

we attain the validity of Eq. (5) as is easily seen from (21), (23), (19a), and (20a).

Substituting (22) into (6), and using (23) if $a_j(-\infty) \times a_k(+\infty)$ depends on E, we get the exact formula

$$\int_{-\infty}^{+\infty} [\psi(x)]^2 \frac{\partial}{\partial E} Q^2(x)dx = 2i(-1)^j a_j(-\infty)a_k(+\infty) \times \frac{\partial}{\partial E} F_{3-j,k}(-\infty,+\infty). \tag{25}$$

Since the eigenvalues are nondegenerate, the wavefunction $\psi(x)$ can be written as a real function multiplied by a phase factor, which does not depend on x. Hence, assuming that Q^2 is given by (2) and taking the absolute values of both members in Eq. (25), we get the formula

$$\int_{-\infty}^{+\infty} |\psi(x)|^2 dx = \frac{\hbar^2}{m} \left| a_j(-\infty)a_k(+\infty) \times \frac{\partial}{\partial E} F_{3-j,k}(-\infty,+\infty) \right|. \tag{26}$$

Requiring that

$$\int_{-\infty}^{+\infty} |\psi(x)|^2 dx = 1, \tag{27}$$

we obtain from (26) the normalization condition

$$|a_j(-\infty)a_k(+\infty)|^{-1} = \frac{\hbar^2}{m} \left| \frac{\partial}{\partial E} F_{3-j,k}(-\infty,+\infty) \right|. \tag{28}$$

We emphasize again that the formulas (23)–(28) are only valid when E is an eigenvalue.

Equations (24) and (28) together give exact formulas for $a_j(-\infty)$ and $a_k(+\infty)$. The normalized wavefunction corresponding to an eigenvalue of E can then be obtained by substituting the value of $a_j(-\infty)$ calculated in the above way into (19a) or by substituting the value of $a_k(+\infty)$ into (20a).

3. APPLICATION TO A SINGLE-WELL POTENTIAL

For the case of a single-well potential the function $Q^2(z)$ defined by (2) has precisely two zeros t_1 and t_2 on the real axis. We assume that these zeros are well separated and that no other zeros of $Q^2(z)$ lie on or close to the real axis. The function $q(z)$ in (7a), (7b), and (8) is assumed to be the function $q(z)$ appearing in the higher-order phase-integral approximations discussed in Ref. 2. Precisely as in Ref. 1, we introduce the quantity

$$L = \tfrac{1}{2}\int_{\Gamma_L} q(z)dz, \tag{29}$$

where the contour of integration Γ_L is shown in Fig. 1 of Ref. 1. The quantities L and $\partial L/\partial E$ are positive, and for the first few orders of approximation we can calculate L from (26) in Ref. 1 and $\partial L/\partial E$ from (29) in Ref. 1. According to Chap. 10 in Ref. 7 we shall put $j = k = 2$ in the formulas of Sec. 2 in the present paper. From the symmetry relation (5.9a) in Ref. 7, the inversion formula (15) in the present paper, and the discussion of the single-well potential on pp. 545–47 in Ref. 2, we obtain the symmetry relations

$$F_{12}(x_0,-\infty) = -iF_{22}^*(x_0,-\infty), \tag{30a}$$

$$F_{12}(x_0,+\infty) = i\exp(-2iL)F_{22}^*(x_0,+\infty), \tag{30b}$$

where x_0 is assumed to lie in the classically allowed region. Substituting (30a), (30b) into (21) with $j = k = 2$ and $z = x_0$, we obtain the identity

$$F_{12}(-\infty,+\infty) = iF_{22}^*(x_0,-\infty)F_{22}(x_0,+\infty) \times \left\{1 + \exp\left[-2i\left(L + \arg\frac{F_{22}(x_0,+\infty)}{F_{22}(x_0,-\infty)}\right)\right]\right\}. \tag{31}$$

It should be noted that formulas (29)–(31) are valid for all values of the parameter E. We also remark that in Sec. 2 the point x_0 is arbitrary, but in the present section the position of the point x_0 is restricted to the classically allowed region.

From (31) and (23) with $j = k = 2$ we obtain the exact quantization condition

$$L + \arg[F_{22}(x_0,+\infty)/F_{22}(x_0,-\infty)] = (s + \tfrac{1}{2})\pi, \tag{32}$$

where s is a nonnegative integer. Equation (32) is the same as Eq. (13) in Ref. 2 with $z = x_0$.

Remembering that $j = k = 2$ and using (31) and (32), we obtain from (24) and (28) the exact formulas

$$\frac{a_2(-\infty)}{a_2(+\infty)} = \frac{F_{22}(x_0,+\infty)}{F_{22}(x_0,-\infty)} \tag{33}$$

and

$$|a_2(\pm\infty)|^{-2} = \frac{2\hbar^2}{m}|F_{22}(x_0,\pm\infty)|^2 \times \frac{\partial}{\partial E}\left[L + \arg\frac{F_{22}(x_0,+\infty)}{F_{22}(x_0,-\infty)}\right], \tag{34}$$

where in the last formula we have deleted the absolute value signs since the energy derivative is positive. Formulas (32)–(34) are only valid when E is an eigenvalue. We note that the quantization condition (32) is obtained by putting the quantity to the right of the operator $\partial/\partial E$ in (34) equal to $(s + \frac{1}{2})\pi$. Remembering the inversion rule (15), we see from the estimate (6.13a) in Ref. 7 that we can approximate $F_{22}(x_0,\pm\infty)$ by unity and hence neglect the quantity $\arg[F_{22}(x_0,+\infty)/F_{22}(x_0,-\infty)]$ appearing in (32) and (34). In this way we obtain from (32) and (34) the approximate formulas (16) and (25) in Ref. 1 since the quantity $|C_1|$ in Ref. 1 is equal to $|a_2(-\infty)|$.

4. APPLICATION TO A DOUBLE-WELL POTENTIAL

We now suppose that $Q^2(x)$ has qualitatively the form shown in Fig. 1. The situation shown in Fig. 1a, where $Q^2(z)$ has four real zeros (t_1,t_2,t_3,t_4), is referred to as the sub-barrier case, and the situation shown in Fig. 1b, where $Q^2(z)$ has two real zeros (t_1,t_4) and two complex conjugate zeros (t_2,t_3) close to the real axis, is referred to as the super-barrier case.[8–10] Precisely as in Ref. 9, we introduce the quantities

$$\alpha = -\operatorname{Re}\tfrac{1}{2}\int_{\Gamma_\alpha} q(z)dz, \tag{35a}$$

$$K = \tfrac{1}{2}i\int_{\Gamma_K} q(z)dz, \tag{35b}$$

$$\beta = -\operatorname{Re}\tfrac{1}{2}\int_{\Gamma_\beta} q(z)dz, \tag{35c}$$

where $q(z)$ is given in Refs. 1, 2, or 11 and where the contours of integration are given in our Fig. 1 or in Figs. 1 and 2 of Ref. 9. Since these contours of integration are closed loops, the quantities α, β, K, and $\partial\alpha/\partial E, \partial\beta/\partial E, \partial K/\partial E$ can be obtained from formulas closely analogous to (26) and (29) in Ref. 1. The quantity K is positive in the sub-barrier case and negative in the superbarrier case. In both these cases the quantity $\partial K/\partial E$ is negative, and the quantities $\alpha, \beta, \partial\alpha/\partial E$, and $\partial\beta/\partial E$ are positive. Expressions for the function $w(z)$ on the real axis are given in Sec. 2 of Ref. 9.

According to Ref. 8, we shall put $j = 2$ and $k = 1$ in the formulas in Sec. 2 of the present paper. From the 11 element of the matrix identity (33) in Ref. 8 and the formulas (16)–(18) in Ref. 8, we get, after some calculations, the identity

$$F_{11}(-\infty, +\infty) = 2|F_{11}(-\infty, x_2)F_{11}(x_4, +\infty)| \times \exp[2K + i(\alpha - \beta)][|F_{12}|\cos(\delta_1 + \delta_2) - |F_{22}|\cos(\delta_1 - \delta_2)], \quad (36)$$

where F_{12} stands for $F_{12}(x_2, x_4)$ and F_{22} for $F_{22}(x_2, x_4)$ and where δ_1 and δ_2 are defined by Eqs. (36a), (36b) in Ref. 9, i.e., by

$$\delta_1 = \alpha - \tfrac{1}{2}\pi - \sigma - \arg F_{11}(-\infty, x_2) + \tfrac{1}{2}\arg F_{22}(x_2, x_4), \quad (37a)$$

$$\delta_2 = \beta - \tfrac{1}{2}\pi - \sigma + \arg F_{11}(x_4, +\infty) - \tfrac{1}{2}\arg F_{22}(x_2, x_4), \quad (37b)$$

the quantity σ being defined by Eq. (36c) in Ref. 8, i.e., by

$$\sigma = \tfrac{1}{2}[\arg F_{12}(x_2, x_4) - \tfrac{1}{2}\pi]. \quad (38)$$

Approximate expressions for σ up to the fifth-order phase-integral approximation are given in Ref. 10. According to Eq. (51) in Ref. 11, we have

$$|\sigma| << 1 \quad \text{when } |K| \text{ is sufficiently large.} \quad (39)$$

From the 21 element of the multiplication rule (14) with $z = x_2, z_1 = x_4, z_0 = +\infty$, the inversion rule (15) and the symmetry relations (17) and (18b) in Ref. 8, we obtain the identity

$$\frac{F_{21}(x_2, +\infty)}{F_{22}(x_2, -\infty)} = -i\left|\frac{F_{11}(x_4, +\infty)}{F_{11}(-\infty, x_2)}\right| \times \exp[2K - i(\alpha + \beta)]\{|F_{12}|\exp[i(\delta_1 + \delta_2)] - |F_{22}|\exp[i(\delta_1 - \delta_2)]\}, \quad (40)$$

for the quantity in the right-hand member of (24) with $j = 2, k = 1$, and $z = x_2$.

Formulas (35)–(40) are valid even if E is not an eigenvalue.

With the aid of (36) we see that the exact quantization condition (23) with $j = 2, k = 1$ is

$$|F_{12}|\cos(\delta_1 + \delta_2) - |F_{22}|\cos(\delta_1 - \delta_2) = 0, \quad (41)$$

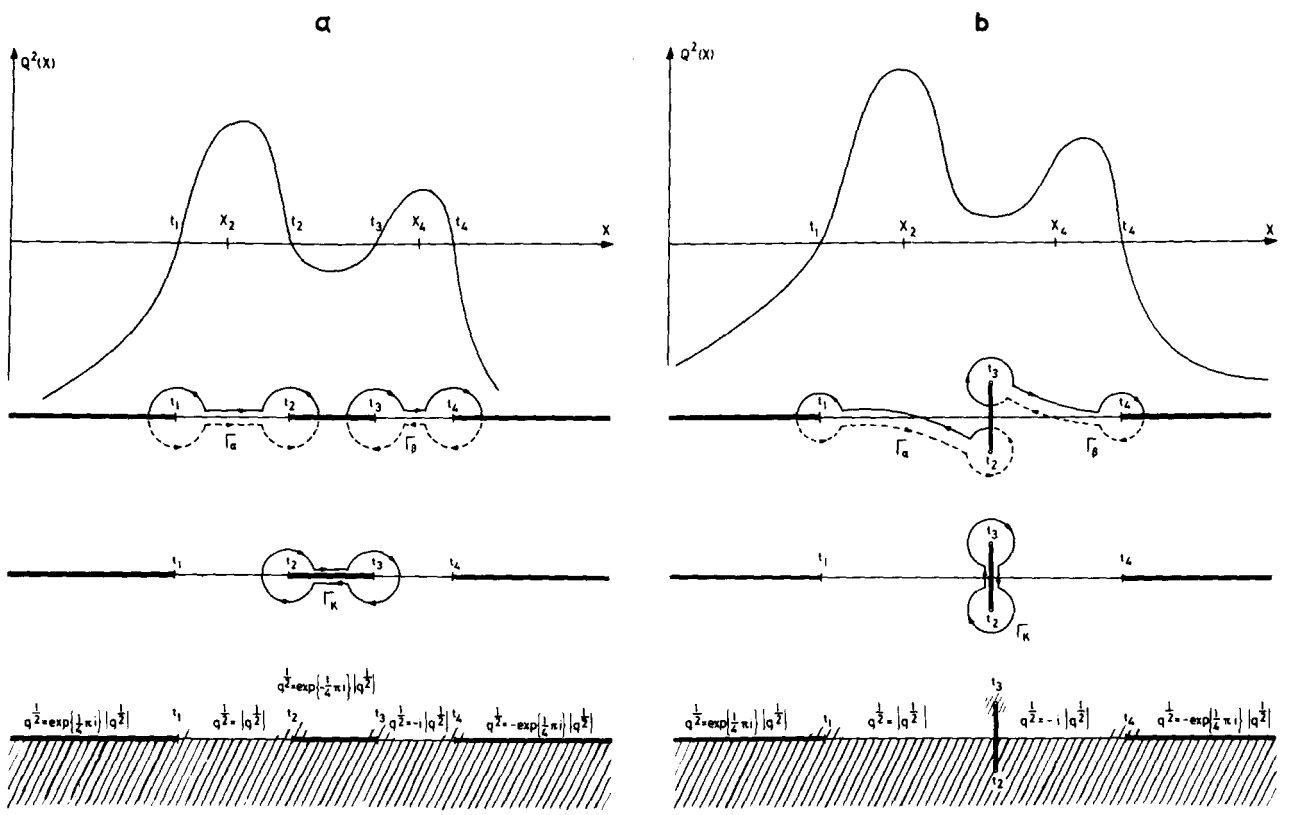


FIG. 1 For the case of a double-well potential this figure shows the qualitative behaviour of $Q^2(x)$ and the contours of integration for obtaining α, β, and K. Cuts are indicated by heavy lines, and the parts of the contours, which lie on the second Riemann sheet, are indicated by dashed lines. For the first Riemann sheet the phase of $q^{1/2}(z)$ is given on the upper lips of the cuts along the real axis. Fig. 1a refers to the sub-barrier case and Fig. 1b to the super-barrier case.

which because of Eq. (55) in Ref. 11 or (29) in Ref. 8 is the same as the quantization condition (35b) in Ref. 8. From (29) in Ref. 8 we find that the following relation holds when (41) is fulfilled:

$$[|F_{12}|\sin(\delta_1+\delta_2)+|F_{22}|\sin(\delta_1-\delta_2)]\times[|F_{12}|\sin(\delta_1+\delta_2)-|F_{22}|\sin(\delta_1-\delta_2)]=\exp(-2K). \tag{42}$$

Remembering that $j=2, k=1$, and using (36), (40), (41), and (42), we obtain from (24) and (28) the exact formulas

$$\frac{a_2(-\infty)}{a_1(+\infty)}=\left|\frac{F_{11}(x_4,+\infty)}{F_{11}(-\infty,x_2)}\right|\exp[2K-i(\alpha+\beta)]\times[|F_{12}|\sin(\delta_1+\delta_2)-|F_{22}|\sin(\delta_1-\delta_2)] \tag{43}$$

and

$$|a_2(-\infty)|^{-2}=\frac{2\hbar^2}{m}|F_{11}(-\infty,x_2)|^2\times\left|\frac{-(\partial/\partial E)[|F_{12}|\cos(\delta_1+\delta_2)-|F_{22}|\cos(\delta_1-\delta_2)]}{|F_{12}|\sin(\delta_1+\delta_2)-|F_{22}|\sin(\delta_1-\delta_2)}\right|, \tag{44a}$$

$$|a_1(+\infty)|^{-2}=\frac{2\hbar^2}{m}|F_{11}(x_4,+\infty)|^2\exp(2K)\times\left|\frac{-(\partial/\partial E)[|F_{12}|\cos(\delta_1+\delta_2)-|F_{22}|\cos(\delta_1-\delta_2)]}{|F_{12}|\sin(\delta_1+\delta_2)+|F_{22}|\sin(\delta_1-\delta_2)}\right|. \tag{44b}$$

We note that the quantization condition (41) is obtained by putting the quantity to the right of the operator $\partial/\partial E$ in (44a), (44b) equal to zero. From (41) it follows that formulas (44a), (44b) can also be written

$$|a_2(-\infty)|^{-2}=\frac{2\hbar^2}{m}|F_{11}(-\infty,x_2)|^2\times\left|\frac{-(\partial/\partial E)(|F_{12}/F_{22}|\cos(\delta_1+\delta_2)-\cos(\delta_1-\delta_2))}{|F_{12}/F_{22}|\sin(\delta_1+\delta_2)-\sin(\delta_1-\delta_2)}\right|, \tag{45a}$$

$$|a_1(+\infty)\exp(K)|^{-2}=\frac{2\hbar^2}{m}|F_{11}(x_4,+\infty)|^2\times\left|\frac{-(\partial/\partial E)(|F_{12}/F_{22}|\cos(\delta_1+\delta_2)-\cos(\delta_1-\delta_2))}{|F_{12}/F_{22}|\sin(\delta_1+\delta_2)+\sin(\delta_1-\delta_2)}\right|. \tag{45b}$$

With the aid of (41) we see that the energy derivative in the numerators of (45a), (45b) can be written

$$-\frac{\partial}{\partial E}\left(\left|\frac{F_{12}}{F_{22}}\right|\cos(\delta_1+\delta_2)-\cos(\delta_1-\delta_2)\right)=\left|\frac{F_{12}}{F_{22}}\right|\left[\sin(\delta_1+\delta_2)\frac{\partial}{\partial E}(\delta_1+\delta_2)+\frac{\partial}{\partial E}\left(\left|\frac{F_{22}}{F_{12}}\right|\cos(\delta_1-\delta_2)\right)\right]. \tag{46}$$

According to (43a), (43b), and (52b) in Ref. 11 and (6.13a) in Ref. 7, we have the approximate formulas

$$F_{22}\approx 1, \tag{47a}$$

$$|F_{12}|\approx[1+\exp(-2K)]^{1/2} \tag{47b}$$

$$\left|\frac{F_{12}}{F_{22}}\right|\approx[1+\exp(-2K)]^{1/2}, \tag{47c}$$

$$F_{11}(-\infty,x_2)\approx 1, \tag{47d}$$

$$F_{11}(x_4,+\infty)\approx 1. \tag{47e}$$

Equation (47c) is a particularly satisfactory approximation when K is positive. Furthermore, we mention that the approximate formulas (47a) and (47c) are in general not very accurate when K takes too large negative values. It is obvious that we cannot have infinitely large positive values of K, since we have two finite potential wells separated by a finite barrier. Similarly we are not interested in taking the limit $K\to-\infty$, since the double-well character of the problem would then be completely lost. Hence, in subsection A below, where we use the estimates (47) and particularize to the case of large absolute values of K, we consider fairly large but not infinitely large values of $|K|$. Furthermore in subsection A below we do not need any accurate formula for $|F_{12}/F_{22}|$ or F_{22} in the discussion of large negative K, but only the rough estimate that $|F_{12}/F_{22}|$ is much greater than unity when K takes fairly large negative values.

Using (47a)–(47e), we obtain from our Eqs. (43), (45a), (45b), (37a), and (37b) the approximate formulas

$$\frac{a_2(-\infty)}{a_1(+\infty)}\approx-\exp[2K-i(\alpha+\beta)]\{[1+\exp(-2K)]^{1/2}\sin(\alpha+\beta-2\sigma)+\sin(\alpha-\beta)\} \tag{48}$$

and

$$|a_2(-\infty)|^{-2}\approx\frac{2\hbar^2}{m}\left|\frac{-(\partial/\partial E)\{[1+\exp(-2K)]^{1/2}\cos(\alpha+\beta-2\sigma)+\cos(\alpha-\beta)\}}{[1+\exp(-2K)]^{1/2}\sin(\alpha+\beta-2\sigma)+\sin(\alpha-\beta)}\right| \tag{49a}$$

$$|a_1(+\infty)\exp(K)|^{-2}\approx\frac{2\hbar^2}{m}\left|\frac{-(\partial/\partial E)\{[1+\exp(-2K)]^{1/2}\cos(\alpha+\beta-2\sigma)+\cos(\alpha-\beta)\}}{[1+\exp(-2K)]^{1/2}\sin(\alpha+\beta-2\sigma)-\sin(\alpha-\beta)}\right|. \tag{49b}$$

Introducing the approximations (47a)–(47e) into the exact quantization condition (41) and Eqs. (37a), (37b), we see that the energy eigenvalues are approximately equal to the zeros of the function to the right of the operator $\partial/\partial E$ in (49a), (49b).

A. Particularization to the Case of Large Absolute Values of K

When K is negative and sufficiently large to its absolute value, retaining the discussion following Eqs. (47), we see from Eqs. (41) and (47c) that the absolute

value of the quantity $\cos(\delta_1 + \delta_2)$ is much smaller than unity. Hence $\delta_1 + \delta_2$ is approximately equal to a half-integral multiple of π and therefore

$$\sin(\delta_1 + \delta_2) \approx -i \exp[i(\delta_1 + \delta_2)] \tag{50}$$

when E is an eigenvalue. We further obtain from (47b) the approximate equation

$$|F_{12}| \approx \exp(-K). \tag{51}$$

From (47c) and the fact that $|\sin(\delta_1 + \delta_2)| \approx 1$ when E is an eigenvalue [cf. Eq. (50)] we see that we can neglect the last term in the denominator of (45a) and (45b), respectively. Similarly, neglecting the last term inside the brackets on the right-hand side of (43) and (46), respectively, substituting (46) into (45a), (45b), neglecting the terms discussed above, and using (47d), (47e), (50), and (51), we obtain the approximate formulas

$$\frac{a_2(-\infty)}{a_1(+\infty)} \approx -i \exp(K) \exp[i(\delta_1 + \delta_2 - \alpha - \beta)] \tag{52}$$

and

$$|a_2(-\infty)|^{-2} \approx |a_1(+\infty) \exp(K)|^{-2} \approx \frac{2\hbar^2}{m} \left|\frac{\partial}{\partial E}(\delta_1 + \delta_2)\right|. \tag{53}$$

Using (37a), (37b), (39), (47d), and (47e), we obtain from (52) and (53) the approximate formulas

$$a_2(-\infty) \approx i \exp(K) a_1(+\infty) \tag{54}$$

and

$$|a_2(-\infty)|^2 \approx |a_1(+\infty) \exp(K)|^{-2} \approx \frac{2\hbar^2}{m} \frac{\partial}{\partial E}(\alpha + \beta), \tag{55}$$

where we have deleted the absolute value signs since the energy derivative of $\alpha + \beta$ is positive. The approximate quantization condition (39) in Ref. 8 is obtained by putting the quantity to the right of the operator $\partial/\partial E$ in (55) equal to a half-integral multiple of π.

To conclude the above discussion, we remark that when K takes sufficiently large negative values, we can consider the double-well potential as a single-well potential. However, since we are regarding a double-well problem, our present phases and definitions are different from those in Sec. 3 of the present paper and in Ref. 1. From the definitions (19) and (21) of α and β in Ref. 8, which are equivalent to our equations (35a), (35c) and from Fig. 2 in Ref. 8 we realize that the right-hand side of (55) is the same as the right-hand side of the approximate normalization condition (25) in Ref. 1 for a single-well potential.

When K is positive and sufficiently large, we see from (47a), (47b) that F_{22} and $|F_{12}|$ are both approximately equal to 1, and hence we find from (41) that either $\sin\delta_1$ or $\sin\delta_2$ is small. We shall exclude the case when both $\sin\delta_1$ and $\sin\delta_2$ are small. Remembering that we can approximate F_{22} and $|F_{12}|$ by unity and using (37a), (37b), (39), and (47d), (47e), we obtain from (43) and (45a) when $\sin\delta_1$ is small, but $\sin\delta_2$ is not too small, the approximate formulas

$$a_1(+\infty) \approx -ia_2(-\infty) \frac{\exp(i\beta - 2K)}{2\cos\beta}, \tag{56a}$$

$$|a_2(-\infty)|^{-2} \approx \frac{2\hbar^2}{m} \frac{\partial\alpha}{\partial E}. \tag{56b}$$

Analogously with the aid of (42) we obtain from (43) and (45b) when $\sin\delta_2$ is small, but $\sin\delta_1$ is not too small, the approximate formulas

$$a_2(-\infty) \approx ia_1(+\infty) \frac{\exp(-i\alpha)}{2\cos\alpha} \tag{57a}$$

$$|a_1(+\infty) \exp(K)|^{-2} \approx \frac{2\hbar^2}{m} \frac{\partial\beta}{\partial E}. \tag{57b}$$

The approximate quantization condition (38a) and (38b) in Ref. 8 are obtained by putting the quantity to the right of the operator $\partial/\partial E$ in (56b) and (57b), respectively, equal to a half-integral multiple of π. We also note that the right-hand members of (56b) and (57b) are the same as the right-hand member of the approximate normalization condition (25) in Ref. 1 with Γ_L replaced by Γ_α and Γ_β, respectively.

B. Particularization to the Case of a Symmetric Double-Well Potential

We shall now specialize our treatment to the case when $V(-x) = V(x)$. If we put $x_2 = -x_4$, the quantities δ_1 and δ_2 become equal to each other, and are given by (47) in Ref. 8. Remembering this formula, substituting (42) in Ref. 8 and (55) in Ref. 11 together with (46) in Ref. 8 into (43) and (45a), (45b), with $\alpha = \beta$ and $\delta_1 = \delta_2$, and using (41) and (42) with $\delta_1 = \delta_2$, we obtain the exact formulas

$$a_1(+\infty) = \pm \exp(-K + 2i\alpha) a_2(-\infty) \tag{58}$$

and

$$|a_2(-\infty)|^{-2} = |a_1(+\infty) \exp(K)|^{-2} = \frac{4\hbar^2}{m} |F_{11}(x_4, +\infty)|^2 \times \left|\frac{\partial}{\partial E}\left(\alpha - \sigma + \arg F_{11}(x_4, +\infty) \mp \tfrac{1}{2}\arctan\frac{\exp(-K)}{F_{22}(-x_4, x_4)}\right)\right|. \tag{59}$$

As in Refs. 8 and 9, the upper sign corresponds to odd parity and the lower sign to even parity of the eigenfunction. We note that the quantization condition (49) in Ref. 8 is obtained by putting the quantities to the right of the operator $\partial/\partial E$ in (59) equal to $(n + \frac{1}{2})\pi$. Substituting (47a), (47e) into (59), we arrive at the following approximate normalization condition:

$$|a_2(-\infty)|^{-2} = |a_1(+\infty) \exp(K)|^{-2} \approx \frac{4\hbar^2}{m} \times \frac{\partial}{\partial E}[\alpha - \sigma \mp \tfrac{1}{2} \arctan \exp(-K)], \tag{60}$$

where we have deleted the absolute value signs on the right-hand side. We note that we obtain the approximate quantization condition (50) in Ref. 8 by putting the quantities to the right of the operator $\partial/\partial E$ in (60) equal to a half-integral multiple of π.

ACKNOWLEDGMENTS

The author is most grateful to Professor Per Olof Fröman, who has suggested the problem and has given advice, encouragement, and constructive criticism in the course of the work. It is also a pleasure to thank Dr. Nanny Fröman for several valuable comments.

* Work supported in part by the Swedish Natural Science Research Council.

1 S. Yngve, J. Math. Phys. 12, 114 (1971).
2 N. Fröman, Arkiv Fysik 32, 541 (1966).
3 N. Fröman, Ann. Phys. (N.Y.) 61, 451 (1970).
4 W. Pauli, "Die allgemeinen Prinzipien der Wellenmechanik," in *Handbuch der Physik*, edited by H. Geiger and K. Scheel (Springer-Verlag, Berlin, 1933), Vol. XXIV/1, p. 173.
5 L. D. Landau and E. M. Lifshitz, *Quantum Mechanics, Nonrelativistic Theory* (Moscow, 1948) [Eng. transl. (Pergamon, London 1958)].
6 W. H. Furry, Phys. Rev. 71, 366 (1947).
7 N. Fröman and P. O. Fröman, *JWKB-Approximation. Contributions to the Theory* (North-Holland, Amsterdam, 1965) [Russian transl. (NER, Moscow, 1967)].
8 N. Fröman, Arkiv Fysik 32, 79 (1966).
9 N. Fröman and U. Myhrman, Arkiv Fysik 40, 497 (1970).
10 N. Fröman, P. O. Fröman, U. Myhrman, and R. Paulsson, to be published.
11 N. Fröman and P. O. Fröman, Nucl. Phys. A147, 606 (1970).

Constants of Motion and Lie Group Actions

M. Andrié
Institut für Theoretische Kernphysik der Universität Bonn, Bonn, Germany

and

D. J. Simms
Department of Pure Mathematics, Trinity College, Dublin, Ireland

(Received 19 October 1970)

A classical Hamiltonian dynamical system with $2N$-dimensional phase space is studied in the case when a Lie algebra of constants of the motion exists which contains $2N-1$ functionally independent elements and when each constant of motion generates a complete integral curve. It is proved that a connected (global) Lie group G acts on the phase space and acts transitively on each connected component of each surface of constant energy. When G is compact, each component of the space of time orbits corresponding to a fixed energy is shown to be a $(2N-2)$-dimensional compact symplectic manifold diffeomorphic to an orbit of G in the dual of the adjoint representation. It is shown that a (global) Lie group does act in the case of the harmonic oscillator, but does not act in the case of the negative energy Kepler problem.

1. INTRODUCTION

Much work has been done on dynamical groups in classical mechanics at the Lie algebra level, but apart from Ref. 1 there appears to have been no attempt to investigate the situation when the "infinitesimal" Lie algebra action can be integrated to give a "global" Lie group action. This is what we propose to do under the following general hypothesis.

We study a classical Hamiltonian dynamical system under the assumption that constants of motion exist which are sufficient to parametrize the set of motions and which span a finite-dimensional Lie algebra L under the Poisson bracket. Such a family exists, of course, in the Kepler problem and the three-dimensional harmonic oscillator. For a global Lie group action to exist, it is necessary that each of the constants of motion generate a vector field which is complete, in the sense that its integral curves are parametrized by $(-\infty, \infty)$. Work of Palais[2] shows that this condition is also sufficient.

Under the above assumptions, then, a connected Lie group G with Lie algebra isomorphic to L acts on the phase space. We prove in this paper that

(i) G acts transitively on each connected component of each surface of constant energy.

(ii) G acts on the time orbits (set of possible motions).

(iii) The time orbits can be mapped into the vector space L' (the dual of L) by a mapping which commutes with the action of G, where G acts on L' by the dual of the adjoint representation.

(iv) When G is compact, each connected component of each surface of constant energy is compact, and the corresponding space of time orbits is a compact symplectic manifold which is diffeomorphic to an orbit of G in the vector space L'. This diffeomorphism commutes with the action of G.

These results are the global equivalents of the results in Ref. 3.

In the case of the harmonic oscillator, a global action of $SU(3)$ does exist, acting transitively on each surface of constant energy. For the negative energy Kepler problem, however, a global action of $SO(4)$ can be obtained only when the phase space is suitably completed, as in Ref. 1. The effect of this completion is to compactify the noncompact surfaces of constant energy.

2. LIE TRANSFORMATION GROUPS

In this section we set out the standard definitions and results on Lie transformation groups which we shall need.

A Lie group G with identity element e is said to be a *Lie transformation group* of a C^∞ (infinitely differentiable) manifold M if a C^∞ map

$$G \times M \to M, \qquad (g,p) \to g \cdot p,$$

is given satisfying

(i) $g_1 \cdot (g_2 \cdot p) = (g_1 g_2) \cdot p,$

(ii) $e \cdot p = p$

for all $g_1, g_2 \in G$ and $p \in M$. We shall say that G *acts on* M when G is a Lie transformation group of M.

Let X be a contravariant vector field (first-order linear differential operator) on M. A differentiable curve $t \mapsto \beta(t)$ in M is said to be *an integral curve of X with initial point p* if

$$\frac{d}{dt} f(\beta(t)) = (Xf)(\beta(t)), \qquad \beta(0) = p,$$

for all $f \in C^\infty(M)$. Here $C^\infty(M)$ denotes the set of real C^∞ functions on M. The vector field X is said to be *complete* if, for each $p \in M$, X has an integral

* Work supported in part by the Swedish Natural Science Research Council.

1 S. Yngve, J. Math. Phys. 12, 114 (1971).
2 N. Fröman, Arkiv Fysik 32, 541 (1966).
3 N. Fröman, Ann. Phys. (N.Y.) 61, 451 (1970).
4 W. Pauli, "Die allgemeinen Prinzipien der Wellenmechanik," in *Handbuch der Physik*, edited by H. Geiger and K. Scheel (Springer-Verlag, Berlin, 1933), Vol. XXIV/1, p. 173.
5 L. D. Landau and E. M. Lifshitz, *Quantum Mechanics, Nonrelativistic Theory* (Moscow, 1948) [Eng. transl. (Pergamon, London 1958)].
6 W. H. Furry, Phys. Rev. 71, 366 (1947).
7 N. Fröman and P. O. Fröman, *JWKB-Approximation. Contributions to the Theory* (North-Holland, Amsterdam, 1965) [Russian transl. (NER, Moscow, 1967)].
8 N. Fröman, Arkiv Fysik 32, 79 (1966).
9 N. Fröman and U. Myhrman, Arkiv Fysik 40, 497 (1970).
10 N. Fröman, P. O. Fröman, U. Myhrman, and R. Paulsson, to be published.
11 N. Fröman and P. O. Fröman, Nucl. Phys. A147, 606 (1970).

Constants of Motion and Lie Group Actions

M. Andrié
Institut für Theoretische Kernphysik der Universität Bonn, Bonn, Germany

and

D. J. Simms
Department of Pure Mathematics, Trinity College, Dublin, Ireland

(Received 19 October 1970)

A classical Hamiltonian dynamical system with $2N$-dimensional phase space is studied in the case when a Lie algebra of constants of the motion exists which contains $2N-1$ functionally independent elements and when each constant of motion generates a complete integral curve. It is proved that a connected (global) Lie group G acts on the phase space and acts transitively on each connected component of each surface of constant energy. When G is compact, each component of the space of time orbits corresponding to a fixed energy is shown to be a $(2N-2)$-dimensional compact symplectic manifold diffeomorphic to an orbit of G in the dual of the adjoint representation. It is shown that a (global) Lie group does act in the case of the harmonic oscillator, but does not act in the case of the negative energy Kepler problem.

1. INTRODUCTION

Much work has been done on dynamical groups in classical mechanics at the Lie algebra level, but apart from Ref. 1 there appears to have been no attempt to investigate the situation when the "infinitesimal" Lie algebra action can be integrated to give a "global" Lie group action. This is what we propose to do under the following general hypothesis.

We study a classical Hamiltonian dynamical system under the assumption that constants of motion exist which are sufficient to parametrize the set of motions and which span a finite-dimensional Lie algebra L under the Poisson bracket. Such a family exists, of course, in the Kepler problem and the three-dimensional harmonic oscillator. For a global Lie group action to exist, it is necessary that each of the constants of motion generate a vector field which is complete, in the sense that its integral curves are parametrized by $(-\infty, \infty)$. Work of Palais[2] shows that this condition is also sufficient.

Under the above assumptions, then, a connected Lie group G with Lie algebra isomorphic to L acts on the phase space. We prove in this paper that

(i) G acts transitively on each connected component of each surface of constant energy.

(ii) G acts on the time orbits (set of possible motions).

(iii) The time orbits can be mapped into the vector space L' (the dual of L) by a mapping which commutes with the action of G, where G acts on L' by the dual of the adjoint representation.

(iv) When G is compact, each connected component of each surface of constant energy is compact, and the corresponding space of time orbits is a compact symplectic manifold which is diffeomorphic to an orbit of G in the vector space L'. This diffeomorphism commutes with the action of G.

These results are the global equivalents of the results in Ref. 3.

In the case of the harmonic oscillator, a global action of $SU(3)$ does exist, acting transitively on each surface of constant energy. For the negative energy Kepler problem, however, a global action of $SO(4)$ can be obtained only when the phase space is suitably completed, as in Ref. 1. The effect of this completion is to compactify the noncompact surfaces of constant energy.

2. LIE TRANSFORMATION GROUPS

In this section we set out the standard definitions and results on Lie transformation groups which we shall need.

A Lie group G with identity element e is said to be a *Lie transformation group* of a C^∞ (infinitely differentiable) manifold M if a C^∞ map

$$G \times M \to M, \qquad (g,p) \to g \cdot p,$$

is given satisfying

(i) $g_1 \cdot (g_2 \cdot p) = (g_1 g_2) \cdot p,$

(ii) $e \cdot p = p$

for all $g_1, g_2 \in G$ and $p \in M$. We shall say that G *acts on* M when G is a Lie transformation group of M.

Let X be a contravariant vector field (first-order linear differential operator) on M. A differentiable curve $t \mapsto \beta(t)$ in M is said to be *an integral curve of X with initial point p* if

$$\frac{d}{dt} f(\beta(t)) = (Xf)(\beta(t)), \qquad \beta(0) = p,$$

for all $f \in C^\infty(M)$. Here $C^\infty(M)$ denotes the set of real C^∞ functions on M. The vector field X is said to be *complete* if, for each $p \in M$, X has an integral

curve $t \mapsto \beta(t)$ with initial point p and parameter domain $-\infty < t < \infty$. When M is compact, every vector field on M is complete.

When G acts on M and G has Lie algebra L, then each $A \in L$ defines a contravariant vector field X_A on M by

$$(X_A f)(p) = \frac{d}{dt} f([\exp(-tA)]\cdot p)\Big|_{t=0}$$

for each $f \in C^\infty(M)$. This satisfies the basic properties

(i) $X_{aA_1+bA_2} = aX_{A_1} + bX_{A_2}$

(ii) $X_{[A_1,A_2]} = X_{A_1}X_{A_2} - X_{A_2}X_{A_1}$

(iii) $(X_A f)([\exp(-tA)]\cdot p) = \frac{d}{dt} f([\exp(-tA)]\cdot p)$.

Thus $A \mapsto X_A$ is a Lie algebra homomorphism from L into the contravariant vector fields on M and each X_A is a complete vector field with integral curve $t \mapsto [\exp(-tA)]\cdot p$ having initial point p.

The preceding paragraph has the following converse due to Palais,[2] which is the global equivalent of a theorem of Lie. Let $A \mapsto X_A$ be a Lie algebra homomorphism from a (real) Lie algebra L into the Lie algebra of contravariant vector fields on M. Suppose that $A_1, \ldots, A_n$ is a basis for L and that $X_{A_1}, \ldots, X_{A_n}$ are complete vector fields. Then there is a connected Lie group G with Lie algebra L and an action of G on M such that

$$(X_A f)(p) = \frac{d}{dt} f([\exp(-tA)]\cdot p)\Big|_{t=0}$$

for all $f \in C^\infty(M)$, $A \in L$, and $p \in M$.

Let X be a complete vector field on M. For each $p \in M$ denote by $t \mapsto t \cdot p$, $\mathbb{R} \to M$, the integral curve of X with initial point p. Then

$$\mathbb{R} \times M \to M, \qquad (t, p) \mapsto t\cdot p,$$

is an action on M of the Lie group $\mathbb{R}$ of real numbers. We call this the *action of* $\mathbb{R}$ *generated by* X.

Suppose that a Lie group G acts on two differentiable manifolds M and N and that $\tau: M \to N$ is a differentiable map which commutes with the action of G:

$$g\cdot\tau(p) = \tau(g\cdot p)$$

for all $g \in G, p \in M$. Let $A \mapsto X_A^M$ and $A \mapsto X_A^N$ be the corresponding homomorphisms from L into the contravariant vector fields on M and N, respectively. Then we have

$$X_A^M(f \circ \tau) = (X_A^N f) \circ \tau \tag{2.1}$$

for each $f \in C^\infty(N)$ and $A \in L$. Conversely, if (2.1) holds for all $A \in L$ and all f belonging to some fixed coordinate system at each point, and if G is connected, then it can be shown that τ commutes with the action of G.

When G acts on M and $p \in M$, then the set

$$G\cdot p = \{g\cdot p \mid g \in G\}$$

of all points of M which can be obtained from p by the action of G is called the G-orbit of p. We denote by

$$M/G = \{G\cdot p \mid p \in M\}$$

the set of all G-orbits. Each point of M belongs to exactly one G-orbit. Each orbit $G\cdot p$ is a submanifold of M whose dimension is equal to the rank at p of the family of vector fields

$$\{X_A \mid A \in L\}.$$

The map $g \mapsto g\cdot p$ is a differentiable map of G onto the orbit $G\cdot p$. When G is compact, it follows that each orbit $G\cdot p$ is compact since it is a continuous image of the compact space G.

3. HAMILTONIAN DYNAMICAL SYSTEMS

Let M be a real C^∞-manifold of even dimension $2N$. A *symplectic form* on M is a nondegenerate differential 2-form Ω (skew-symmetric second-order covariant tensor) whose exterior derivative is zero, $d\Omega = 0$. The pair (M, Ω) is called a *symplectic manifold*. A local coordinate system $q_1, \ldots, q_N, p_1, \ldots, p_N$ on M is called a *canonical coordinate system* if the expression of Ω in these coordinates is

$$\sum_{i=1}^{N} dq_i \wedge dp_i.$$

Each point of M lies in the domain of some canonical coordinate system.

For each $f \in C^\infty(M)$ the differential df is a covariant vector field, and by using Ω as a metric tensor we may pass to the corresponding contravariant vector field which we denote by X_f. We call X_f the vector field generated by f. It is uniquely defined by the equation

$$\Omega(X_f, Y) = df(Y) = Yf$$

for all contravariant Y. In terms of a canonical coordinate system we have

$$df = \sum_{i=1}^{N}\left(\frac{\partial f}{\partial q_i}\, dq_i + \frac{\partial f}{\partial p_i}\, dp_i\right)$$

and

$$X_f = \sum_{i=1}^{N}\left(\frac{\partial f}{\partial p_i}\frac{\partial}{\partial q_i} - \frac{\partial f}{\partial q_i}\frac{\partial}{\partial p_i}\right).$$

For each $f, h \in C^\infty(M)$ we write

$$X_f h = \{f, h\}$$

and call it the *Poisson bracket of* f *and* h. In canonical coordinates we have, locally,

$$\{f, h\} = \sum_{i=1}^{N}\left(\frac{\partial f}{\partial p_i}\frac{\partial h}{\partial q_i} - \frac{\partial f}{\partial q_i}\frac{\partial h}{\partial p_i}\right).$$

Under the Poisson bracket the real vector space $C^\infty(M)$ *becomes a Lie algebra.* The map $f \mapsto X_f$ is a Lie algebra homomorphism of $C^\infty(M)$ into the Lie algebra of contravariant vector fields on M.

Thus

$$X_{af+bh} = aX_f + bX_h,$$

$$X_{\{f,h\}} = X_f X_h - X_h X_f$$

for all $f, h \in C^\infty(M)$.

We shall call a function $H \in C^\infty(M)$ a *Hamiltonian function* if it has no critical points (dH nowhere zero)

and if X_H is a complete vector field on M. We shall call the triple (M, Ω, H) a (conservative time-independent) *Hamiltonian dynamical system*. M is called the *phase space* of the system. The level sets of H are called the surfaces of constant energy. The action of $\mathbb{R}$ on M which is generated by X_H is called the motion in time. The corresponding orbit

$$\mathbb{R}\cdot p = \{t\cdot p \mid t \in \mathbb{R}\}$$

is called the *time orbit* with initial point p. Thus we have

$$\frac{d}{dt} f(t\cdot p) = (X_H f)(t\cdot p) = \{H, f\}(t\cdot p) \tag{3.1}$$

for all $t \in \mathbb{R}$, $f \in C^\infty(M)$, and $p \in M$.

4. LIE ALGEBRAS OF CONSTANTS OF MOTION

In the following two sections we shall repeatedly use the standard results and definitions of the previous two sections.

Let (M, Ω, H) be a Hamiltonian dynamical system with $\dim M = 2N$. Let L be a finite-dimensional Lie subalgebra of $C^\infty(M)$ and let $f_1, \ldots, f_n$ be a basis of L. Suppose that

(i) $\{H, f_k\} = 0$ for $k = 1, \ldots, n$,

(ii) $df_1, \ldots, df_n$ span a $(2N-1)$-dimensional vector space at each point of M,

(iii) each f_k generates a complete vector field.

Condition (i) ensures by (3.1) that the elements of L are constant on the time orbits and are thus *constants of motion*. Condition (ii) ensures that the level sets of the family of functions in L are one dimensional and therefore coincide locally with the time orbits. We shall call L a *full Lie system of constants of motion*.

A basis-free formulation of condition (ii) may be given as follows. For each $f \in L$ and each $l \in L'$, the space of linear forms on L, denote by $\langle f, l\rangle$ the value of l at f. Let

$$\tau: M \to L$$

be the map defined by

$$\langle f, \tau(p)\rangle = f(p)$$

for all $p \in M, f \in L$.

We note that for each $f \in L$

$$(f \circ \tau)(p) = \langle f, \tau(p)\rangle = f(p)$$

so that

$$f \circ \tau = f. \tag{4.1}$$

When $f_1, \ldots, f_n$ is taken as a basis for L, we see that $\tau(p)$ has coordinates

$$(f_1(p), \ldots, f_n(p))$$

relative to the dual basis for L' since $\langle f_i, \tau(p)\rangle = f_i(p)$. Condition (ii) is therefore equivalent to

$\tau: M \to L'$ has rank $2N-1$ at each point.

Theorem 1: Let L be a full Lie system of constants of motion of a Hamiltonian dynamical system (M, Ω, H). Then:

(a) A connected Lie group G with Lie algebra L acts on the phase space M.

(b) Each G-orbit is a connected component of a surface of constant energy.

(c) The map $\tau: M \to L'$ commutes with the action of G, where G acts on L' via the dual of the adjoint representation.

Proof:

(a) Consider the Lie algebra homomorphism $f \mapsto X_f$ from L into the Lie algebra of vector fields on M. Since $X_{f_1}, \ldots, X_{f_n}$ are complete vector fields, it follows from the theorem of Palais that there is a connected Lie group G with Lie algebra L and an action of G on M such that

$$\frac{d}{dt} h([\exp(-tf)]\cdot p) = (X_f h)([\exp(-tf)]\cdot p) = \{f, h\}([\exp(-tf)]\cdot p)$$

for all $f \in L$, $h \in C^\infty(M)$, $p \in M$, and $t \in \mathbb{R}$.

(b) Let $p \in M$ and consider the orbit $G\cdot p$ under the action of G. For each $f \in L$ we have

$$\frac{d}{dt} H([\exp(-tf)]\cdot p) = \{f, H\}([\exp(-tf)]\cdot p) = 0$$

by (i). Thus $H([\exp(-tf)]\cdot p) = H(p)$. Each element of G is a product of exponentials since G is connected, and therefore $H(g\cdot p) = H(p)$ for all $g \in G$. This proves that H is constant on each orbit $G\cdot p$, and therefore G acts on each surface of constant energy.

Since Ω is nondegenerate, the rank of the family of contravariant vector fields $\{X_f \mid f \in L\}$ is the same as the rank of the family of covariant vector fields $\{df \mid f \in L\}$ at each point and this is equal to $2N-1$ by (ii). Therefore each orbit $G\cdot p$ is a $(2N-1)$-dimensional manifold and is thus a connected open submanifold of the corresponding surface of constant energy. This proves that each orbit $G\cdot p$ is a connected component of a surface of constant energy.

(c) Finally for all $f, h \in L$ we have

$$X_f(h \circ \tau) = X_f h = \{f, h\} = \{f, h\} \circ \tau$$

by (4.1). Since G is connected and since any basis for L is a coordinate system for L', if follows that τ commutes with the action of G, where G acts on L' via the dual of the adjoint representation.

This completes the proof of the theorem.

5. THE STRUCTURE OF THE SPACE OF TIME ORBITS

Let (M, Ω, H) be a Hamiltonian dynamical system and let L be a full Lie system of constants of motion. Since

$$\frac{d}{dt} H(t\cdot p) = \{H, H\}(t\cdot p) = 0$$

for all $t \in \mathbb{R}$ and $p \in M$ it follows that H is constant on the time orbit $\mathbb{R}\cdot p$. The motion in time therefore gives an action of $\mathbb{R}$ on each connected component $G\cdot p$ of each surface of constant energy.

The map $\tau: M \to L'$ introduced in the previous section is constant on each time orbit since each element of L is a constant of motion. Since τ has rank $2N-1$ at each point, it follows that the level set of τ through p coincides locally with the time orbit $\mathbb{R}\cdot p$ which is contained in $G\cdot p$. The restriction of τ to the $(2N-1)$-dimensional submanifold $G\cdot p$ therefore has rank $2N-2$ at each point. We can therefore choose an open neighborhood U of p in $G\cdot p$ which is cubical with respect to a coordinate system of the form

$$f_{i_1}, \ldots, f_{i_{2N-2}}, h.$$

The time orbits in U are contained in the level sets of $f_1, \ldots, f_n$, and it follows that each time orbit intersects U at most once and in a flat one-dimensional segment. An application of a result of Palais[4] now gives:

Theorem 2: Let L be a full Lie system of constants of motion of a Hamiltonian dynamical system, and let G be the connected Lie group of Theorem 1. Suppose that G is compact. Then:

(a) The space $G\cdot p/\mathbb{R}$ of time orbits contained in any connected component $G\cdot p$ of any surface of constant energy is a compact manifold of dimension $2N-2$. The time orbits contained in $G\cdot p$ are all compact and diffeomorphic to a circle.

(b) The natural projection

$$\pi: G\cdot p \to G\cdot p/\mathbb{R}$$

is a fibering, the fibres being the time orbits.

(c) The functions $f_{i_1}, \ldots, f_{i_{2N-2}}$ provide a local coordinate system on $G\cdot p/\mathbb{R}$ in a neighborhood of $\pi(p)$.

Proof: Since

$$X_f X_H - X_H X_f = X_{\{f,H\}} = 0$$

for all $f \in L$, it follows that the actions on M of G and of $\mathbb{R}$ commute. We therefore have an action of G on the space $M/\mathbb{R}$ of time orbits. Since the map $\tau: M \to L'$ is constant on time-orbits and commutes with the action of G, we have an induced map

$$M/\mathbb{R} \to L', \qquad \mathbb{R}\cdot p \mapsto \tau(p),$$

which commutes with the action of G.

If $p \in M$ and $\tau(p) = l$, then the orbit $G\cdot p$ in M is mapped by τ onto the orbit $G\cdot l$ in L'. The resulting map

$$G\cdot p/\mathbb{R} \to G\cdot l$$

has rank $2N-2$, which is equal to the dimension of the manifold $G\cdot p/\mathbb{R}$.

It follows that $G\cdot p/\mathbb{R}$ is a covering space of the orbit $G\cdot l$. By a proposition we shall prove in the next paragraph, $G\cdot l$ is simply connected and the map $G\cdot p/\mathbb{R} \to G\cdot l$ is therefore a diffeomorphism (bijective). Since $G\cdot l$ is a submanifold of the vector space L', this gives a "linearization" of the space $G\cdot p/\mathbb{R}$ of time orbits. This is the global equivalent of the results of Ref. 3.

The following proposition together with its proof was pointed out to us by Professor Steven Halperin. For a compact semisimple group it is a known result.

Proposition: Let G be a compact connected Lie group, and let G act on the dual L' of its Lie algebra L by the dual of the adjoint representation. Then the orbits of G in L' are all simply connected.

Proof: Since G is compact, there exists a G-invariant inner product on L, and this provides a G-invariant isomorphism between L and L'. We therefore need only show that the G-orbits in L are all simply connected.

We have[5] $L = L_1 \oplus Z$ where Z is the center of L and $L_1 = [L, L]$ is a compact semisimple subalgebra. Let H be the connected Lie subgroup of G having L_1 as Lie algebra. Then H is a semisimple Lie group with a compact simply connected covering group $\tilde{H}$.[6]

For any x in L we have $x = y + z$, where $y \in L_1$ and $z \in Z$. The translation of L by z maps y to x and commutes with the action of G since $\mathrm{ad}(g)z = z$ for all g in G. The translation by z therefore maps the orbit $G\cdot y$ diffeomorphically onto the orbit $G\cdot x$. Now each g in G is of the form $g = hk$, where $h \in H$ and k is in the center of G. Therefore

$$\mathrm{ad}(g)y = \mathrm{ad}(h)\mathrm{ad}(k)y = \mathrm{ad}(h)y.$$

This shows that $G\cdot y = H\cdot y$, which in turn is equal to $\tilde{H}\cdot y$, which is diffeomorphic to $\tilde{H}/\tilde{H}_y$, where $\tilde{H}_y$ is the isotropy group of y in $\tilde{H}$. Since $\tilde{H}$ is simply connected, the quotient $\tilde{H}/\tilde{H}_y$ will be simply connected provided that $\tilde{H}_y$ is connected.[7] It remains therefore to prove that $\tilde{H}_y$ is connected. Since

$$\begin{aligned}\tilde{H}_y &= \{h \in \tilde{H} \mid \mathrm{ad}(h)y\} \\ &= \{h \in \tilde{H} \mid h(\exp(ty))h^{-1} = \exp(ty), \text{ all } t \in \mathbb{R}\},\end{aligned}$$

we see that $\tilde{H}_y$ is the centralizer of

$$\{\exp(ty) \mid t \in \mathbb{R}\}.$$

The closure of the latter is a compact Abelian connected group (torus) and its centralizer $\tilde{H}_y$ is connected.[8] This completes the proof of the proposition.

Each contravariant tangent vector X to $G\cdot p$ at p is mapped by π to a tangent vector $\pi_* X$ to $G\cdot p/\mathbb{R}$ at $\pi(p)$. Moreover $\pi_* X$ is zero if and only if Xf is zero for all $f \in L$, since the local coordinate functions $f_{i_1}, \ldots, f_{i_{2N-2}}$ belong to L. It follows that for each $h \in L$ we have $\Omega(X_f, X_h) = X_h f = 0$ for all $f \in L$ if and only if $\pi_* X_h = 0$. Therefore, Ω induces a nondegenerate symplectic form $\tilde{\Omega}$ on $G\cdot p/\mathbb{R}$ defined by

$$\tilde{\Omega}(\pi_* X_f, \pi_* X_h)(\pi(p)) = \Omega(X_f, X_h)(p) = \{h, f\}(p) \tag{5.1}$$

We have thus proved:

Theorem 3: Let the conditions of Theorem 2 hold. Then:

(a) there is a diffeomorphism

$$G\cdot p/\mathbb{R} \to G\cdot l$$

commuting with the action of G, from the space $G\cdot p/\mathbb{R}$ onto an orbit of G in the vector space L', where G acts on L' by the dual of the adjoint representation.

(b) $G\cdot p/\mathbb{R}$ is a symplectic manifold of dimension $2N-2$.

We note that we have a commutative diagram

$$\begin{array}{ccc} G\cdot p & & \\ {\scriptstyle\pi}\downarrow & \searrow^{\tau} & \\ G\cdot p/\mathbb{R} & \to & G\cdot l \end{array}$$

and (5.1) may be expressed as

$$\tilde{\Omega}(\pi_* X_f, \pi_* X_h)(\pi(p)) = \langle \{h, f\}, \tau(p) \rangle.$$

This shows that the symplectic form $\tilde{\Omega}$ on $G\cdot p/\mathbb{R}$ corresponds to the G-invariant symplectic form on $G\cdot l$ which has been introduced by Kostant and Souriau.[9]

6. THE KEPLER PROBLEM AND THE HARMONIC OSCILLATOR

We now discuss two well-known cases in the light of the foregoing analysis.

In the *Kepler problem*, with negative energy, the phase space is a subset of $\mathbb{R}^6$ with canonical coordinates

$$(\mathbf{q}, \mathbf{p}) = (q_1, q_2, q_3, p_1, p_2, p_3)$$

and Hamiltonian

$$H = \tfrac{1}{2}\mathbf{p}^2 - 1/|\mathbf{q}|.$$

The phase space is the subset given by $H < 0$. The angular momentum functions

$$\mathbf{L} = (L_1, L_2, L_3) = \mathbf{q} \times \mathbf{p}$$

and the Pauli–Lenz functions

$$\mathbf{B} = (B_1, B_2, B_3) = (1/\sqrt{-2H})(\mathbf{L} \times \mathbf{p} + \mathbf{q}/|\mathbf{q}|)$$

form the basis of a Lie algebra of constants of motion under the Poisson bracket. This algebra is isomorphic to the algebra of the orthogonal group $SO(4)$, and any connected Lie group having this Lie algebra is necessarily compact. The surfaces of constant energy

$$\tfrac{1}{2}\mathbf{p}^2 - 1/|\mathbf{q}| = -k^2$$

are, however, connected and noncompact and therefore cannot admit a transitive action of a compact group. It follows from Theorem 1 that $\mathbf{L}$ and $\mathbf{B}$ do not span a full Lie system of constants of motion. The reason why is that the Pauli–Lenz functions $\mathbf{B}$ do not generate complete vector fields, and the local Lie group action does not, therefore, extend to a global action.

It is, however, interesting to note that it is possible to adjoin points to the phase space of the negative energy Kepler problem in such a way that the compact group $SO(4)$ does act on the resulting extended phase space. This adjunction of points may be carried out by parametrizing the phase space for large $|\mathbf{p}|$ and small $|\mathbf{q}|$ by the seven functions

$$\mathbf{L}(\mathbf{q}, \mathbf{p}), \quad \mathbf{B}(\mathbf{q}, \mathbf{p}), \quad \exp[i\sigma(\mathbf{q}, \mathbf{p})]$$

subject to $\mathbf{L}\cdot\mathbf{B} = 0$. Here $\sigma(\mathbf{q}, \mathbf{p})$, denotes the time taken by a moving body to reach the perihelion from an initial state $(\mathbf{q}, \mathbf{p})$, normalized so that $\sigma(q, p)$ is defined modulo 2π. The adjoined points are then defined to be those parametrized by

$$\mathbf{L} = \mathbf{0}, \quad \sigma = 0.$$

These may be thought of as limiting states as particles moving on rectilinear orbits approach the centre of force. It is shown by Bacry, Ruegg and Souriau[1] that $SO(4)$ acts on this extended phase space.

In the case of the *harmonic oscillator* the phase space is $\mathbb{R}^6 - \{\mathbf{o}\}$ with canonical coordinates

$$(\mathbf{q}, \mathbf{p}) = (q_1, q_2, q_3, p_1, p_2, p_3)$$

and Hamiltonian

$$H = \tfrac{1}{2}(\mathbf{p}^2 + \mathbf{q}^2).$$

The angular momentum functions

$$\mathbf{L} = (L_1, L_2, L_3) = \mathbf{q} \times \mathbf{p}$$

and the functions

$$t_{ij} = p_i p_j + q_i q_j, \quad i, j = 1, 2, 3,$$

span a Lie algebra of constants of motion under the Poisson bracket, isomorphic to the algebra of $U(3)$, and their differentials have rank 5 at each point. The angular momentum functions generate complete vector fields since they correspond to the action of the rotation group $SO(3)$. The function t_{ij} generates a vector field whose integral curves are solutions of

$$\frac{dp_k}{ds} = \{t_{ij}, p_k\} = -\delta_{jk}q_i - \delta_{ik}q_j,$$

$$\frac{dq_k}{ds} = \{t_{ij}, q_k\} = \delta_{jk}p_i + \delta_{ik}p_j.$$

This is a system of linear first-order differential equations with constant coefficients and the integral curves are therefore parametrized by $-\infty < s < \infty$. Thus each t_{ij} generates a complete vector field, and we have a full Lie system of constants of motion. Theorem 1 shows that $U(3)$ acts transitively on each surface of constant nonzero energy:

$$\mathbf{p}^2 + \mathbf{q}^2 = k^2.$$

ACKNOWLEDGMENTS

We would like to thank Professors K. Bleuler and S. Halperin for stimulating discussions. One of us (D. J. S.) would like to thank the Instutut für Theoretische Kernphysik at Bonn for hospitality at the time of this work.

---

[1] H. Bacry, H. Ruegg, and J. Souriau, Commun. Math. Phys. 3, 323 (1966).
[2] R. S. Palais, Mem. Am. Math. Soc. No. 22 (1957), Theorem III, p. 95, and Theorem IV, p. 98.
[3] M. Andrié and K. Bleuler, Commun. Phys. Math. Soc. Sci. Fenn. **34**, No. 15 (1969).
[4] See Ref. 2, Theorem V, p. 14, and Corollary 5, p. 22.
[5] S. Helgason, *Differential Geometry and Symmetric Spaces*

(Academic, New York, 1962), p. 122, Proposition 6. 6.
[6] See Ref. 5, p. 123, Theorem 6. 9.
[7] A simple proof of this standard fact is contained in the last nine lines of p. 178 of Ref. 5.
[8] See Ref. 5, p. 247, Corollary 2. 8.
[9] B. Kostant, Proceedings of the United States Japan Seminar in Differential Geometry (Kyoto, 1965); J. M. Souriau, *Structures des systèmes dynamique* (Dunod, Paris, 1970), p. 116; or D. J. Simms, Projective Representations, Symplectic Manifolds and Extensions of Lie Algebras (C. N. R. S. notes 69/P. 300, Marseille, 1969), p. 34.

Scattering from a Periodic Corrugated Structure. II. Thin Comb with Hard Boundaries

John A. DeSanto
Naval Research Laboratory, Washington, D.C. 20390
(Received 15 August 1971)

The scattered field is calculated for plane wave incidence on a periodic rectangularly corrugated surface (thin comb grating) with hard (Neumann) boundary conditions. Except for the hard boundary (and the consequent representation of the field in the comb wells), the formalism is similar to that of a previous paper [J. Math. Phys. 12, 1913 (1971)]. Reflection coefficients are plotted, grating anomalies illustrated, and a correspondence between reflection coefficients and amplitude phases (as a function of corrugation depth) is illustrated.

1. INTRODUCTION

In a previous paper,[1] we discussed the calculation of the scattered field when a plane wave was incident on a (one-dimensional) periodic corrugated surface consisting of an infinite number of periodically spaced, infinitesimally thin parallel plates having a finite depth (diffraction by a thin comb grating). The soft (Dirichlet) boundary condition was used. In this paper, exactly the same geometric surface is again used, but the hard (Neumann) boundary value problem is considered. Details and notation are similar in both papers. Hurd[2] has previously discussed this problem approximately, and from a surface wave point of view. He used the residue calculus method to obtain his analytically approximate solution. It is possible to remove the more restrictive analytical approximation by replacing it with a numerical approximation having a uniformly high accuracy over any parameter domain. The numerical approximation consists in a modification of the residue calculus method due to Mittra[3] and is used here as in Paper I. For a discussion of some of the grating anomalies to be expected from surfaces like this, reference is made to the results of Stewart and Gallaway,[4] Hessel and Oliner,[5] and Tseng.[6]

The formalism of the problem is presented in Sec. 2 and is similar to I except for the representation of the fields in the wells and the hard (Neumann) boundary condition. Linear equations are derived which relate the field amplitude coefficients in the different geometric regions. The modified residue calculus method enables us to solve the exact set of equations approximately, whereas Hurd[2] had to approximate the original equations, as well as further approximate their solution. The amplitude coefficients are expressed as either values or residues of a constructed meromorphic function.

Section 3 contains the results and discussion. Many of the parameter values are chosen to compare and contrast with results in I. Reflection coefficients and amplitude phases are plotted with respect to incident frequency, incident angle, and corrugation depth. Observable grating anomalies include the Rayleigh anomaly, Wood S-anomaly, and Brewster angle anomaly,[4-6] as well as the correspondence between reflection coefficients and amplitude phases, as a function of depth. The latter behavior was also observed in I. Finally, the periodicity of the reflection coefficients and amplitude phases as a function of depth is illustrated.

2. FORMALISM

The problem is to calculate the full velocity potential field ψ when a plane wave ψ_i (at angle θ_i) is incident on the periodic (period $2l$) rectangularly corrugated surface illustrated in Fig. 1, and consisting of thin parallel plates of finite depth d. The surface $S(x)$ is given by

$$S(x) = \begin{cases} -d & x \neq 2ml \quad m = 0, \pm 1, \cdots \\ 0 & x = 2ml \end{cases} \tag{2.1}$$

Region A is $z \geq 0$ and B is $-d \leq z \leq 0$. In both regions, ψ satisfies the Helmholtz equation ($e^{-i\omega t}$ assumed, $k = 2\pi/\lambda$, λ = incident wavelength)

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2} + k^2\right)\psi(x, z) = 0. \tag{2.2}$$

The geometry and general procedure are similar to I, but here the hard (Neumann) boundary condition is used, viz.,

$$\frac{\partial}{\partial n}\,\psi[x, S(x)] = 0, \tag{2.3}$$

where n is the surface normal. ψ has the same restrictions as in Sec. 2 of I.

The field representation for $z \geq 0$ (region A) is

$$\psi_A(x, z) = e^{ik(\alpha_0 x - \beta_0 z)} + \sum_{n=-\infty}^{\infty} A_n^h e^{ik(\alpha_n x + \beta_n z)}. \tag{2.4}$$

The notation is explained in I. The superscript "h" on A_n indicates the hard boundary condition.

The field representation for $0 \geq z \geq -d$ (region B) and $0 \leq x \leq 2l$ is

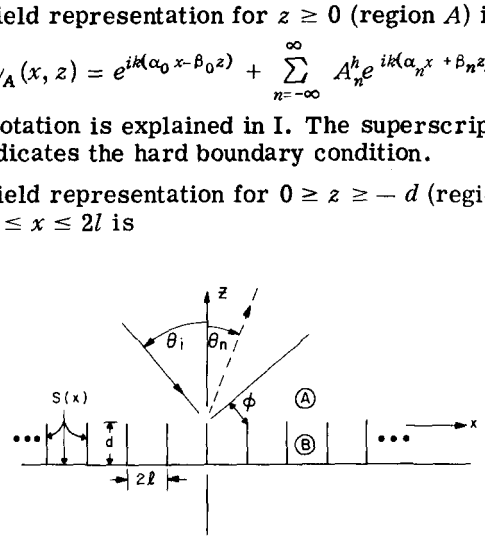


FIG. 1. Plane wave (angle θ_i) incident on a corrugated thin comb grating with period $2l$ and depth d. The angles of the scattered plane waves are θ_n. Region A is $z \geq 0$, region B is $-d \leq z \leq 0$, and ϕ is a phase lag.

(Academic, New York, 1962), p. 122, Proposition 6. 6.
[6] See Ref. 5, p. 123, Theorem 6. 9.
[7] A simple proof of this standard fact is contained in the last nine lines of p. 178 of Ref. 5.
[8] See Ref. 5, p. 247, Corollary 2. 8.
[9] B. Kostant, Proceedings of the United States Japan Seminar in Differential Geometry (Kyoto, 1965); J. M. Souriau, *Structures des systèmes dynamique* (Dunod, Paris, 1970), p. 116; or D. J. Simms, Projective Representations, Symplectic Manifolds and Extensions of Lie Algebras (C. N. R. S. notes 69/P. 300, Marseille, 1969), p. 34.

Scattering from a Periodic Corrugated Structure. II. Thin Comb with Hard Boundaries

John A. DeSanto
Naval Research Laboratory, Washington, D.C. 20390
(Received 15 August 1971)

The scattered field is calculated for plane wave incidence on a periodic rectangularly corrugated surface (thin comb grating) with hard (Neumann) boundary conditions. Except for the hard boundary (and the consequent representation of the field in the comb wells), the formalism is similar to that of a previous paper [J. Math. Phys. 12, 1913 (1971)]. Reflection coefficients are plotted, grating anomalies illustrated, and a correspondence between reflection coefficients and amplitude phases (as a function of corrugation depth) is illustrated.

1. INTRODUCTION

In a previous paper,[1] we discussed the calculation of the scattered field when a plane wave was incident on a (one-dimensional) periodic corrugated surface consisting of an infinite number of periodically spaced, infinitesimally thin parallel plates having a finite depth (diffraction by a thin comb grating). The soft (Dirichlet) boundary condition was used. In this paper, exactly the same geometric surface is again used, but the hard (Neumann) boundary value problem is considered. Details and notation are similar in both papers. Hurd[2] has previously discussed this problem approximately, and from a surface wave point of view. He used the residue calculus method to obtain his analytically approximate solution. It is possible to remove the more restrictive analytical approximation by replacing it with a numerical approximation having a uniformly high accuracy over any parameter domain. The numerical approximation consists in a modification of the residue calculus method due to Mittra[3] and is used here as in Paper I. For a discussion of some of the grating anomalies to be expected from surfaces like this, reference is made to the results of Stewart and Gallaway,[4] Hessel and Oliner,[5] and Tseng.[6]

The formalism of the problem is presented in Sec. 2 and is similar to I except for the representation of the fields in the wells and the hard (Neumann) boundary condition. Linear equations are derived which relate the field amplitude coefficients in the different geometric regions. The modified residue calculus method enables us to solve the exact set of equations approximately, whereas Hurd[2] had to approximate the original equations, as well as further approximate their solution. The amplitude coefficients are expressed as either values or residues of a constructed meromorphic function.

Section 3 contains the results and discussion. Many of the parameter values are chosen to compare and contrast with results in I. Reflection coefficients and amplitude phases are plotted with respect to incident frequency, incident angle, and corrugation depth. Observable grating anomalies include the Rayleigh anomaly, Wood S-anomaly, and Brewster angle anomaly,[4-6] as well as the correspondence between reflection coefficients and amplitude phases, as a function of depth. The latter behavior was also observed in I. Finally, the periodicity of the reflection coefficients and amplitude phases as a function of depth is illustrated.

2. FORMALISM

The problem is to calculate the full velocity potential field ψ when a plane wave ψ_i (at angle θ_i) is incident on the periodic (period $2l$) rectangularly corrugated surface illustrated in Fig. 1, and consisting of thin parallel plates of finite depth d. The surface $S(x)$ is given by

$$S(x) = \begin{cases} -d & x \neq 2ml \quad m = 0, \pm 1, \cdots \\ 0 & x = 2ml \end{cases} \tag{2.1}$$

Region A is $z \geq 0$ and B is $-d \leq z \leq 0$. In both regions, ψ satisfies the Helmholtz equation ($e^{-i\omega t}$ assumed, $k = 2\pi/\lambda$, λ = incident wavelength)

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2} + k^2\right)\psi(x, z) = 0. \tag{2.2}$$

The geometry and general procedure are similar to I, but here the hard (Neumann) boundary condition is used, viz.,

$$\frac{\partial}{\partial n}\,\psi[x, S(x)] = 0, \tag{2.3}$$

where n is the surface normal. ψ has the same restrictions as in Sec. 2 of I.

The field representation for $z \geq 0$ (region A) is

$$\psi_A(x, z) = e^{ik(\alpha_0 x - \beta_0 z)} + \sum_{n=-\infty}^{\infty} A_n^h e^{ik(\alpha_n x + \beta_n z)}. \tag{2.4}$$

The notation is explained in I. The superscript "h" on A_n indicates the hard boundary condition.

The field representation for $0 \geq z \geq -d$ (region B) and $0 \leq x \leq 2l$ is

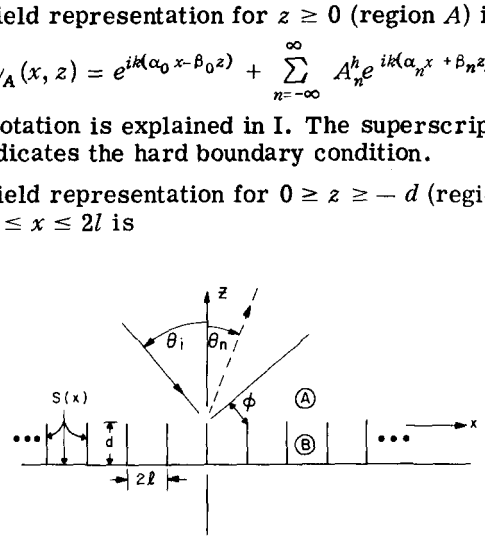


FIG. 1. Plane wave (angle θ_i) incident on a corrugated thin comb grating with period $2l$ and depth d. The angles of the scattered plane waves are θ_n. Region A is $z \geq 0$, region B is $-d \leq z \leq 0$, and ϕ is a phase lag.

$$\psi_B(x, t) = \sum_{j=0}^{\infty} B_j^h \cos(kp_j x) \cos\{kq_j(z + d)\}, \quad (2.5)$$

where $p_j = j\Lambda/2$ and $p_j^2 + q_j^2 = 1$. This satisfies Eq. (2.3) at $x = 0, 2l$ and at $z = -d$. For values of x outside this range, ψ_B is given by the Floquet condition in I.

The continuity conditions are

$$\psi_A(x, 0) = \psi_B(x, 0), \quad (2.6)$$

$$\frac{\partial \psi_A}{\partial z}(x, 0) = \frac{\partial \psi_B}{\partial z}(x, 0). \quad (2.7)$$

These are expressions derived from the physical continuity conditions of pressure and normal velocity.

Substituting the field representations in Eqs. (2.6) and (2.7), projecting out the B_j^h amplitudes, and rearranging the resulting equations yield the following set of equations relating the amplitude coefficients

$$\sum_{n=-\infty}^{\infty} \alpha_n A_n^h \left(\frac{e^{-iq_j kd}}{\beta_n - q_j} \pm \frac{e^{iq_j kd}}{\beta_n + q_j} \right) - \alpha_0 \left(\frac{e^{-iq_j kd}}{\beta_0 + q_j} \pm \frac{e^{iq_j kd}}{\beta_0 - q_j} \right) = \begin{pmatrix} 0 \\ \tilde{B}_j^h \end{pmatrix}, \quad (2.8)$$

where

$$\tilde{B}_j^h \equiv 2\pi i q_j B_j^h \Lambda^{-1} [1 - (-)^j e^{2\pi i \alpha_0/\Lambda}]^{-1}. \quad (2.9)$$

The upper (lower) sign in Eq. (2.8) is to be read with the 0 or $\tilde{B}_j^h$ term on the right-hand side.

Consider contour integrals of the form

$$(2\pi i)^{-1} \int_C d\omega F(\omega) \left(\frac{e^{-iq_j kd}}{\omega - q_j} \pm \frac{e^{iq_j kd}}{\omega + q_j} \right) = 0, \quad (2.10)$$

where C is a closed contour at infinity, and the meromorphic function $F(\omega)$ has properties

(a) simple poles at $\omega = \beta_n$, $n = 0, \pm 1, \pm 2, \cdots$, and $\omega = -\beta_0$,

(b) simple zeroes at $\omega = q_j' = q_j + \delta_j$, $j = 0, 1, 2, \cdots$, which are found from the condition

$$F(q_j) + e^{2iq_j kd} F(-q_j) = 0, \quad (2.11)$$

(c) $F(\omega) = O(\omega^{-1/2})$ as $|\omega| \to \infty$.

Substituting $F(\omega)$ into Eq. (2.10), and performing the integration yields a residue series similar to Eq. (2.8) if the following identifications are made

$$\alpha_n A_n^h = R(\beta_n), \quad (2.12)$$

$$B_j^h = (i\Lambda/\pi q_j) e^{-iq_j kd} (1 - (-)^j e^{2\pi i \alpha_0/\Lambda}) F(q_j), \quad (2.13)$$

$$\alpha_0 = R(-\beta_0), \quad (2.14)$$

where $R(\beta)$ is the residue of $F(\omega)$ at $\omega = \beta$. Equation (2.14) is used in the construction of the function which is given by

$$F(\omega) = \frac{2\alpha_0\beta_0}{\omega^2 - \beta_0^2} \frac{\omega - q_0'}{\beta_0 + q_0'} \frac{\Pi(\omega, q')}{\Pi(-\beta_0, q')} \frac{\Pi_{12}(-\beta_0, \beta)}{\Pi_{12}(\omega, \beta)} \times e^{2i(\omega + \beta_0) \ln 2/\Lambda}, \quad (2.15)$$

where the infinite products are

$$\Pi(\omega, q') = \prod_{m=1}^{\infty} (1 - \omega/q_m')(2q_m'/im\Lambda) e^{2\omega/im\Lambda}, \quad (2.16)$$

$$\Pi_{\{1\atop 2\}}(\omega, \beta) = \prod_{n=1}^{\infty} (1 - \omega/\beta_{1n})(\beta_{1n}/in\Lambda) e^{\omega/in\Lambda}, \quad (2.17)$$

and $\Pi_{12} \equiv \Pi_1 \Pi_2$. These products are discussed in I. Using Eqs. (2.12) and (2.15) and techniques in I, it is seen that

$$A_n^h = \frac{\beta_0}{\beta_n} \frac{\beta_n - q_0'}{\beta_n + q_0'} \frac{\Pi(\beta_n, q')}{\Pi(-\beta_0, q')} \frac{\Pi_{12}(-\beta_n, \beta)}{\Pi_{12}(\beta_0, \beta)} \times e^{2i(\beta_n + \beta_0) \ln 2/\Lambda}. \quad (2.18)$$

The edge condition follows similarly to the edge condition in I if we note that for large n, $A_n^h = O(n^{-3/2})$. Also, the flux conservation relation can be derived and is similar to that in I, viz.

$$\sum_n R_n \equiv \sum_n |A_n^h|^2 \left(\frac{\beta_n}{\beta_0} \right) = 1, \quad (2.19)$$

where the sum runs over integer n such that β_n is real (real scattering orders), and R_n are the individual spectral reflection coefficients. R_n as well as the amplitude phases ϕ_n are plotted later. The latter are defined via $A_n^h = |A_n^h| e^{i\phi_n}$.

Lastly, it is necessary to write the iterative scheme used to find the δ_j, and thereby the zeroes of $F(\omega)$ shifted from the known q_j values. This follows by substituting $F(\omega)$ into Eq. (2.11), rearranging terms, and introducing an iteration index. A similar procedure was used in I. Only the final equation will be written here. It is

$$\frac{q_j - 1 - \delta_0^{(m+1)}}{q_j + 1 + \delta_0^{(m+1)}} \frac{\delta_j^{(m+1)} e^{4q_j/ij\Lambda}}{\delta_j^{(m+1)} + 2q_j} \prod_{n=1}^{j-1} \frac{\delta_n^{(m+1)} + q_n - q_j}{\delta_n^{(m+1)} + q_n + q_j} e^{4q_j/in\Lambda} = \text{rhs} \prod_{n=j+1}^{\infty} \frac{\delta_n^{(m)} + q_n + q_j}{\delta_n^{(m)} + q_n - q_j} e^{-4q_j/in\Lambda}, \quad (2.20)$$

where m is the iteration index, rhs is

$$\text{rhs} \equiv \frac{\Pi_{12}(q_j, \beta)}{\Pi_{12}(-q_j, \beta)} e^{2iq_j(kd - 2(\ln 2)/\Lambda)}, \quad (2.21)$$

which is the same rhs as in I, and $\delta_n^{(m)} \sim 0$ for n large. The iteration procedure is similar to the procedure in I, except here the calculation of δ_j starts with $j = 0$. Once the δ_j are known, the amplitudes are evaluated from Eqs. (2.12) and (2.13). Results are presented in the next section.

3. RESULTS AND DISCUSSION

It has already been remarked that the iterative procedure is similar to that in I; So is the evaluation of the reflection coefficients R_n and amplitude phases ϕ_n. The errors on both procedures are the same as in I although generally the iterative procedure converged slower here than in I. Equation (2.19) was again used as a check on the accuracy, and was satisfied to at least three decimal places.

Figures 2–10 are plots of R_n and ϕ_n vs kd, α_0 or θ_i, or Λ. Many of the parameter values were chosen to compare with those in I.

Figures 2–4 are plots of R_n and ϕ_n (from $-\pi$ to π) vs kd with $\Lambda = 0.63$ and $\alpha_0 = \sin\theta_i = 0.1$ (near normal incidence, Fig. 2), 0.707 (45° incidence, Fig. 3), and 0.99619 (near grazing incidence, Fig. 4). The values are considerably different from those in I, and the oscillatory behavior in I is not as evident here.

Figure (2a) illustrates that specular return can remain near maximum over a very broad range of kd values, whereas Fig. (3a) illustrates a resonance-type effect in R_0 at $kd \cong 4.2$ and a subsequent rapid interchange of energy between specular and backscatter. The latter is an illustration of the Wood S-anomaly. Figure (4a) shows more oscillatory behavior than the other figures. In all three figures the far backscatter reflection coefficients were generally larger than the other reflection coefficients with the exception of the specular one. Also noticed is the correlation between kd points where R_n go through minimum values and the kd points where $\partial^2\phi_n/\partial(kd)^2 = 0$. The magnitude of an R_n at its minimum also correlates to the value of $\partial\phi_n/\partial(kd)$ at the R_n-minimum point. These effects were also noted in I and will be more thoroughly discussed elsewhere. Finally, the kd values imposed no restriction on the evaluation of the model.

In addition to the periodicity of R_n and ϕ_n as a function of kd, Fig. (5a) illustrates the Brewster angle effect at $kd \approx 7.5$ and $kd \approx 23.5$. This is the vanishing of the specular reflection coefficient R_0 and, since, there is no energy lost, the full conversion of the energy into the backscatter direction, the only other available propagating grating order for the indicated parameters. The phase correlation discussed in the previous paragraph is also observed in Fig. (5b).

Figures 6–9 illustrate the variation in R_n and ϕ_n as the incident angle θ_i (or its sine, α_0) is changed. Again note in Figs. (6a) and (8a) that, except for the specular reflection coefficient, the far backscatter amplitude predominates. Also at $\alpha_0 \approx 0.9$, both figures exhibit Rayleigh anomalies. For $\alpha_0 > 0.9$, the energy interchange between R_0 and the far backscatter reflection coefficient is an example of the Wood S-anomaly. Figure (7a) illustrates the Wood S-anomaly with a partial Brewster angle effect. A near standing wave pattern is set up. The corresponding figure in I showed R_0 nearly constant over the full α_0 range. The phase behavior in Figs. (6b), (7b), and (8b) is much less predictable than in previous figures. Several phase jumps of π or approximately π are shown, but they do not appear to be correlated with significant reflection coefficient effects. Figures (9a) and (9b) are plots of the backscatter reflection coefficients R_{-2} and R_{-4}, respectively, as the angle of the incident plane wave approaches grazing. The maxima in the reflection coefficients occur when the incident and scattered propagation vectors are parallel and of course oppositely directed. A similar result occurred in I.

Finally, Fig. 10 illustrates the variation in R_n and ϕ_n as Λ changes. A Rayleigh anomaly occurs at $\Lambda = 0.8535$ and an S-anomaly around $\Lambda = 0.99$.

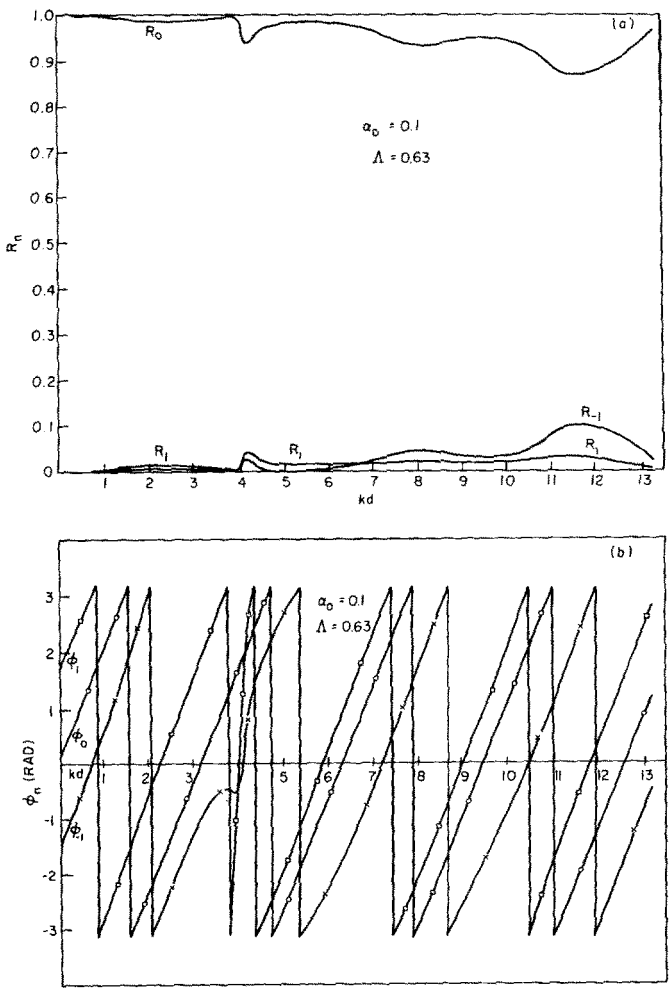


FIG. 2. Reflection coefficients R_n, (a), and phases ϕ_n, (b), plotted vs kd with $\Lambda = 0.63$ and $\alpha_0 = 0.1$ ($\theta_i \approx 5.8°$).

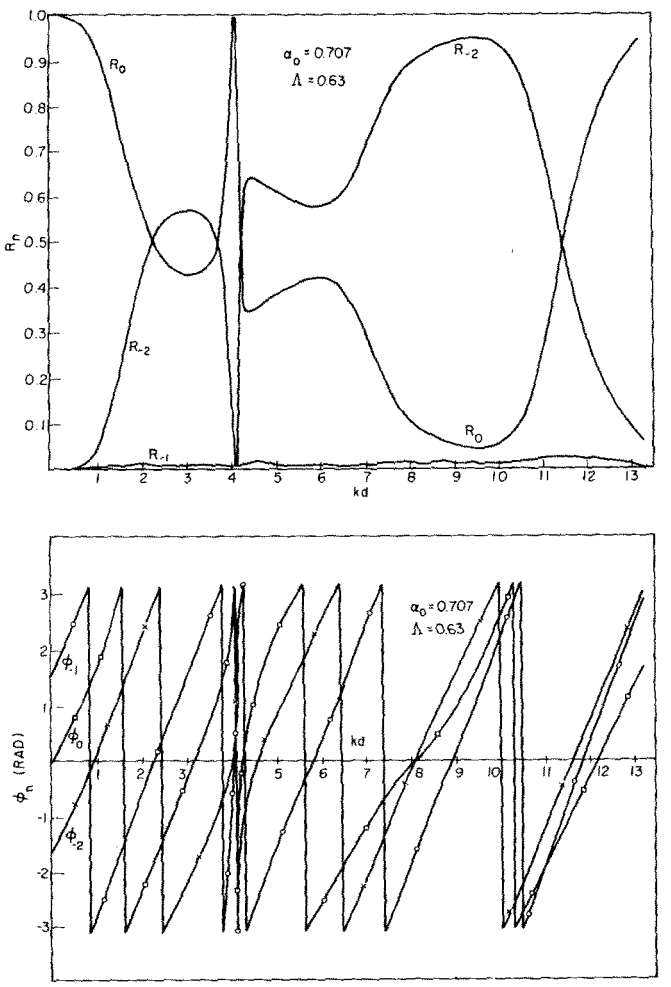


FIG. 3. Plots of R_n, (a), and ϕ_n, (b), vs kd with $\Lambda = 0.63$ and $\alpha_0 = 0.707$ ($\theta_i = 45°$). Resonance and S-anomaly effects are illustrated.

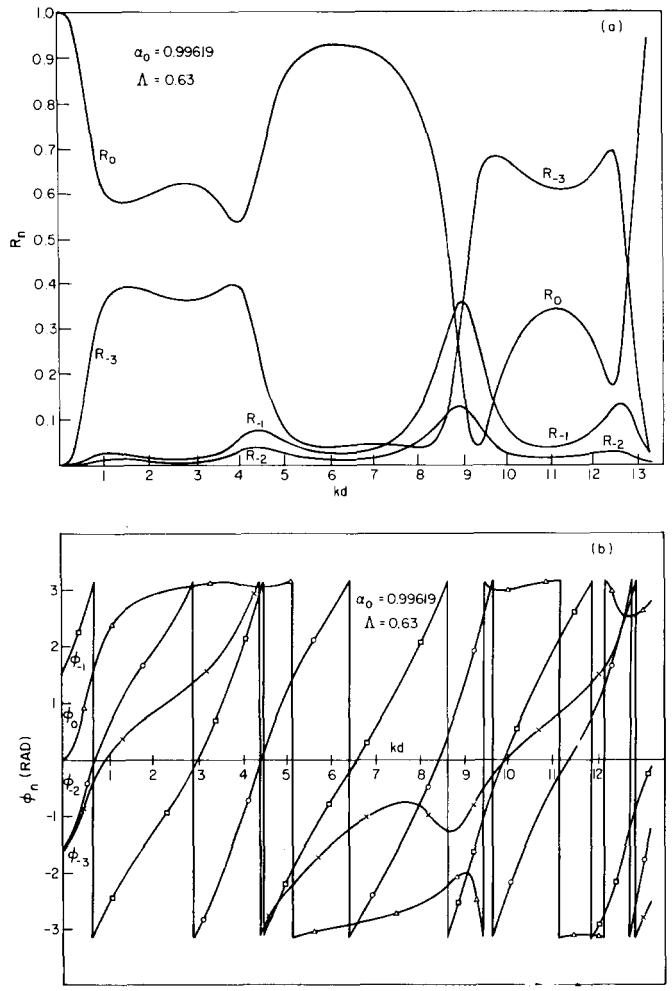


FIG. 4. Plots of R_n, (a), and ϕ_n, (b), vs kd with $\Lambda = 0.63$ and $\alpha_0 = 0.99619$ ($\theta_i = 85°$).

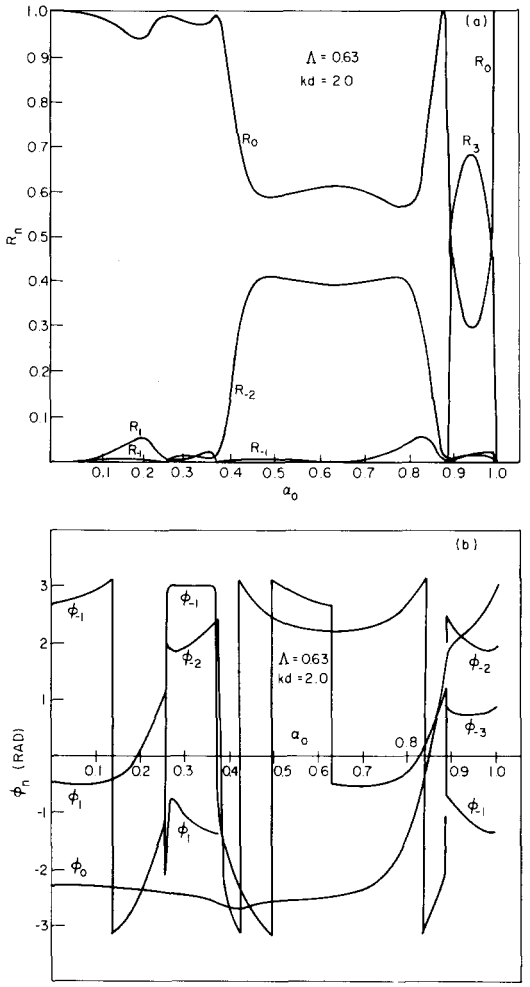


FIG. 6. Plots of R_n, (a), and ϕ_n, (b), vs α_0 for $\Lambda = 0.63$ and $kd = 2$. The Wood S-anomaly and Rayleigh anomaly are shown.

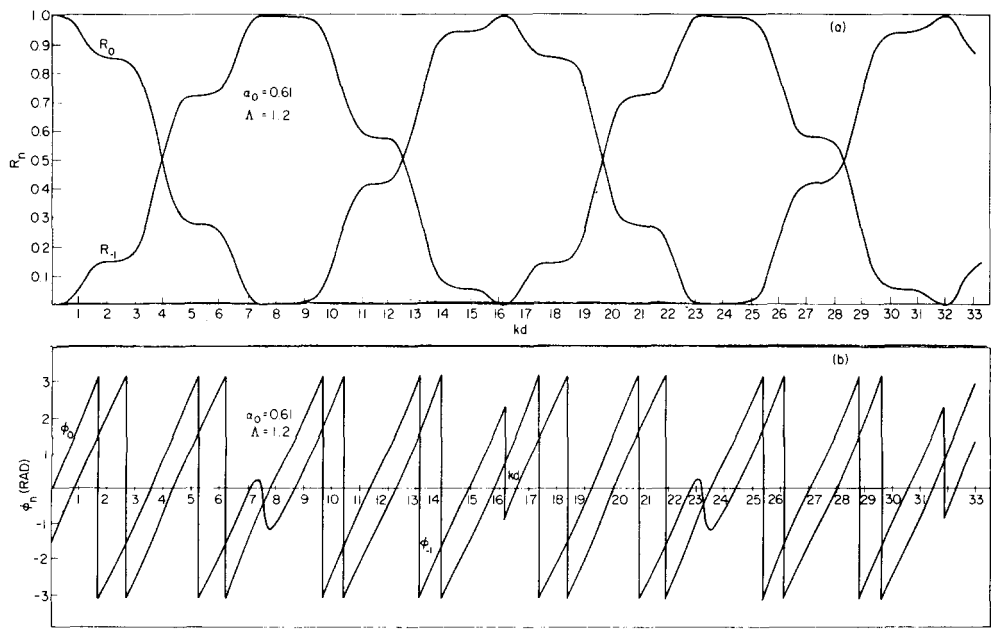


FIG. 5. Plots of R_n, (a), and ϕ_n, (b), vs kd for $\Lambda = 1.2$ and $\alpha_0 = 0.61$. The Brewster angle anomaly and kd-periodicity of R_n and ϕ_n are shown.

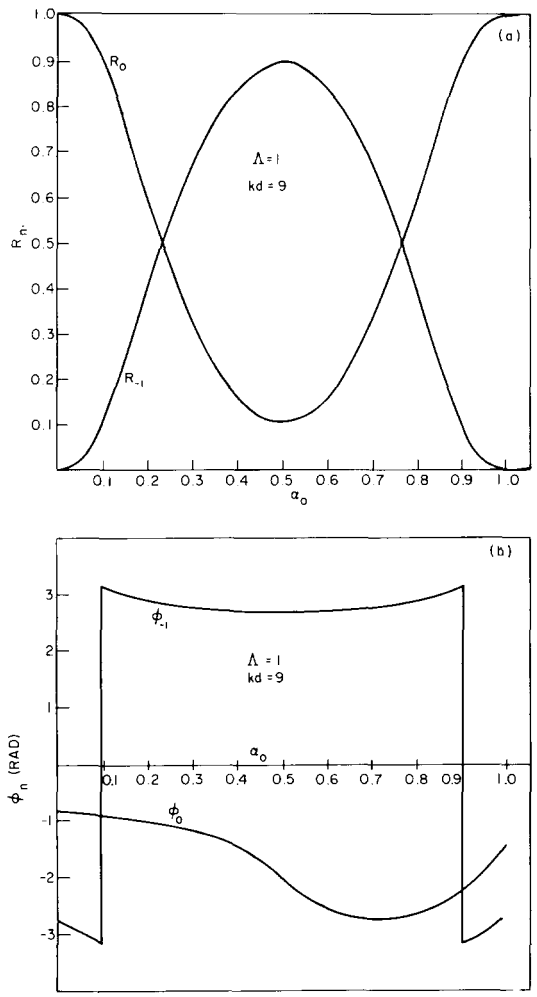


FIG. 7. Plots of R_n, (a), and ϕ_n, (b), vs α_0 for $\Lambda = 1$ and $kd = 9$. The Wood S-anomaly and a partial Brewster angle anomaly are shown.

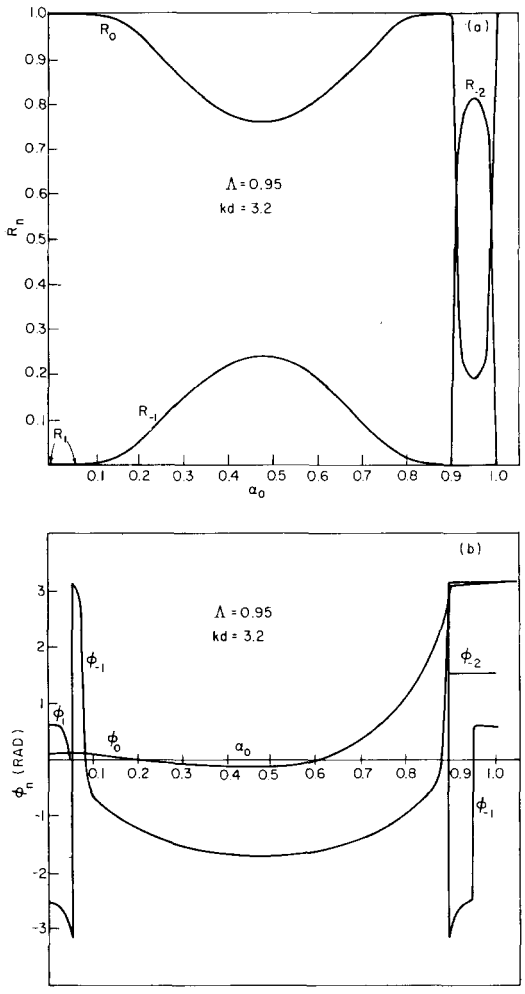


FIG. 8. Plots of R_n, (a), and ϕ_n, (b), vs α_0 for $\Lambda = 0.95$ and $kd = 3.2$. The Wood S-anomaly and the Rayleigh anomaly are shown.

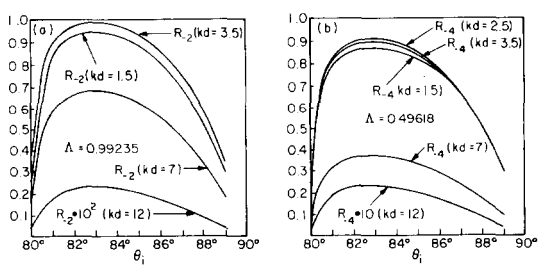


FIG. 9. Plots of backscatter reflection coefficients at near grazing incidence for (a) $\Lambda = 0.99235$ and (b) $\Lambda = 0.49618$.

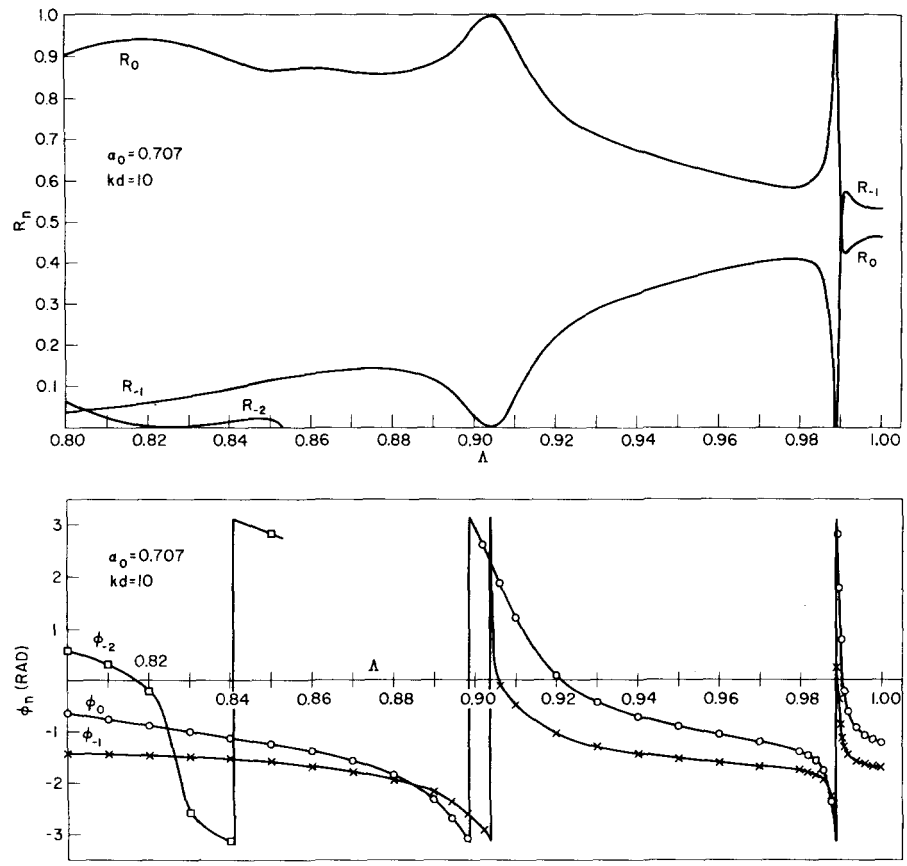


FIG. 10. Plots of R_n, (a), and ϕ_n, (b), vs Λ for $\alpha_0 = 0.707$ ($\theta_i = 45°$) and $kd = 10$. A Rayleigh anomaly and a Wood S-anomaly are shown.

4. SUMMARY AND CONCLUSIONS

It has been shown how to calculate the fields scattered and diffracted by a thin comb grating with Neumann boundary conditions and plane wave incidence. The solution requires an iterative procedure and complex function theory. Reflection coefficients R_n and amplitude phases ϕ_n were evaluated as a function of incident angle, frequency, and corrugation depth. The various grating anomalies, the Brewster angle anomaly, Wood S-anomaly, and Rayleigh anomaly were illustrated as well as correlations between R_n minima and behavior of derivatives of ϕ_n with respect to the depth. The periodicity of the scattered field as a function of corrugation depth was also illustrated. Finally, the procedure was highly accurate, and different parameter values imposed no restrictions on the evaluation of the model.

ACKNOWLEDGMENTS

I am grateful to Ned Wright and George Frisk for the numerical work and to Frances Vandivier for her careful typing.

[1] J. A. DeSanto, J. Math. Phys. **12**, 1913 (1971). This paper will be referred to as I.
[2] R. A. Hurd, Can. J. Phys. **32**, 727 (1954).
[3] R. Mittra, S. W. Lee, and G. F. Vanblaricum, Intern J. Eng. Sci. **6**, 395 (1968) and R. Mittra and S. W. Lee, *Analytical Techniques in the Theory of Guided Waves* (Macmillan, New York, 1971).
[4] J. E. Stewart and W. S. Gallaway, Appl. Opt. **1**, 421 (1962).
[5] A. Hessel and A. A. Oliner, Appl. Opt. **4**, 1275 (1965).
[6] D. Y. Tseng, Ph. D. Thesis (Polytechnic Institute of Brooklyn, 1967).

On Projective Unitary-Antiunitary Representations of Finite Groups

T. Janssen
Institute for Theoretical Physics, Katholieke Universiteit, Nijmegen, The Netherlands
(Received 25 May 1970; Revised Manuscript Received 10 October 1971)

A generalization of Schur's treatment of projective representations is discussed for a very general class of representations occurring in physical theories: the projective representations by unitary and antiunitary operators. It is shown that for every finite group the projective unitary-antiunitary (PUA) representations can be obtained from the ordinary unitary-antiunitary representations of another finite group. The construction of such a representation group is treated. As an example we apply the theory for the determination of irreducible representations of subgroups of the Poincaré group. The classes of PUA representations for all finite crystallographic groups in spaces of dimension up to four are explicitly given.

1. INTRODUCTION

As is well known, not only representations of groups are in use in physics, but also "representations up to a factor," equally often called "projective representations" or "ray representations" (see, for example, Hamermesh[1]). For example, projective representations play a role in the theory of representations of nonsymmorphic space groups and are quite generally of importance for many quantum-mechanical systems because quantum-mechanical states are described by rays (rather than by vectors) of a Hilbert space. Such systems are systems of particles with half-integral spins (in connection with double groups), and those of charged particles in an electromagnetic field (in connection with gauge transformations[2]).

It is also well known that the time-reversal transformation leads to antiunitary operators in quantum mechanics. This gives rise to "corepresentations," which in the present paper will often be called "unitary-antiunitary representations". They occur in symmetry considerations of systems with time-reversal symmetry, for systems having a magnetic group as symmetry, and in the determination of corepresentations of space-time groups. The theory of projective representations of finite groups stems from Schur,[3] that of continuous groups was given by Bargmann.[4] Corepresentations were introduced by Wigner.[5] However, a still more general kind of representations occurring in physics are projective representations by unitary and antiunitary operators, that is, projective corepresentations. A general discussion of these "PUA representations" was given by Parthasarathy[6] for locally compact groups. In the present paper we want to discuss the general theory of PUA representations for finite groups in more detail. This case has already been considered by Murthy,[7] who has not, however, formulated a general method for obtaining the so-called "factor systems" for the PUA representations, so that his treatment of the problem is incomplete. Our treatment is a generalization of that given by Schur.[3] We will give a general constructive method to determine all projective representations and corepresentations of a finite group. In particular, for the crystallographic magnetic groups the classes of PUA representations are explicitly given. Finally, we comment on two applications of the general theory.

We start with a discussion of the role of PUA representations in physics. The states of a quantum-mechanical system are in one-to-one correspondence with one-dimensional subspaces of a Hilbert space $\mathfrak{K}$, called rays. The rays are elements of the projective Hilbert space $\mathfrak{K}$. Notice that this is not a linear vector space. For any nonzero element $\Psi \in \mathfrak{K}$ the ray $\mathbf{\Psi}$ is defined by

$$\mathbf{\Psi} = \{\lambda\Psi \mid \text{all complex numbers } \lambda\}. \tag{1.1}$$

The set of all rays $\mathbf{\Psi}$ with $\Psi \in \mathfrak{K}$ is the projective Hilbert space $\mathfrak{K}$. In $\mathfrak{K}$ one defines a kind of inner product by

$$(\mathbf{\Psi}_1, \mathbf{\Psi}_2) = \frac{|\langle\Psi_1|\Psi_2\rangle|}{(\langle\Psi_1|\Psi_1\rangle\langle\Psi_2|\Psi_2\rangle)^{1/2}}, \qquad \Psi_1, \Psi_2 \in \mathfrak{K}. \tag{1.2}$$

This definition does not depend on the choice of the representative Ψ of $\mathbf{\Psi}$. An automorphism of $\mathfrak{K}$ is an invertible mapping of $\mathfrak{K}$ onto itself which preserves this "inner product." For any unitary or antiunitary operator A on $\mathfrak{K}$ an automorphism $\mathbf{A}$ of $\mathfrak{K}$ is defined by $\mathbf{A}\mathbf{\Psi} = (A\Psi)$. On the other hand Wigner has shown[5] that any automorphism $\mathbf{A}$ of $\mathfrak{K}$ can be obtained in this way from a unitary or antiunitary operator. We denote by $\mathfrak{U}(\mathfrak{K})$ the group of unitary operators on $\mathfrak{K}$, by $\mathfrak{A}(\mathfrak{K})$ the group of unitary and antiunitary operators on $\mathfrak{K}$, by $\mathbf{\mathfrak{A}}(\mathfrak{K})$ the group of automorphisms of $\mathfrak{K}$, by $\mathbf{\mathfrak{U}}(\mathfrak{K})$ those obtained from a unitary operator, and by π the epimorphism $\mathfrak{A}(\mathfrak{K}) \to \mathbf{\mathfrak{A}}(\mathfrak{K})$ defined by $(\pi A)\mathbf{\Psi} = (\mathbf{A}\mathbf{\Psi})$. (Since the theory of projective representations is closely related to that of group extensions, we will use a fair amount of terminology from group extension and cohomology theory. We refer the reader not familiar with these notions to Hall[8] or McLane.[9] Some notions are defined in Appendix A, where we discuss also some concepts from the theory of finite groups.) If we only consider normalized states we can restrict the values of λ in (1.1) to $U(1)$, the group of unimodular complex numbers, and leave out the denominator in (1.2). That this does not have consequences is shown in Appendix B. In the following we only consider normalized states. Then we have the commutative diagram

$$\begin{array}{ccccccccc} 1 & \longrightarrow & U(1) & \longrightarrow & \mathfrak{U}(\mathfrak{K}) & \xrightarrow{\pi} & \mathbf{\mathfrak{U}}(\mathfrak{K}) & \longrightarrow & 1, \\ & & \| & & \downarrow & & \downarrow & & \\ 1 & \longrightarrow & U(1) & \longrightarrow & \mathfrak{A}(\mathfrak{K}) & \xrightarrow{\pi} & \mathbf{\mathfrak{A}}(\mathfrak{K}) & \longrightarrow & 1. \end{array} \tag{1.3}$$

This means that any element $\mathbf{A} \in \mathbf{\mathfrak{U}}(\mathfrak{K})$ [respectively $\in \mathbf{\mathfrak{A}}(\mathfrak{K})$] is the image under π of some $A \in \mathfrak{U}(\mathfrak{K})$ [respectively $\in \mathfrak{A}(\mathfrak{K})$]. If two operators A and A' satisfy $\pi A = \pi A'$, one has $A' = e^{i\varphi}A$ for some $e^{i\varphi} \in U(1)$.

We define a projective unitary-antiunitary representation (PUA rep) of the couple (G, H), where H is a subgroup of index 1 or 2 in the group G, as a mapping $P: G \to \mathfrak{A}(\mathfrak{K})$ such that (i) $\mathbf{P} = \pi P$ is a homomorphism of G in $\mathbf{\mathfrak{A}}(\mathfrak{K})$, and (ii) $P(H) \subset \mathfrak{U}(\mathfrak{K})$, $P(G\text{-}H) \not\subset \mathfrak{U}(\mathfrak{K})$. In an analogous way one can define a projective linear-antilinear representation of (G, H) as a mapping of G on a set of linear and antilinear operators, which mapping is a homomorphism up to a nonzero constant and which maps H on linear operators, $G\text{-}H$ on antilinear ones. An ordinary unitary-antiuni-

tary representation (UA rep) of (G,H) is a PUA rep of (G,H) which is a homomorphism. A projective representation of G is a PUA rep of (G,G). Sometimes $\mathbf{P} = \pi P$ is also called a projective representation (in $\mathfrak{K}$).

When P is a PUA rep of (G,H) in $\mathfrak{K}$, R a group, $\sigma: R \to G$ an epimorphism, a UA rep D of R in $\mathfrak{K}$ is called a lifting of P if $D = P\sigma$. Schur has shown that for a finite group G there exists a group R such that any projective unitary representation P of G can be lifted to an ordinary unitary representation of R. We will show that the same holds for PUA reps of finite groups. We will give a construction of such a group R. The kernel of σ is called the comultiplicator $M(G,H)$. For PU reps it is just Schur's multiplicator. One has the commutative diagram (whose rows are exact sequences)

$$\begin{array}{ccccccccc} 1 & \longrightarrow & M(G,H) & \longrightarrow & R & \xrightarrow{\sigma} & G & \longrightarrow & 1, \\ & & \downarrow \chi & & D \downarrow & \swarrow P & P \downarrow & & \\ 1 & \longrightarrow & U(1) & \longrightarrow & \mathfrak{A}(\mathfrak{K}) & \xrightarrow{\pi} & \mathfrak{A}(\mathfrak{K}) & \longrightarrow & \end{array} \tag{1.4}$$

Any irreducible UA rep D of R will give a PUA rep P of G by $P = D\gamma$, if γ is a section $\gamma: G \to R$.

The importance of PUA reps for physics will be clear. As Wigner has shown a symmetry group G of space–time transformations for a physical system induces a group of automorphisms of $\mathfrak{K}$ which forms a PUA rep of G. More generally, the symmetry group of an operator A on $\mathfrak{K}$ is the group of all unitary and antiunitary operators on $\mathfrak{K}$ commuting with A. This symmetry group is of infinite order because it contains the group of all scalar unimodular operators (phase factors). The projection of this symmetry group on $\mathfrak{A}(\mathfrak{K})$ is a group G of automorphisms of $\mathfrak{K}$. Choosing a section $s: \mathfrak{A}(\mathfrak{K}) \to \mathfrak{A}(\mathfrak{K})$, one obtains a PUA rep of G. This group G may very well be finite.

2. PUA REPRESENTATIONS

Consider in $\mathfrak{K}$ a PUA rep P of a couple (G,H) or, as we also say, of a group G with respect to a group H. Since πP is a homomorphism, one has for any $\alpha, \beta \in G$ an element $\omega(\alpha,\beta) \in U(1)$ such that

$$P(\alpha)P(\beta) = \omega(\alpha,\beta)P(\alpha\beta). \tag{2.1}$$

The mapping $\omega: G \times G \to U(1)$ is called a factor system. Because of the associativity of the product of operators one has for any $\alpha, \beta, \gamma \in G$

$$\begin{aligned} \omega(\alpha,\beta)\omega(\alpha\beta,\gamma) &= \omega(\alpha,\beta\gamma)\omega(\beta,\gamma) \quad \text{if } \alpha \in H \\ &= \omega(\alpha,\beta\gamma)\omega^*(\beta,\gamma) \quad \text{if } \alpha \in G\text{-}H. \end{aligned} \tag{2.2}$$

Now we define a mapping ϕ of G on the group of automorphisms of $U(1)$ by

$$\begin{aligned} \phi(\alpha)\lambda \equiv \lambda^\alpha &= \lambda \quad \text{if } \alpha \in H \\ &= \lambda^{-1} \quad \text{if } \alpha \in G\text{-}H \end{aligned} \tag{2.3}$$

For any $\lambda \in U(1)$. The mapping ϕ is a homomorphism. Using Eq. (2.3) one can write Eq. (2.2) as

$$\omega(\alpha,\beta)\omega(\alpha\beta,\gamma) = \omega(\alpha,\beta\gamma)\omega^\alpha(\beta,\gamma), \qquad \alpha,\beta,\gamma \in G \tag{2.4}$$

which expression means that ω is a 2-cocycle (cf. Appendix A for definitions of some cohomological concepts). We denote the group of n-cocycles with arguments in G, values in $U(1)$, and the action (2.3) of G on $U(1)$ by $Z^n_H(G)$. From the action of G follows the action of the integral group ring ZG on $U(1)$ by

$$\lambda^{\alpha+\beta} = \lambda^\alpha\lambda^\beta, \quad \lambda^{m\alpha} = (\lambda^\alpha)^m, \quad \text{any } \lambda \in U(1), \text{integer } m.$$

When $H = G$ the action of G on $U(1)$ is the trivial one. In that case we have a PU rep.

The PUA rep P gives a homomorphism $\mathbf{P} = \pi P$. The same $\mathbf{P}$ is obtained from a PUA rep P' if and only if there is a 1-cochain $c: G \to U(1)$ such that

$$P'(\alpha) = c(\alpha)P(\alpha), \qquad \forall\alpha \in G. \tag{2.5}$$

The PUA rep P' determines a factor system ω' related to ω by

$$\omega'(\alpha,\beta) = c(\alpha)c^\alpha(\beta)c^{-1}(\alpha\beta)\omega(\alpha,\beta), \ \forall\alpha,\beta \in G. \tag{2.6}$$

This means that $\omega'\omega^{-1}$ is a 2-coboundary: $\omega'\omega^{-1} = \delta c$. The group of n-coboundaries is denoted by $B^n_H(G)$. The reps P and P' as well as their factor systems ω and ω' are called associated. Any homomorphism $\mathbf{P}: G \to \mathfrak{A}(\mathfrak{K})$ determines a class of associated PUA reps. If $\omega \in B^2_H(G)$, there is an associated factor system which is identically 1, i.e., the PUA rep P is associated with a UA rep. Then there is a section $s: \mathfrak{A}(\mathfrak{K}) \to \mathfrak{A}(\mathfrak{K})$ such that the PUA rep $s\mathbf{P}$ is a homomorphism.

The classes of associated factor systems are elements of the quotient group

$$H^2_H(G) = Z^2_H(G)/B^2_H(G). \tag{2.7}$$

$H^2_H(G)$ is isomorphic to the comultiplicator $M(G,H)$ of G with respect to H which is defined as a representative of the isomorphism class of $H^2_H(G)$. If $H = G$ it is just Schur's multiplicator.

Theorem 1: In each equivalence class of factor systems of a finite group of order N there is a factor system consisting of Nth roots of unity.

Proof: This follows from a theorem in cohomology theory which states that any element of $H^n_\phi(G,A)$ for a finite group G, a G-module A and $n > 0$ has order dividing N. This implies that ω^N is associated with the trivial factor system:

$$\omega(\alpha,\beta)^N = \frac{c(\alpha)c^\alpha(\beta)}{c(\alpha\beta)}$$

for some $c \in C^1(G)$. Then choosing $c'(\alpha) = c(\alpha)^{-1/N}$ one defines $\omega' = \omega\delta c$ which is associated with ω and has the property $\omega'^N(\alpha,\beta) = 1$ $(\forall\alpha,\beta \in G)$.

The notions of reducibility and equivalence are analogous to those for ordinary representations. The PUA rep $P: G \to \mathfrak{A}(\mathfrak{K})$ is irreducible if $\mathfrak{K}$ does not have a proper G-invariant subspace. The PUA reps P_1 and P_2 are equivalent if there is a unitary operator $U \in \mathfrak{U}(\mathfrak{K})$ such that $U^{-1}P_1(\alpha)U = P_2(\alpha)$ $(\forall\alpha \in G)$. The PUA reps P_1 and P_2 are similar if there are a unitary operator U and a 1-cochain $c \in C^1(G)$ such that $c(\alpha)U^{-1}P_1(\alpha)U = P_2(\alpha)$. Equivalent representations have the same factor system. The factor systems of

similar PUA reps are associated. To find all nonsimilar PUA reps of a finite group G with respect to H one calculates first $H^2_H(G)$ and determines for one factor system from each equivalence class the nonequivalent PUA reps with these factor systems. The first step is discussed in Sec. 3, the second one in Sec. 4.

For two linear operators A_1 and A_2, or two antilinear operators B_1 and B_2 on vector spaces V_1 and V_2, respectively, one defines the Kronecker products $A_1 \otimes A_2$ and $B_1 \otimes B_2$ on $V_1 \otimes V_2$ by $(A_1 \otimes A_2)(x \otimes y) = (A_1 x) \otimes (A_2 y)$ and $(B_1 \otimes B_2)(x \otimes y) = (B_1 x) \otimes (B_2 y)$. Then $A_1 \otimes A_2$ is a linear operator and $B_1 \otimes B_2$ an antilinear one. This implies that if P_1 and P_2 are PUA reps of (G, H), the operators $(P_1 \otimes P_2)(\alpha) = P_1(\alpha) \otimes P_2(\alpha)$ form also a PUA rep of (G, H). If ω_i is the factor system for $P_i (i = 1, 2)$, the factor system of $P_1 \otimes P_2$ is $\omega_1 \omega_2$. Consider a one-dimensional space $\mathfrak{K}$, an antilinear operator θ on $\mathfrak{K}$ with $\theta^2 = 1$, and a 1-cocycle $c \in Z^1_H(G)$. Then $P_c(\alpha) = c(\alpha)$ for $\alpha \in H$, $P_c(\alpha) = c(\alpha)\theta$ for $\alpha \in G\text{-}H$ gives a one-dimensional PUA rep P_c of (G, H). Then we see that two PUA reps P_1 and P_2 with the same factor system are similar if and only if P_1 is equivalent with the product PUA rep $P_c \otimes P_2$ for some 1-cocycle c.

3. DETERMINATION OF THE COMULTIPLICATOR

We will determine all nonassociated factor systems for PUA reps of a finite group G with respect to H. We use the close relationship between projective representations and group extensions. When A is an Abelian group, B an arbitrary finite group, and ϕ a homomorphism of B into the group of automorphisms of A, any extension of A by B determines a factor system $m \in Z^2_\phi(B, A)$ and conversely each 2-cocycle m determines a group extension with product $(a, \alpha)(b, \beta) = (a \cdot \phi(\alpha)b \cdot m(\alpha, \beta), \alpha\beta)$. Now we consider an extension of $U(1)$ by G where the action of G on $U(1)$ is given by Eq. (2.3):

$$1 \longrightarrow U(1) \longrightarrow K \xrightarrow{\sigma} G \longrightarrow 1. \tag{3.1}$$

Any extension (3.1) gives a factor system. On the other hand, equivalent extensions determine factor systems which differ by a 2-coboundary. Therefore, there is a one-to-one correspondence between the nonassociated factor systems of PUA reps of (G, H) and the nonequivalent extensions (3.1).

Now suppose that G is generated by ν generators $\alpha_1, \ldots, \alpha_\nu$. Furthermore, G is defined by r defining relations $\Phi_i(\alpha_1, \ldots, \alpha_\nu) = \epsilon$ $(i = 1, \ldots, r)$. Take a section $\tau: G \to K$ and replace in the relations $\Phi_i = \epsilon$ everywhere α by $\tau(\alpha)$. Because $\sigma\tau(\alpha) = \alpha \in G$ one has $\sigma\Phi_i(\tau(\alpha_1), \ldots, \tau(\alpha_\nu)) = \sigma\tau(\epsilon) = \epsilon$. This implies that $\Phi_i(\tau(\alpha_1), \ldots, \tau(\alpha_\nu))$ belongs to the kernel of σ, i.e., to $U(1)$:

$$\Phi_i(\tau(\alpha_1), \ldots, \tau(\alpha_\nu)) = g_i \in U(1), \qquad i = 1, \ldots, r. \tag{3.2}$$

For another section τ' there is a 1-cochain $u \in C^1(G)$ such that $\tau'(\alpha) = u(\alpha)\tau(\alpha)$. Using the relations $\tau(\alpha)c\tau(\alpha)^{-1} = c^\alpha (c \in U(1), \alpha \in G)$, one has elements $\pi_i(\alpha_j)$ from the integral group ring ZG such that

$$\Phi_i(\tau'(\alpha_1), \ldots, \tau'(\alpha_\nu)) = \Phi_i(u(\alpha_1)\tau(\alpha_1), \ldots, u(\alpha_\nu)\tau(\alpha_\nu))$$
$$= g_i \prod_{j=1}^{\nu} u(\alpha_j)^{\pi_i(\alpha_j)} = g_i'. \tag{3.3}$$

According to a theorem by Hall[8] elements $g_1, \ldots, g_r \in U(1)$ determine an extension if and only if

$$\prod_{i=1}^{r} g_i^{h_i} = 1 \tag{3.4}$$

whenever elements $h_1, \ldots, h_r \in ZG$ are solutions of the equations

$$\sum_{i=1}^{r} h_i \pi_i(\alpha_j) = 0, \qquad j = 1, \ldots, \nu. \tag{3.5}$$

Moreover, equivalent extensions are determined by sets $\{g_i\}_{i=1,\ldots,r}$ and $\{g_i'\}_{i=1,\ldots,r}$ if and only if there are elements $u(\alpha_1), \ldots, u(\alpha_\nu) \in U(1)$ satisfying Eq. (3.3).

The solutions of (3.5) form a left ZG module. We suppose that this module is generated by the sets $\{h_i^1\}, \ldots, \{h_i^k\}$. The sets $\Phi = \{g_i\}$ form an Abelian group by

$$\Phi\Phi' = \{g_i\}\{g_i'\} = \{g_i'' = g_i g_i'\}. \tag{3.6}$$

A subgroup is formed by those solutions Φ which are equivalent to $\{g_i = 1\} = \Phi_0$. Because of the action of G on $U(1)$ one has with a mapping $\epsilon: ZG \to Z$ defined by

$$\epsilon\left(\sum_{\alpha \in G} m_\alpha \alpha\right) = \sum_{\alpha \in H} m_\alpha - \sum_{\beta \in G-H} m_\beta \tag{3.7}$$

the relation

$$g^{(\Sigma m_\alpha \alpha)} = g^{\epsilon(\Sigma m_\alpha \alpha)}.$$

The mapping of the group ring elements h_i^j and $\pi_i(\alpha_j)$ gives the integers

$$m_{ij} = \epsilon(\pi_i(\alpha_j)), \qquad i = 1, \ldots, r; \quad j = 1, \ldots, \nu,$$
$$k_{ij} = \epsilon(h_i^j), \qquad i = 1, \ldots, r; \quad j = 1, \ldots, k.$$

Then Hall's theorem can be stated in the following form.

Theorem 2: The group $H^2_H(G)$ is isomorphic to the quotient group

$$\frac{\{\Phi \mid \prod_{i=1}^{r} g_i^{k_{ij}} = 1,\ j = 1, \ldots, k\}}{\{\Phi \mid g_i = \prod_{j=1}^{\nu} c_j^{m_{ij}},\ i = 1, \ldots, r, \text{ some } c_1, \ldots, c_\nu \in U(1)\}}. \tag{3.8}$$

The elements $\pi_i(\alpha_j)$ are found from the defining relations, the elements h_i^j from Eqs. (3.5). They make it possible to determine arbitrary extensions for any ZG module A. For all four-dimensional crystallographic magnetic point groups (i.e., also for n-dimensional crystallographic point groups for $n < 4$) they are given in Ref. 10. From the tables given there the integers m_{ij} and k_{ij} are immediately found using Eq. (3.7). Then $H^2_H(G)$ follows from Theorem 2.

As an example we determine the comultiplicator of the dihedral group D_2 with respect to a subgroup C_2. The group D_2 is generated by α_1 and α_2 with defining relations $\alpha_1^2 = \alpha_2^2 = (\alpha_1\alpha_2)^2 = \epsilon$. For the subgroup C_2 we take the group generated by α_1. The integers m_{ij} and k_{ij} are $m_{11} = k_{22} = k_{33} = k_{24} = -k_{34} = 2$ and otherwise zero. This means that $g_2^2 = g_3^2 =$

TABLE I. Comultiplicators of crystallographic point groups and Shubnikov point groups.[a]

G	H	$G = H + aH$	$M(G,H) = C_2^k$	Defining relations									
		a	k	α^n g_1	β^l g_2	$(\alpha\beta)^m$ g_3	γ^2 g_4	$\alpha\gamma\alpha^{-1}\gamma$ g_5	$\beta\gamma\beta^{-1}\gamma$ g_6	δ^2 g_7	$\alpha\delta\alpha^{-1}$ g_8	$\beta\delta\beta^{-1}\delta$ g_9	$\gamma\delta\gamma^{-1}\delta$ g_{10}
C_{2n+1}	C_{2n+1}	ϵ	0	u^{2n+1}									
C_{2n}	C_{2n}	ϵ	0	u^{2n}									
C_{2n}	C_n	α	1	± 1									
D_{2n+1}	D_{2n+1}	ϵ	0	u^{2n+1}	v^2	u^2v^2							
D_{2n}	D_{2n}	ϵ	1	$\pm u^{2n}$	v^2	u^2v^2							
D_3	C_3	β	1	u^3	$\pm 1 = \pm 1$								
D_{2n}	C_{2n}	β	2	u^{2n}	± 1	± 1							
D_{2n}	D_n	α	2	± 1	v^2	± 1							
T	T	ϵ	1	u^3	$\pm v^2$	u^3v^3							
O	O	ϵ	1	$\pm u^4$	v^3	u^2v^2							
O	T	α	2	± 1	v^3	± 1							
$C_{2n} \times C_2$	$C_{2n} \times C_2$	ϵ	1	u^{2n}			v^2	$\pm v^2$					
$C_{2n} \times C_2$	C_{2n}	γ	2	$\pm u^{2n}$			± 1	u^2					
$C_{2n} \times C_2$	$C_n \times C_2$	α	2	± 1			v^2	± 1					
$D_{2n} \times C_2$	$D_{2n} \times C_2$	ϵ	3	$\pm u^{2n}$	v^2	u^2v^2	w^2	$\pm w^2$	$\pm w^2$				
$D_{2n} \times C_2$	D_{2n}	γ	4	$\pm u^{2n}$	$\pm v^2$	$\pm u^2v^2$	± 1	u^2	v^2				
$D_4 \times C_2$	$D_2 \times C_2$	α	4	± 1	v^2	± 1	w^2	± 1	$\pm w^2$				
$D_{2n} \times C_2$	$C_{2n} \times C_2$	β	4	u^{2n}	± 1	± 1	w^2	$\pm w^2$	± 1				
$T \times C_2$	$T \times C_2$	ϵ	1	u^3	$\pm v^2$	u^3v^3	w^2	w^2	w^2				
$T \times C_2$	T	γ	3	$(-1)^\delta u^3$	$\pm(-1)^\epsilon v^2$	$(-1)^{\epsilon+\delta}u^3v^3$	$(-1)^\epsilon$	$(-1)^\epsilon u^2$	$(-1)^\epsilon v^2$				
$O \times C_2$	$O \times C_2$	ϵ	2	$\pm u^4$	v^3	u^2v^2	w^2	$\pm w^2$	w^2				
$O \times C_2$	O	γ	3	$\pm u^4$	$\pm v^3$	u^2v^2	$(-1)^\delta$	u^2	$(-1)^\delta v^2$				
$O \times C_2$	$T \times C_2$	α	3	± 1	v^3	± 1	w^2	± 1	w^2				
$C_{2n} \times C_2 \times C_2$	$C_{2n} \times C_2 \times C_2$	ϵ	3	u^{2n}			w^2	$\pm w^2$		z^2	$\pm z^2$		$\pm z^2$
$C_{2n} \times C_2 \times C_2$	$C_{2n} \times C_2$	δ	4	$\pm u^{2n}$			w^2	$\pm w^2$		± 1	u^2		$\pm w^2$
$C_4 \times C_2 \times C_2$	$C_2 \times C_2 \times C_2$	α	4	± 1			w^2	± 1		z^2	± 1		$\pm z^2$
$D_{2n} \times C_2 \times C_2$	$D_{2n} \times C_2 \times C_2$	ϵ	6	$\pm u^{2n}$	v^2	u^2v^2	w^2	$\pm w^2$	$\pm w^2$	z^2	$\pm z^2$	$\pm z^2$	$\pm z^2$
$D_{2n} \times C_2 \times C_2$	$D_{2n} \times C_2$	δ	7	$\pm u^{2n}$	$\pm v^2$	$\pm u^2v^2$	w^2	$\pm w^2$	$\pm w^2$	± 1	u^2	v^2	$\pm w^2$
$D_{2n} \times C_2 \times C_2$	$C_{2n} \times C_2 \times C_2$	β	7	u^{2n}	± 1	± 1	w^2	$\pm w^2$	± 1	z^2	$\pm z^2$	± 1	$\pm z^2$
$D_4 \times C_2 \times C_2$	$D_2 \times C_2 \times C_2$	α	7	± 1	v^2	± 1	w^2	± 1	$\pm w^2$	z^2	± 1	$\pm z^2$	$\pm z^2$
$T \times C_2 \times C_2$	$T \times C_2 \times C_2$	ϵ	2	u^3	$\pm v^2$	u^3v^3	w^2	w^2	w^2	z^2	z^2	z^2	$\pm z^2$
$O \times C_2 \times C_2$	$T \times C_2$	δ	4	$(-1)^\delta u^3$	$\pm(-1)^\epsilon v^2$	$(-1)^{\epsilon+\delta}u^3v^3$	w^2	w^2	w^2	$(-1)^\epsilon$	$(-1)^\epsilon u^2$	$(-1)^\epsilon v^2$	$\pm w^2$
$O \times C_2 \times C_2$	$O \times C_2 \times C_2$	ϵ	4	$\pm u^4$	v^3	u^2v^2	w^2	$\pm w^2$	w^2	z^2	$\pm z^2$	z^2	$\pm z^2$
$O \times C_2 \times C_2$	$O \times C_2$	δ	5	$\pm u^4$	$\pm v^3$	u^2v^2	w^2	$\pm w^2$	w^2	$(-1)^\delta$	u^2	$(-1)^\delta v^2$	$\pm w^2$
$O \times C_2 \times C_2$	$T \times C_2 \times C_2$	α	6	± 1	v^3	± 1	w^2	± 1	w^2	z^2	± 1	$\pm z^2$	$\pm z^2$

[a] $u, v, w, z \in U(1)$, representatives Φ from all classes are found by putting $u = v = w = z = 1$, the triple n, l, m is $n22$ for D_n, 323 for T, 432 for O, $\delta, \epsilon = 0$ or 1. Example: There are $2^2 = 4$ classes of nonassociated PUA reps of $T \times C_2$ with respect to T. They are given by $P(\alpha)^3 = [P(\alpha)P(\beta)]^3 = P(\beta)P(\gamma)P(\beta)^{-1}P(\gamma) = P(\alpha)P(\gamma)P(\alpha)^{-1}P(\gamma) = 1$ and $P(\beta)^2 = \pm 1$. The class given by $P(\alpha)^3 = u^3 1$, $P(\beta)^2 = v^2 1$, $[P(\alpha)P(\beta)]^3 = u^3v^3 1$, $P(\gamma)^2 = 1$, $P(\alpha)P(\gamma)P(\alpha)^{-1}P(\gamma) = u^2 1$, $P(\beta)P(\gamma)P(\beta)^{-1}P(\gamma) = v^2 1$ is associated to a UA rep. The operator $P(\gamma)$ is antiunitary.

$g_2^2 g_3^{-2} = 1$ and g_1 is arbitrary. A set $\{g_1, g_2 = \pm 1, g_3 = \pm 1\}$ is equivalent to $\{1, \pm 1, \pm 1\}$ choosing $g_1' = u(\alpha)^2 g_1$ with $u(\alpha)^2 = g_1^{-1}$. Consequently the classes of nonequivalent Φ's have representatives $\{g_1 = 1, g_2 = \pm 1, g_3 = \pm 1\}$. Hence $M(G,H) \cong C_2 \times C_2$ has four elements.

In the same way one finds the comultiplicators for the other crystallographic groups. They are given in Table I. Notice that for a cyclic group the comultiplicator is only trivial if $H = G$, i.e., for PU reps. Moreover, any $M(G,H)$ in the list is a direct product of k cyclic groups C_2. Hence any factor system is of order 2. The PU reps of three-dimensional crystallographic point groups were already determined by Döring[11] and Hurley.[12] In the latter paper explicit matrix representations are given.

Although the group (3.8) is isomorphic to $M(G,H)$, one particular solution Φ does not determine a factor system uniquely. Define a subgroup of $C^1(G)$ by

$$\tilde{C}^1(G) = \{c \in C^1(G) \mid c(\alpha_j) = 1, j = 1, \ldots, \nu\}.$$

The group $\tilde{B}_H^2(G) = \delta\tilde{C}^1(G)$ is a subgroup of $B_H^2(G)$. According to Ref. 7, p. 52, [case 2i] one has

Theorem 3: An element Φ satisfying

$$\prod_{i=1}^{r} g_i^{k_{ij}} = 1, \quad j = 1, \ldots, \nu,$$

determines an element of $Z_H^2(G)/\tilde{B}_H^2(G)$. This means that Φ determines a factor system up to an element of $\tilde{B}^2(G)$. This proves that it is justified to give the nonequivalent factor systems by nonequivalent Φ's. To determine one factor system corresponding to Φ one chooses for any element $\alpha \in G$ a word $\alpha = w_\alpha(\alpha_1, \ldots, \alpha_\nu)$ in the generators. Then take a section r in such a way that $r(\alpha) = w_\alpha(r(\alpha_1), \ldots, r(\alpha_\nu))$. One has for the factor system $\omega(\alpha,\beta) = r(\alpha)r(\beta)r(\alpha\beta)^{-1} = w_\alpha(r(\alpha_1), \ldots, r(\alpha_\nu))w_\beta(r(\alpha_1), \ldots, r(\alpha_\nu))w_{\alpha\beta}^{-1}(\ldots) \equiv W_{\alpha,\beta}(r(\alpha_1), \ldots, r(\alpha_\nu))$. As σ maps $W_{\alpha,\beta}$ on ϵ, it is the product of certain $\Phi_i(r(\alpha_1), \ldots, r(\alpha_\nu))$ and conjugates $r(\alpha)\Phi_i(r(\alpha_1), \ldots, r(\alpha_\nu))r(\alpha)^{-1} = \Phi_i^\alpha(r(\alpha_1), \ldots, r(\alpha_\nu))$. Then one has

$$\omega(\alpha,\beta) = \prod_{i=1}^{r} \Phi_i(r(\alpha_1), \ldots, r(\alpha_\nu))^{c_i(\alpha,\beta)} = \prod_{i=1}^{r} g_i^{c_i(\alpha,\beta)}$$

$$\text{for } c_i(\alpha,\beta) \in ZG. \qquad (3.9)$$

As an example we determine a factor system ω corresponding to a given Φ for the PUA reps of (D_2, C_2).

Choosing $r(\alpha_1\alpha_2) = r(\alpha_1)r(\alpha_2)$ one has

$$\omega(\alpha_1, \alpha_1) = r(\alpha_1)^2 = g_1 = 1,$$
$$\omega(\alpha_1, \alpha_2) = r(\alpha_1)r(\alpha_2)r(\alpha_2)^{-1}r(\alpha_1)^{-1} = 1,$$
$$\omega(\alpha_1, \alpha_1\alpha_2) = r(\alpha_1)r(\alpha_1\alpha_2)r(\alpha_2)^{-1} = 1,$$
$$\omega(\alpha_2, \alpha_1) = r(\alpha_2)r(\alpha_1)r(\alpha_2)^{-1}r(\alpha_1)^{-1}$$
$$= r(\alpha_1)^{-1}r(\alpha_1)r(\alpha_2)r(\alpha_1)r(\alpha_2)r(\alpha_2)^{-2}r(\alpha_1)$$
$$= r(\alpha_1)^{-1}g_3g_2^{-1}r(\alpha_1)$$
$$= g_3^{\alpha_1}g_2^{-\alpha_1} = (\pm 1)(\pm 1) = \pm 1,$$
$$\omega(\alpha_2, \alpha_2) = r(\alpha_2)^2 = g_2 = \pm 1,$$
$$\omega(\alpha_2, \alpha_1\alpha_2) = r(\alpha_2)r(\alpha_1\alpha_2)r(\alpha_1)^{-1}$$
$$= r(\alpha_1)^{-1}r(\alpha_1)r(\alpha_2)^2r(\alpha_1)r(\alpha_1)^{-2}$$
$$= g_3^{\alpha_1}g_1^{-1} = \pm 1,$$
$$\omega(\alpha_1\alpha_2, \alpha_1) = r(\alpha_1)r(\alpha_2)r(\alpha_1)r(\alpha_2)^{-1} = g_3g_2^{-1} = \pm 1,$$
$$\omega(\alpha_1\alpha_2, \alpha_2) = r(\alpha_1)r(\alpha_2)^2r(\alpha_1)^{-1} = g_2^{\alpha_1} = \pm 1,$$
$$\omega(\alpha_1\alpha_2, \alpha_1\alpha_2) = g_3 = \pm 1.$$

The nine values $\omega(\alpha, \beta)$ are determined by the three components of Φ once one has chosen $r(\alpha_1\alpha_2) = r(\alpha_1)r(\alpha_2)$. Another choice would give an associated factor system.

A second method to determine $M(G,H)$ which does not require solution of Eqs. (3.5) is the following.

Lemma 1: $H_H^2(G) \cong H_\phi^3(G, Z)$.

Proof: The group $U(1)$ is isomorphic to the factor group of the additive group R of real numbers and the additive group of integers. One has

$$0 \to Z \to R \to R/Z \to 0$$

and consequently a long exact sequence[9]

$$\cdots \to H_\phi^n(G,R) \to H_\phi^n(G,R/Z) \to H_\phi^{n+1}(G,Z) \to H_\phi^{n+1}(G,R) \to \cdots, \quad (3.10)$$

where the action ϕ on an element a of R, Z or G/Z is $\phi(\alpha)a = a (a \in H), \phi(\alpha)a = -a (\alpha \in G\text{-}M)$. Because G is finite and R divisible and torsion-free $H_\phi^n(G,R) = 0$ for $n > 0$. Hence

$$H_\phi^n(G, R/Z) \cong H_\phi^{n+1}(G, Z). \quad (3.11)$$

As $U(1) \simeq (R/Z)$ the lemma follows by taking $n = 2$.

We now apply a reduction theorem by Eilenberg and McLane.[13] If $C^1(G,Z)$ is the group of 1-cochains with integral values, one has the isomorphism

$$H_\phi^{n+1}(G,Z) \cong H_\sigma^n(G, C^1(G,Z)), \quad n > 0, \quad (3.12)$$

where $C^1(G,Z)$ is a G module with action σ defined by

$$[\sigma(\alpha)f](\beta) = f(\beta\alpha) - f^\beta(\alpha), \quad f \in C^1(G,Z), \alpha,\beta \in G. \quad (3.13)$$

Because f is an integral-valued function with $f(\epsilon) = 0$, the group $C^1(G,Z)$ is isomorphic to Z^{N-1}, where N is the order of G. It is the subring of ZG of all functions with $f(\epsilon) = 0$. Then we have the following lemma.

Lemma 2: For a finite group G one has the following isomorphisms:

$$M(G,H) \cong H_H^2(G) \cong H_\phi^3(G,Z) \cong H_\sigma^2(G, Z^{N-1}). \quad (3.14)$$

The determination of $M(G,H)$ is equivalent to that of $H_\sigma^2(G, Z^{N-1})$. In Ref. 10 a technique was developed to determine $H_\sigma^2(G, Z^n)$ for finite groups and an arbitrary homomorphism $\sigma: G \to \mathrm{Aut}(Z^n)$. The matrices $\sigma(G)$ can be found by choosing a basis $\alpha_1, \ldots, \alpha_N$ in the integral group ring. The elements $\alpha_2, \ldots, \alpha_N$ form a basis for $C^1(G,Z)$. With respect to this basis

$$\sigma(\alpha)\cdot\alpha_i = \alpha_i\alpha^{-1} - \delta_{\alpha,\alpha_i}\sum_{\beta\in G}\beta, \quad \forall\alpha \in G. \quad (3.15)$$

4. REPRESENTATION GROUPS

In this section we show the existence of a group R such that any PUA rep of (G,H) can be lifted to an ordinary UA rep of R. First we treat a generalization of Schur's lemma.

Lemma 3: If a unitary operator S satisfies $SU = US$, $AS^{-1} = SA$ for any unitary U and antiunitary A from a set $\mathcal{J}$ of operators on an $\mathcal{J}$-invariant irreducible space $\mathcal{H}$, it is a scalar multiple of the identity operator.

Proof: Define a new operator $T = S - \lambda\mathbb{1}$ for some eigenvalue λ of S. We show that the kernel of T is $\mathcal{J}$-invariant. If x is an eigenvector of S with eigenvalue λ, it belongs to the kernel of T and it is an eigenvector of S^{-1} with eigenvalue λ^*. Then for any $A, U \in \mathcal{J}$ one has $TUx = UTx = 0$ and $TAx = SAx - \lambda Ax = (AS^{-1} - A\lambda^*)x = (AS^{-1} - \lambda^*\mathbb{1})x = 0$. Since $\mathcal{H}$ is irreducible, the kernel is either $\mathcal{H}$ or 0. The latter is impossible as any eigenvector with eigenvalue λ belongs to the kernel. Hence the kernel is $\mathcal{H}$, i.e., $T = 0$ and $S = \lambda\mathbb{1}$.

Now we consider an extension $1 \to A \to R \xrightarrow{\sigma} G \to 1$ such that the action ϕ of G on the Abelian group A is given by $\phi(\alpha)a = a^\alpha = a$, $\phi(\beta)a = a^\beta = a^{-1}$ for any $a \in A$, $\alpha \in G$, $\beta \in G\text{-}H$. We denote the subgroup R which is mapped by σ on $H \subseteq G$ by U. Taking a section $r: G \to R$ one has $r(\alpha)r(\beta) = m(\alpha,\beta)r(\alpha\beta)$ and $r(\alpha)ar(\alpha)^{-1} = a^\alpha (a \in A; \alpha, \beta \in G)$, where m is a 2-cocycle. Another section r' given by $r(\alpha) = u(\alpha)r(\alpha)$ for some $u \in C^1(G,A)$ gives $r'(\alpha)ar'(\alpha)^{-1} = a^\alpha$ and $m' = m\delta u$.

If D is an irreducible UA representation of (R, U) in a space $\mathcal{H}$, one has for any $a \in A$, $\alpha \in H$, $\beta \in G\text{-}H$ the relations $D(\alpha)D(a) = D(a)D(\alpha)$ and $D(a)D(\beta) = D(\beta)D(a)^{-1}$. Notice that A considered as subgroup of R belongs to U. Therefore, $D(a)$ is a unitary operator satisfying the conditions of lemma 3. This means that $D(a) = \chi(a)\mathbb{1}$ for some 1-cocycle (homomorphism) $\chi \in Z_A^1(A)$ with values in $U(1)$, where the action of A on $U(1)$ is trivial. The UA rep D of R gives a PUA rep P of (G,H) by choosing a section $r: G \to R$ and the definition $P(\alpha) = D(r(\alpha))$. The factor system is given by

$$\omega(\alpha,\beta)\mathbb{1} = P(\alpha)P(\beta)P(\alpha\beta)^{-1} = D(r(\alpha)r(\beta)r(\alpha\beta)^{-1})$$
$$= D(m(\alpha,\beta)) = \chi(m(\alpha,\beta))\mathbb{1}$$

or

$$\omega(\alpha,\beta) = \chi(m(\alpha,\beta)). \quad (4.1)$$

Because

$$\omega^{\alpha}(\beta,\gamma)\omega(\alpha,\beta\gamma)\omega^{-1}(\alpha,\beta)\omega^{-1}(\alpha\beta,\gamma)$$
$$=\chi(m^{\alpha}(\beta,\gamma)m(\alpha,\beta\gamma)m^{-1}(\alpha,\beta)m^{-1}(\alpha\beta,\gamma))$$
$$=\chi(e)=1,$$

ω is a 2-cocycle: $\omega\in Z_H^2(G)$.

Another section r' with $r'(\alpha)=u(\alpha)r(\alpha)$ gives a factor system $\omega'=\omega\delta c$ with $c(\alpha)=\chi(u(\alpha))$. Therefore, χ maps the cohomology class $[m]$ on the cohomology class $[\omega]$.

The group $Z_A^1(A)$ is the same as the character group A^* which is isomorphic with A: By this isomorphism to any $a\in A$ corresponds a 1-cocycle a^*. For any element of A one defines $\rho(a)=[a^*(m)]$, the cohomology class of $a^*(m)=\omega$. Since a^* maps the cohomology class of m on that of ω, ρ does not depend on the section r. The mapping $\rho: A\to H_H^2(G)$ is a homomorphism because

$$\rho(a)\rho(b)=[a^*(m)][b^*(m)]=[(ab)^*(m)]=\rho(ab). \tag{4.2}$$

Some elements of $Z_A^1(A)$ are restrictions to A of elements of $Z_U^1(R)$, crossed homomorphisms $R\to U(1)$. We define a subgroup $B\subseteq A$ by

$$B=\{a\in A\,|\,a^*\in Z_A^1(A) \text{ is the restriction to } A \text{ of a } \chi\in Z_U^1(R)\}. \tag{4.3}$$

If $\rho(a)=\rho(b)$ for $a,b\in A$, one has $a^*(m)=\delta c\cdot b^*(m)$ or $(ab^{-1})^*(m)\in B_H^2(G)$. Then the mapping $\chi: R\to U(1)$ defined by

$$\chi(ur(\gamma))=(ab^{-1})^*(u)\cdot c(\gamma),\qquad u\in A,\ \gamma\in G$$

is a crossed homomorphism since

$$\chi(ur(\gamma)vr(\delta))=\chi(uv^{\gamma}m(\gamma,\delta)r(\gamma\delta))$$
$$=(ab^{-1})^*(uv^{\gamma})\cdot c(\gamma)c^{\gamma}(\delta)c(\gamma\delta)^{-1}\cdot c(\gamma\delta)$$
$$=(ab^{-1})^*(u)\cdot c(\gamma)\cdot(ab^{-1})^*(v)^{\gamma}\cdot c^{\gamma}(\delta)$$
$$=\chi(ur(\gamma)\chi^{\gamma}(vr(\delta)).$$

Since $\chi\in Z_U^1(R)$ and $\chi(A)=(ab^{-1})^*(A)$, one has $ab^{-1}\in B$. On the other hand if $ab^{-1}\in B$, there exists a $\chi\in Z_U^1(R)$ such that $\chi(A)=(ab^{-1})^*(A)$. Define $c(\alpha)=\chi(r(\alpha))$. Then

$$\omega(\alpha,\beta)=(ab^{-1})^*(m(\alpha,\beta))=\chi(r(\alpha)r(\beta)r(\alpha\beta)^{-1})$$
$$=c(\alpha)c^{\alpha}(\beta)c(\alpha\beta)^{-1}=\delta c(\alpha,\beta). \tag{4.4}$$

Consequently a^* and b^* determine associated factor systems if and only if $ab^{-1}\in B$. This means that $\mathrm{Ker}_{\rho}=B$. Hence

$$\mathrm{Im}_{\rho}=\frac{A}{\mathrm{Ker}_{\rho}}=\frac{A}{B}. \tag{4.5}$$

Any irreducible UA rep D of (R,U) gives a PUA rep P of (G,H) by $P(\alpha)=D(r(\alpha))$. The factor system of P is given by $\omega=a^*(m)$ for some $a\in A$. If any PUA rep of (G,H) can be lifted, this means that $\rho: A\to H_H^2(G)$ is an epimorphism. Then one has for the orders of the groups: $|H_H^2(G)|=|\mathrm{Im}_{\rho}|=|A|\cdot|B|^{-1}$. For the order of R one has $|R|=|A|\cdot|G|=|H_H^2(G)|\cdot|B|\cdot|G|\geq|H_H^2(G)|\cdot|G|$. Hence if R gives all PUA reps of (G,H), it is at least of order $|H_H^2(G)|\cdot|G|$. We will construct a group R with this property for which the order is minimal. We call such a group a (co)representation group. It is an extension of $M(G,H)$ by G.

Suppose G is generated by $\alpha_1,\ldots,\alpha_{\nu}$ with defining relations $\Phi_i(\alpha_1,\ldots,\alpha_{\nu})=\epsilon(i=1,\ldots,r)$. The Abelian group $H_H^2(G)$ is determined by Theorem 2. As any finite Abelian group it is isomorphic to a direct product of cyclic groups:

$$H_H^2(G)\simeq\prod_{j=1}^{s}C_{d_j}, \tag{4.6}$$

where C_n is a cyclic group of order n. If the generator of C_{d_j} corresponds via this isomorphism to a solution Φ_j' of (3.4), the elements of C_{d_j} correspond to the solutions $(\Phi_j')^m(0\leq m<d_j)$. Because $(\Phi_j')^{d_j}$ is equivalent to $\Phi_0=\{g_i=1,i=1,\ldots,r\}$ there is a solution Φ_j equivalent to Φ_j' such that $\{\Phi_j\}^{d_j}=\Phi_0$. Then the r components of Φ_j are d_jth roots of unity. If e_j is a primitive d_jth root there is an integer n_{ij} such that the ith component g_{ij} of $\{\Phi_j\}$ satisfies

$$g_{ij}=e_j^{n_{ij}}.$$

An arbitrary solution of (3.4) is equivalent to

$$\Phi=\prod_{j=1}^{s}\Phi_j^{p_j},\qquad 0\leq p_j<d_j. \tag{4.7}$$

Consider a group A which is isomorphic to $H_H^2(G)$ (4.6) generated by $a_1,\ldots,a_s$. Define r elements of A by

$$f_i=\prod_{j=1}^{s}a_j^{n_{ij}},\qquad i=1,\ldots r. \tag{4.8}$$

Because Φ_j is a solution of (3.4), one has

$$\prod_{i=1}^{r}g_{ij}^{k_{i\rho}}=\prod_{i=1}^{r}e_j^{n_{ij}k_{i\rho}}=1,\qquad \text{any } j,\rho$$

or

$$\sum_{i=1}^{r}n_{ij}k_{i\rho}\equiv 0,\qquad \mathrm{mod}\,d_j.$$

Consequently

$$\prod_{i=1}^{r}f_i^{k_{i\rho}}=\prod_{i=1}^{r}\prod_{j=1}^{s}a_j^{n_{ij}k_{i\rho}}=1,\qquad \forall\rho. \tag{4.9}$$

This means that $\{f_i\}$ defines an extension of A by G:

$$1\to A\to R\to G\to 1.$$

When P is a PUA rep of (G,H) with factor system ω, one can define an element $u\in Z_A^1(A)$ by $u(a_j)=e_j^{p_j}$. Then

$$u(f_i)=\prod_{j=1}^{s}u(a_j)^{n_{ij}}=\prod_{j=1}^{s}e_j^{n_{ij}p_j}=\prod_{j=1}^{s}g_{ij}^{p_j}=g_i,$$

which is the ith component of Φ (4.7). As Φ determines a class $[\omega]\in H_H^2(G)$ and $\{f_i\}$ a class $[m]\in H_{\Phi}^2(G,A)$, there are elements $\omega\in Z_H^2(G)$ and $m\in Z_{\Phi}^2(G,A)$ such that $\omega(\alpha,\beta)=u(m(\alpha,\beta))$. Then the mapping $D: R\to\mathfrak{A}(\mathfrak{K})$ defined by $D(ar(\alpha))=\chi(a)P(\alpha)$ is a UA rep of (R,U) because

$$D(ar(\alpha))D(br(\beta))=\chi(a)P(\alpha)\chi(b)P(\beta)$$
$$=\chi(a)\chi^{\alpha}(b)P(\alpha)P(\beta)=\chi(ab^{\alpha})\omega(\alpha,\beta)P(\alpha\beta)$$
$$=\chi(ab^{\alpha}m(\alpha,\beta))P(\alpha\beta)=D[ab^{\alpha}m(\alpha\beta)r(\alpha\beta)]$$
$$=D[\alpha r(\alpha)br(\beta)].$$

TABLE II. Character table for the dihedral group D_4.

	$\{e\}$	$\{a\}$	$\{\bar{\alpha}_1\bar{\alpha}_2, a\bar{\alpha}_1\bar{\alpha}_2\}$	$\{\bar{\alpha}_1, a\bar{\alpha}_1\}$	$\{\bar{\alpha}_2, a\bar{\alpha}_2\}$
$\Gamma 1$	1	1	1	1	1
$\Gamma 2$	1	1	1	-1	-1
$\Gamma 3$	1	1	-1	1	-1
$\Gamma 4$	1	1	-1	-1	1
$\Gamma 5$	2	-2	0	0	0

Hence the PUA rep P can be lifted to the UA rep D. As $|R| = |H_H^2(G)| \cdot |G|$ it is a group of minimal order. Hence it is a (co)representation group. Now we have proved the following theorems.

Theorem 4: For any finite group G there exists a finite group R such that any PUA rep of G with respect to a subgroup H can be lifted to a UA rep of R with respect to a subgroup U. The group R is an extension of the comultiplicator $M(G,H)$ by G, where the action of G on $M(G,H)$ is given by $a^\alpha = a$, $a^\beta = a^{-1}$ for any $a \in M(G,H)$, $\alpha \in H$, $\beta \in G-H$. The canonical epimorphism $\sigma: R \to G$ maps U on H.

Theorem 5: If G is generated by $\alpha_1, \ldots, \alpha_\nu$ with defining relations $\Phi_i(\alpha_1, \ldots, \alpha_\nu) = \epsilon (i = 1, \ldots, r)$ and if

$$M(G,H) \simeq \prod_{j=1}^{s} C_{d_j},$$

a (co)representation group R is generated by $a_1, \ldots, a_s, \bar{\alpha}_1, \ldots, \bar{\alpha}_\nu$, with defining relations

$$\begin{aligned} &a_j^{d_j} = e, && j = 1, \ldots, s, \\ &a_j a_k = a_k a_j, && 1 \leq j, k \leq s, \\ &\Phi_i(\bar{\alpha}_1, \ldots, \bar{\alpha}_\nu) = f_i, && i = 1, \ldots, r, \\ &\bar{\alpha}_\rho a_j \bar{\alpha}_\rho^{-1} = a_j^{\alpha_\rho}, && j = 1, \ldots, s; \rho = 1, \ldots, \nu, \end{aligned} \tag{4.10}$$

where f_i is given by Eq. (4.8). Any irreducible PUA rep of (G,H) can be lifted to an irreducible UA rep of $(R, \sigma^{-1}(H))$.

The UA reps of (R,U) are found from the unitary reps of U.[5,14] When D is an irreducible UA rep of (R,U) in $\mathfrak{K}$, either $\mathfrak{K}$ is irreducible under U (case 1) or it is reducible in two nonequivalent (case 2) or equivalent (case 3) components. When a U-irreducible subspace carries a representation with character χ, one distinguishes the three cases by $(R = U + \beta U)$:

$$\sum_{\alpha \in U} \chi((\beta\alpha)^2) = \begin{cases} \text{order } N \text{ of } U & \text{in case 1} \\ 0 & \text{in case 2} \\ -N & \text{in case 3.} \end{cases} \tag{4.11}$$

On the other hand, if one has a unitary representation of U in $\mathfrak{K}$ for which Eq. (4.11) gives N, the space $\mathfrak{K}$ carries an irreducible UA rep of R. If it gives 0 or $-N$, there is an antiunitary operator T_β such that $\mathfrak{K} + T_\beta \mathfrak{K} \neq \mathfrak{K}$ carries an irreducible UA rep of R. From this it follows that one has to consider the unitary representations of U. We can choose the generators and relations of G in such a way that $\alpha_1, \ldots, \alpha_{\nu-1}$ generate H and $\alpha_\nu \in G-H$. Then U is generated by $a_1, \ldots, a_s, \bar{\alpha}_1, \ldots, \bar{\alpha}_{\nu-1}$ with defining relations as in (4.10), leaving out the relations in which $\bar{\alpha}_\nu$ occurs.

As a first *example* we consider the PUA reps of the cyclic group C_2 with respect to its unit element. From Table I it follows that the comultiplicator has two elements. Hence the corepresentation group is generated by a and $\bar{\alpha}$ with relations $a^2 = e$, $\bar{\alpha}^2 = a$, $\bar{\alpha} a \bar{\alpha}^{-1} = a^{-1} = a$. This is the cyclic group of order 4 generated by $\bar{\alpha}$. The subgroup U is generated by a and is the cyclic group of order 2. This group has two irreducible representations: $D_\pm(a) = \pm 1$. As $\chi(\bar{\alpha}^2) + \chi(\bar{\alpha} a \bar{\alpha} a) = 2\chi(a) = \pm 2$, the representation space (one-dimensional) for D_+ is irreducible for the UA rep, whereas the one-dimensional space for D_- gives a two-dimensional UA rep. Hence there are two nonsimilar PUA reps for C_2: one one-dimensional and one two-dimensional. The first is realized by the complex-conjugation operator θ_0, the other by $P(\alpha) = \sigma_2 \theta_0$ with σ_2 the second Pauli matrix. As is well known the first occurs for an even number of fermions, the second for an odd number. These are the only two possibilities to represent the time reversal transformation by an antiunitary operator.

A second *example* is the determination of the nonsimilar PUA reps of the dihedral group D_2 with respect to itself (PU reps). The group is generated by α_1 and α_2 with defining relations $\alpha_1^2 = \alpha_2^2 = (\alpha_1\alpha_2)^2 = \epsilon$. According to Table I, the multiplicator $M(G,G)$ is given by $\{g_1 = g_2 = 1, g_3 = \pm 1\}$. Hence the representation group R is generated by $a, \bar{\alpha}_1, \bar{\alpha}_2$ with defining relations

$$a^2 = \bar{\alpha}_1^2 = \bar{\alpha}_2^2 = e, \qquad (\bar{\alpha}_1\bar{\alpha}_2)^2 = a, \qquad \bar{\alpha}_1 a \bar{\alpha}_1^{-1} = \bar{\alpha}_2 a \bar{\alpha}_2^{-1} = a.$$

It is the dihedral group D_4 with eight elements and five classes. Its character is given in Table II. Choosing $r(\alpha_1) = \bar{\alpha}_1$, $r(\alpha_2) = \bar{\alpha}_2$, $r(\alpha_1\alpha_2) = \bar{\alpha}_1\bar{\alpha}_2$, $r(\epsilon) = e$ one obtains five nonequivalent PU reps of D_2. However, the four one-dimensional representations are all associated (cf. remark at the end of Sec 2). Hence there are two classes of nonsimilar irreducible PU reps of D_2: one one-dimensional, one two-dimensional. The one-dimensional representation has trivial factor system.

5. LIFTING OF A GIVEN PUA REP

In Sec. 4 we showed that any PUA rep of (G,H) can be lifted to a UA rep of a corepresentation group. A related problem is the following. Suppose a PUA rep of G is given, e.g., as a set of operators commuting with a Hamiltonian (See, e.g., Ref. 2.) To apply the group-theoretical methods one wants to extend this set to a group by adding appropriate phase factors. This means that one wants to lift the PUA rep. The smallest group which gives this lifting is, in general, only a subgroup of a representation group.

If P is a PUA rep with factor system ω, the class of ω generates a cyclic subgroup of $M(G,H)$. As $\omega^d \in B_H^2(G)$ for some d, there is an equivalent factor system consisting of dth roots. In the same way as in Sec. 4 one can define an extension of the cyclic group C_d by G:

$$1 \longrightarrow C_d \longrightarrow K \xrightarrow{\sigma} G \longrightarrow 1. \tag{5.1}$$

If again $G = \{\alpha_1, \ldots, \alpha_\nu \mid \Phi_i(\alpha_1, \ldots, \alpha_\nu) = \epsilon, i = 1, \ldots, r\}$, one has

$$\Phi_i(P(\alpha_1), \ldots, P(\alpha_\nu)) = g_i\, 1. \tag{5.2}$$

In order to determine K one determines the smallest

nonnegative integer d such that

$$g_i^d = \prod_{j=1}^{\nu} u(\alpha_j)^{m_{ij}} \text{ for some } u \in C^1(G).$$

Then define

$$g_i' = g_i \prod_{j=1}^{\nu} u(\alpha_j)^{-m_{ij}/d} = e_d^{n_i} \tag{5.3}$$

which are dth roots. If a cyclic group C_d is generated by a, the elements

$$f_j = a^{n_j} \in C_d \tag{5.4}$$

determine an extension (5.1). Any PUA rep with factor system equivalent to ω^p for some $0 \leq p < d$ can be lifted to a UA rep of K. Then quite analogous to Theorem 5 one has the following.

Theorem 6: A PUA rep of (G, H) with factor system ω of order d can be lifted to a UA rep of (K, U), where K is generated by $a, \bar{\alpha}_1, \ldots, \bar{\alpha}_\nu$ with defining relations

$$\begin{aligned} &a^d = e, \\ &\Phi_i(\bar{\alpha}_1, \ldots, \bar{\alpha}_\nu) = a^{n_i}, \quad \text{cf. Eq. (5.3)}, \ i = 1, \ldots, r, \\ &\bar{\alpha}_j a \bar{\alpha}_j^{-1} = a^{\alpha_j}, \ j = 1, \ldots, \nu, \end{aligned} \tag{5.5}$$

and U is the subgroup of K mapped on H by the canonical epimorphism $\sigma: K \to G$.

6. UA REPS OF SUBGROUPS OF THE INHOMOGENEOUS LORENTZ GROUP

Important symmetry groups for physics are subgroups of the inhomogeneous Lorentz group. Among these groups are the ordinary space groups [subgroups of the Euclidean group IO(3)], the magnetic space groups [subgroups of $\mathbf{IO}(3) \times J$, when J is the group generated by time reversal] and space-time groups. They have a translation subgroup which is an invariant Abelian subgroup. The UA rep can be found by the method of induction[15] sketched below. Here we will consider the UA reps of those groups which have a finite point group. Among others this includes the cases of unitary representations of space groups discussed, e.g., in Ref. 16, and the UA reps of magnetic space groups.[14]

We denote the group by G, its translation subgroup by T, its point group by $K \simeq G/T$. The subgroup of G represented by unitary operators has point group $H \subseteq K$, whereas the elements $\{\alpha \mid t\} \in G$ with $\alpha \in K\text{-}H$ are represented by antiunitary operators. The translation subgroup T is represented by unitary operators. Consider an irreducible representation $D(G)$. Since the irreducible representations of T are one dimensional, the carrier space $\mathcal{K}$ of $D(G)$ decomposes into a sum of carrier spaces $\mathcal{K}_k$ of representations of T characterized by a vector k in the Brillouin zone. If $\psi \in \mathcal{K}_k$ and $\{\alpha \mid t\} \in G$, the element $D(\{\alpha \mid t\})\psi$ belongs to $\mathcal{K}_{k'}$ with $k' = \dot{\alpha}k (= \alpha k$ for $\alpha \in H, = -\alpha k$ for $\alpha \in K\text{-}H)$, because $D(\{\epsilon \mid a\})D(\{\alpha \mid t\})\psi = \exp(\pm ika) D(\{\alpha \mid t\})\psi$. Hence $\mathcal{K} = \oplus_i \mathcal{K}_{k_i}$, where $k_1, \ldots, k_s$ are the nonequivalent k vectors from $\{\dot{\alpha}k \mid \forall\ \alpha \in K\}$. Moreover, if we define the group of k by

$$G_k = \{\{\alpha \mid t\} \in G \mid \dot{\alpha}k \text{ equivalent to } k\}, \tag{6.1}$$

the space $\mathcal{K}_k$ is invariant under G_k. By the decomposition

$$G = G_k + g_2 G_k + \cdots + g_s G_k \tag{6.2}$$

one has $D(g)D(g_i)\mathcal{K}_k \subseteq D(g_j)\mathcal{K}_k$ if $gg_i = g_j h \in g_j G_k$. Hence $\mathcal{K}_{k_1}, \ldots, \mathcal{K}_{k_s}$ is a system of imprimitivity. The representation $D(G)$ is found by induction from the irreducible UA reps D_k of G_k with the property $D_k(\{\epsilon \mid a\}) = \exp(ika) D_k(\{\epsilon \mid 0\})$. Hence one can write

$$D_k(\{\alpha \mid t\}) = \exp(ikt) P(\alpha), \quad \{\alpha \mid t\} \in G_k, \tag{6.3}$$

where the operators $P(\alpha)$ form a PUA rep of the point group K_k of G_k.

$$P(\alpha)P(\beta) = \exp(i[\alpha^{-1}(k - \dot{\alpha}k)]t_\beta) P(\alpha\beta) \tag{6.4}$$

because $P(\alpha)$ does not depend on the nonprimitive translation t_α. It is readily verified that $\omega(\alpha, \beta) = \exp\{i[\alpha^{-1}(k - \dot{\alpha}k)]t_\beta\}$ forms a factor system. Consequently the irreducible UA reps of G are determined by the "stars" $k_1, \ldots, k_s$ and the irreducible PUA reps (6.4) of K_k. To find all nonequivalent irreducible PUA reps one can use Theorem 6. The preceding sketch of the method of induction is only meant to indicate the line and to establish the factor system (6.4).

If G has elements $\{\alpha \mid t_\alpha + a\}$, with $\alpha \in K$, $a \in T$ and nonprimitive translation t_α, an equivalent (affine conjugated) group has elements $\{\alpha \mid a + \bar{t}_\alpha\}$, with $\bar{t}_\alpha = (1 - \alpha)v + a_\alpha + t_\alpha$, where v is an arbitrary translation, $a_\alpha \in T$. For this equivalent group $\bar{G}$ one finds the UA reps from PUA reps of K with factor system

$$\bar{\omega}(\alpha, \beta) = \exp(i[\alpha^{-1}(k - \dot{\alpha}k)]\bar{t}_\beta) = \omega(\alpha, \beta) \frac{u(\alpha)u^\alpha(\beta)}{u(\alpha\beta)}, \tag{6.5}$$

with $u(\alpha) = \exp(i\alpha^{-1}(k - \dot{\alpha}k)v)$. Hence equivalent groups determine equivalent factor systems. Consequently Eq. (6.4) gives a homomorphism ψ_k from the group of nonequivalent systems of nonprimitive translations corresponding to the arithmetic point group $\phi(k)$ to $H^2_\beta(K)$. This homomorphism ψ_k depends on k. For a symmorphic group $\mathrm{Im}\psi_k = 0$.

As an example we consider UA reps of a two-dimensional nonsymmorphic magnetic space group from the arithmetic class $R2'm'm$. It has a point group K generated by

$$\alpha_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \alpha_2 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

from the isomorphism class D_2, and associated nonprimitive translations

$$t_{\alpha_1} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}, \quad t_{\alpha_2} = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}, \quad t_{\alpha_1\alpha_2} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

The elements α_2 and $\alpha_1\alpha_2$ are combined with time reversal. The Brillouin zone is rectangular. For $k = (\pi, \pi)$ on the border of the zone the group G_k is the whole space group. Hence $K_k = K \simeq D_2$. A group from the isomorphism class D_2 is generated by α_1, α_2 with

$$\begin{aligned} g_1 &= P(\alpha_1)^2 = \omega(\alpha_1, \alpha_1) \\ &= -1, \text{ because } (\alpha_1 - 1)k \cdot t_{\alpha_1} = -\pi, \end{aligned}$$

$g_2 = P(\alpha_2)^2 = \omega(\alpha_2, \alpha_2)$
$= -1$, because $(\alpha_2 + 1)k \cdot t_{\alpha_2} = \pi$,
$g_3 = [P(\alpha_1)P(\alpha_2)]^2 = \omega(\alpha_1\alpha_2, \alpha_1\alpha_2)$
$= 1$, as $\alpha_1\alpha_2 + 1 = 0$.

Hence the PUA reps of D_2 with factor system (6.4) are obtained from the UA reps of the extension B of C_2 by K generated by $a, \bar{\alpha}_1, \bar{\alpha}_2$ with defining relations $a^2 = (\bar{\alpha}_1\bar{\alpha}_2)^2 = 1$, $\bar{\alpha}_1^2 = \bar{\alpha}_2^2 = a$, $\bar{\alpha}_1 a = a\bar{\alpha}_1$, $\bar{\alpha}_2 a = a\bar{\alpha}_2$. It is the group $C_4 \times C_2$ of order 8, generated by $\bar{\alpha}_1$ and $\bar{\alpha}_1\bar{\alpha}_2$. Its subgroup U mapped on ϵ and α_1 of K is the cyclic group generated by $\bar{\alpha}_1$. It has four nonequivalent irreducible unitary representations, all one dimensional: $\Gamma_\nu(\bar{\alpha}_1) = i^{\nu-1}(\nu = 1, \ldots, 4)$. The corresponding UA reps of B are found using $\sum_{h\in U}\chi((\bar{\alpha}_1\bar{\alpha}_2 h)^2) = 2\chi(e) + 2\chi(\bar{\alpha}_1^2) = 0(\nu = 2, 4)$ or $4(\nu = 1, 3)$. The representations Γ_1 and Γ_3 give one-dimensional PUA reps of K with trivial factor system. The representations Γ_2 and Γ_4 give together a two-dimensional PUA rep of K with the nontrivial factor system determined by g_1, g_2, g_3. Hence for $k = (\pi, \pi)$ there is exactly one two-dimensional UA rep of $R2'm'm$.

Notice that this method can also be used if primitive translations of the magnetic space group are combined with time reversal. In this case one can take for T the intersection of G with the group of three-dimensional translations. The elements of T are called primitive translations. In the point group K there appears the time-reversal operator with a nonprimitive translation. As an example take the magnetic two-dimensional space group generated by the magnetic translation $a_1' = (a, 0)$ and the non-magnetic translation $a_2 = (0, b)$. The group T has elements $2na_1' + ma_2$ (n, m integers), K consists of two elements: the unit ϵ and the time reversal α. The time reversal has nonprimitive translation $t_\alpha = a_1'$. Then one can proceed in quite the same way as before.

ACKNOWLEDGMENT

I thank Professor A. Janner for critical remarks.

APPENDIX A: SOME COHOMOLOGICAL AND GROUP-THEORETICAL NOTIONS

A group G is said to be *generated* by elements $\alpha_1, \ldots, \alpha_\nu$ if any $\alpha \in G$ can be written as a product of generators and their inverses. Any such product is called a *word* in the generators. Any word which is equal to the unit element ϵ is a *relation* in the group. If G is completely determined by the relations $\Phi_i(\alpha_1, \ldots, \alpha_\nu) = \epsilon$, the relations $\Phi_1 = \epsilon, \ldots, \Phi_\nu = \epsilon$ are *defining relations*.

The functions on G with integral values form a ring. Formally these functions can be written as $\Sigma_{\alpha\in G} m_\alpha \alpha$ (m_α integer). Addition and multiplication are defined by $(\Sigma_\alpha m_\alpha\alpha) + (\Sigma_\alpha n_\alpha\alpha) = \Sigma_\alpha(m_\alpha + n_\alpha)\alpha$ and $(\Sigma_\alpha m_\alpha\alpha)(\Sigma_\beta n_\beta\beta) = \Sigma_\alpha m'_\alpha\alpha$ with $m'_\alpha = \Sigma_{\beta\in G} m_\alpha m_{\alpha^{-1}\beta}$. With these rules the set forms a ring called *integral group ring* ZG.

For a ring F an F *module* A is an Abelian group A for which a scalar multiplication with elements of F is defined which has to satisfy some distributive conditions. A vector space over a field is a module for which F is a field. When ϕ is a homomorphic mapping of a group G into the group of automorphisms of an Abelian group A, the image $\phi(\alpha)a$ of a is denoted by a^α. This implies that A is a ZG module if one defines $a^{\alpha+\beta} = a^\alpha a^\beta$ and $a^{m\alpha} = (a^\alpha)^m$ (integer m).

When A is a ZG module, an *n-cochain* u is a mapping $G\times\ldots\times G \to A$: $u(\alpha_1, \ldots, \alpha_n) \in A$. The n-cochains form an Abelian group $C_\phi^n(G, A)$ by $(u_1 \cdot u_2)(\alpha_1, \ldots, \alpha_n) = u_1(\alpha_1, \ldots, \alpha_n)u_2(\alpha_1, \ldots, \alpha_n)$. The groups $C_\phi^n(G, A)$ are connected by homomorphisms $\delta_n: C_\phi^n(G, A) \to C_\phi^{n+1}(G, A)$ if one defines $(\delta_n u)(\alpha_1, \ldots, \alpha_{n+1}) = u^{\alpha_1}(\alpha_2, \ldots, \alpha_{n+1}) \prod_{i=1}^n u(\alpha_1, \ldots, \alpha_i\alpha_{i+1}, \ldots, \alpha_{n+1})^{(-1)^i} u(\alpha_1, \ldots, \alpha_n)^{(-1)^{n+1}}$. The first three homomorphisms are

$$(\delta_0 u)(\alpha) = u^{1-\alpha}, \tag{A1}$$

$$(\delta_1 u)(\alpha_1, \alpha_2) = u^{\alpha_1}(\alpha_2)u(\alpha_1\alpha_2)^{-1}u(\alpha_1), \tag{A2}$$

$$(\delta_2 u)(\alpha_1, \alpha_2, \alpha_3) = u^{\alpha_1}(\alpha_1, \alpha_3)u(\alpha_1\alpha_2, \alpha_3)^{-1} u(\alpha_1, \alpha_2\alpha_3)u(\alpha_1, \alpha_2)^{-1}. \tag{A3}$$

The homomorphisms δ_n satisfy $\delta_{n+1}\delta_n = 0$. The kernel of δ_n is denoted by $Z_\phi^n(G, A)$, its elements are *n-cocycles*. The image of δ_{n-1} is denoted by $B_\phi^n(G, A)$, its elements are *n-coboundaries*. Since $B_\phi^n(G, A)$ is a subgroup of the Abelian group $Z_\phi^n(G, A)$ because of $\delta_n\delta_{n-1} = 0$, the factor group $Z_\phi^n(G, A)/B_\phi^n(G, A)$ is again an Abelian group denoted by $H_\phi^n(G, A)$. Its elements are equivalence classes of n-cocycles: cohomology classes. To give an example: for $n = 2$ the 2-cochains are functions $u(\alpha_1, \alpha_2)$, u is a 2-cocycle if $u^{\alpha_1}(\alpha_2, \alpha_3)u(\alpha_1, \alpha_2\alpha_3) = u(\alpha_1\alpha_2, \alpha_3)u(\alpha_1, \alpha_2)$, it is a 2-coboundary if there is a 1-cochain c such that $u(\alpha_1, \alpha_2) = u^{\alpha_1}(\alpha_2)u(\alpha_1)u(\alpha_1\alpha_2)^{-1}$. For $n = 1$: a 1-cochain u is a 1-cocycle if $u(\alpha)u^\alpha(\beta) = u(\alpha\beta)$. This is called a crossed homomorphism. If the action of G is trivial, it is an ordinary homomorphism $u: G \to A$.

A series of groups $\{A_i\}_{i\in Z}$ connected by homomorphisms $\phi_i: A_i \to A_{i+1}$ is an *exact sequence* if $\mathrm{Ker}_{\phi_i} = \mathrm{Im}_{\phi_{i-1}}$. An example is the short exact sequence

$$1 \xrightarrow{\phi_0} A_1 \xrightarrow{\phi_1} A_2 \xrightarrow{\phi_2} A_3 \xrightarrow{\phi_3} 1. \tag{A4}$$

Exactness means $\mathrm{Im}_{\phi_2} = A_3$ (i.e., ϕ_2 is an epimorphism), $\mathrm{Ker}_{\phi_1} = 1$ (i.e., ϕ_1 is a monomorphism) and $\mathrm{Ker}_{\phi_2} = \phi_1(A_1)$. The latter means that $\phi_1(A_1)$ is an invariant subgroup of A_2 such that the factor group $A_2/\phi_1(A_1) \cong A_3$. Then A_2 is called an *extension* of A_1 by A_3. A *section* r for an epimorphism $\sigma: A \to B$ is a mapping $r: B \to A$ such that σr is the identity on B. We choose always a mapping which maps the unit of B on the unit of A. A section r for ϕ_2 in (A4) determines an automorphism of A_1 by $\phi_1(a^\alpha) = r(\alpha)\phi_1(a)r(\alpha)^{-1}$ ($\forall a \in A_1$). Moreover, r determines a mapping $m: A_3 \times A_3 \to A_1$ by $r(\alpha)r(\beta) = m(\alpha, \beta)r(\alpha\beta)$ ($\alpha, \beta \in A_3$). If A_1 is an Abelian group, it becomes a ZA_3 module. The 2-cochain m is a 2-cocycle, as follows from the associativity of the product, and is called a *factor system*. For any 2-cocycle there is an extension and vice versa. For any 2-cocycle m there is an extension A_2 of A_1 by A_3 with product rule $(a, \alpha)(b, \beta) = (a \cdot \phi(\alpha)b \cdot m(\alpha, \beta), \alpha\beta)$. Two extensions A_2 and A_2' of A_1 by A_3 are called *equivalent* if and only if there is an isomorphism $\psi: A_2 - A_2'$ such

that both the restriction to A_1 and the induced automorphism of A_3 are the identity automorphism. Equivalent extensions determine factor systems differing by a 2-coboundary and vice versa. The classes of nonequivalent extensions are in one-to-one correspondence with the elements of $H^2_\phi(A_3, A_1)$. They form a group for that reason.

A *commutative diagram* is a set of groups connected by homomorphisms such that the composition of several such homomorphisms between two groups determines the same homomorphism between the groups. In particular for a diagram

$$\begin{array}{ccc} A & \xrightarrow{\rho} & B \\ {\scriptstyle\sigma}\downarrow & & \downarrow{\scriptstyle\tau} \\ C & \xrightarrow{\nu} & D \end{array}$$

commutativity means that $\tau\rho = \upsilon\sigma$.

APPENDIX B

In Sec. 1 we restricted ourselves to (anti)unitary operators on $\mathfrak{K}$. A ray of operators consists of (anti) unitary operators differing by a complex number of modulus one. However, the rays of $\mathfrak{K}$ consist of vectors differing by an arbitrary complex number. If we had defined operator rays in an analogous way we would have found an exact sequence:

$$1 \to C^\times \to P(\mathfrak{K}) \to \mathfrak{A}(\mathfrak{K}) \to 1,$$

where $C^\times$ is the multiplicative group of nonzero complex numbers and

$$P(\mathfrak{K}) = \{\lambda A \mid A \in \mathfrak{A}(\mathfrak{K}), \lambda \in C^\times\}.$$

Following the same lines the factor systems are then elements of $Z^2_H(G, C^\times)$ when $\alpha c = c (\alpha \in H), \alpha c = c^* (\alpha \in G\text{-}H)$ for any $c \in C^\times$ and classes of associated factor systems are elements of $H^2_H(G, C^\times)$. Now we have

$$H^2_H(G) \cong H^2_H(G, C^\times).$$

For the proof consider the exact sequence $1 \to U(1) \to C^\times \to R^\times \to 1$, where $R^\times$ is the multiplicative group of positive real numbers. From this follows the exact sequence

$$\cdots \to H^n_H(G, R^\times) \to H^{n+1}_H(G) \to H^{n+1}_H(G, C^\times) \to H^{n+1}_H(G, R^\times) \to \cdots.$$

As G is finite and $R^\times$ torsion-free and divisible $H^n_H(G, R^\times) = 1$. From this follows the statement. This means that the restriction to rays of (anti)unitary operators is not essential.

1 M. Hamermesh, *Group Theory* (Addison-Wesley, Reading, Mass., 1962).
2 A. Janner and T. Janssen, Physica **53**, 1 (1971).
3 I. Schur, J. Reine und Angewandte Math. **127**, 20 (1904); **132**, 85 (1907).
4 V. Bargmann, Ann. Math. **59**, 1 (1954).
5 E. Wigner, *Group Theory* (Academic, New York, 1959).
6 K. R. Parthasarathy, Commun. Math. Phys. **15**, 305 (1969).
7 M. V. Murthy, J. Math. Phys. **7**, 853 (1966).
8 M. Hall, *Group Theory* (MacMillan, New York, 1957).
9 S. McLane, *Homology* (Springer, Berlin, 1963).
10 T. Janssen, A. Janner, and E. Ascher, Physica **42**, 41 (1969).
11 W. Döring, Z. Naturf. **14a**, 343 (1959).
12 A. C. Hurley, Phil. Trans. Roy. Soc. **260**, 1 (1966).
13 S. Eilenberg and S. McLane, Ann. Math. **48**, 51 (1947).
14 C. J. Bradley and B. L. Davies, Rev. Mod. Phys. **40**, 359 (1968).
15 A. H. Clifford, Ann. Math. **38**, 533 (1938).
16 G. Koster, *Solid State Physics* (Academic, New York, 1957), Vol. 5.

Commuting Polynomials in Quantum Canonical Operators and Realizations of Lie Algebras

A. Joseph
Department of Physics and Astronomy, Tel-Aviv University, Ramat-Aviv, Israel

(Received 21 September 1970)

It is pointed out that the problem of realizing Lie algebras through polynomials in quantum canonical operators is not equivalent to its classical counterpart because the polynomial Lie algebras taken with respect to the classical and quantum Lie brackets are not isomorphic. Yet there are still many results which are common to both. To show this, the properties of commuting polynomials in quantum canonical operators are analyzed. This makes possible an extension from the classical to the quantum domain of a number of theorems on realizations of semisimple Lie algebras. At the same time it is stressed that differences can arise in the classical and quantum solutions, and some of these are described.

1. INTRODUCTION

The success of the group theoretical classification of the elementary particles has stimulated a reinvestigation of the dynamical symmetries of mechanical systems.[1,2] This presents the following problem. Given an arbitrary Lie algebra $\mathfrak{L}$, determine up to canonical equivalence, all possible expressions for the generators of $\mathfrak{L}$ as polynomials in a given number n of pairs of canonical operators. This problem can be considered in the context of either the Poisson bracket Lie algebra of classical mechanics[2-4] or the commutator bracket Lie algebra of quantum mechanics.[5,6] Now although a distinction is not always made between these two approaches, they may indeed admit quite different solutions. This arises because on the polynomials these Lie algebras are not isomorphic.[7] This fact can be ignored in the simplest cases which only involve subalgebras of the polynomials that do admit an isomorphism.[8] Yet it must be taken into consideration in general for it can lead to important differences in the dynamical symmetries of the two mechanical systems.

The solution to the above problem in the classical framework is greatly simplifed[4] through the use of certain existence theorems[9-11] in the theory of differential equations. It is our aim to develop a corresponding formalism which applies in the quantum domain. This is achieved through the study of com-

that both the restriction to A_1 and the induced automorphism of A_3 are the identity automorphism. Equivalent extensions determine factor systems differing by a 2-coboundary and vice versa. The classes of nonequivalent extensions are in one-to-one correspondence with the elements of $H^2_\phi(A_3, A_1)$. They form a group for that reason.

A *commutative diagram* is a set of groups connected by homomorphisms such that the composition of several such homomorphisms between two groups determines the same homomorphism between the groups. In particular for a diagram

$$\begin{array}{ccc} A & \xrightarrow{\rho} & B \\ {\scriptstyle\sigma}\downarrow & & \downarrow{\scriptstyle\tau} \\ C & \xrightarrow{\nu} & D \end{array}$$

commutativity means that $\tau\rho = \upsilon\sigma$.

APPENDIX B

In Sec. 1 we restricted ourselves to (anti)unitary operators on $\mathfrak{K}$. A ray of operators consists of (anti) unitary operators differing by a complex number of modulus one. However, the rays of $\mathfrak{K}$ consist of vectors differing by an arbitrary complex number. If we had defined operator rays in an analogous way we would have found an exact sequence:

$$1 \to C^\times \to P(\mathfrak{K}) \to \mathfrak{A}(\mathfrak{K}) \to 1,$$

where $C^\times$ is the multiplicative group of nonzero complex numbers and

$$P(\mathfrak{K}) = \{\lambda A \mid A \in \mathfrak{A}(\mathfrak{K}), \lambda \in C^\times\}.$$

Following the same lines the factor systems are then elements of $Z^2_H(G, C^\times)$ when $\alpha c = c (\alpha \in H), \alpha c = c^* (\alpha \in G\text{-}H)$ for any $c \in C^\times$ and classes of associated factor systems are elements of $H^2_H(G, C^\times)$. Now we have

$$H^2_H(G) \cong H^2_H(G, C^\times).$$

For the proof consider the exact sequence $1 \to U(1) \to C^\times \to R^\times \to 1$, where $R^\times$ is the multiplicative group of positive real numbers. From this follows the exact sequence

$$\cdots \to H^n_H(G, R^\times) \to H^{n+1}_H(G) \to H^{n+1}_H(G, C^\times) \to H^{n+1}_H(G, R^\times) \to \cdots.$$

As G is finite and $R^\times$ torsion-free and divisible $H^n_H(G, R^\times) = 1$. From this follows the statement. This means that the restriction to rays of (anti)unitary operators is not essential.

1 M. Hamermesh, *Group Theory* (Addison-Wesley, Reading, Mass., 1962).
2 A. Janner and T. Janssen, Physica **53**, 1 (1971).
3 I. Schur, J. Reine und Angewandte Math. **127**, 20 (1904); **132**, 85 (1907).
4 V. Bargmann, Ann. Math. **59**, 1 (1954).
5 E. Wigner, *Group Theory* (Academic, New York, 1959).
6 K. R. Parthasarathy, Commun. Math. Phys. **15**, 305 (1969).
7 M. V. Murthy, J. Math. Phys. **7**, 853 (1966).
8 M. Hall, *Group Theory* (MacMillan, New York, 1957).
9 S. McLane, *Homology* (Springer, Berlin, 1963).
10 T. Janssen, A. Janner, and E. Ascher, Physica **42**, 41 (1969).
11 W. Döring, Z. Naturf. **14a**, 343 (1959).
12 A. C. Hurley, Phil. Trans. Roy. Soc. **260**, 1 (1966).
13 S. Eilenberg and S. McLane, Ann. Math. **48**, 51 (1947).
14 C. J. Bradley and B. L. Davies, Rev. Mod. Phys. **40**, 359 (1968).
15 A. H. Clifford, Ann. Math. **38**, 533 (1938).
16 G. Koster, *Solid State Physics* (Academic, New York, 1957), Vol. 5.

Commuting Polynomials in Quantum Canonical Operators and Realizations of Lie Algebras

A. Joseph
Department of Physics and Astronomy, Tel-Aviv University, Ramat-Aviv, Israel

(Received 21 September 1970)

It is pointed out that the problem of realizing Lie algebras through polynomials in quantum canonical operators is not equivalent to its classical counterpart because the polynomial Lie algebras taken with respect to the classical and quantum Lie brackets are not isomorphic. Yet there are still many results which are common to both. To show this, the properties of commuting polynomials in quantum canonical operators are analyzed. This makes possible an extension from the classical to the quantum domain of a number of theorems on realizations of semisimple Lie algebras. At the same time it is stressed that differences can arise in the classical and quantum solutions, and some of these are described.

1. INTRODUCTION

The success of the group theoretical classification of the elementary particles has stimulated a reinvestigation of the dynamical symmetries of mechanical systems.[1,2] This presents the following problem. Given an arbitrary Lie algebra $\mathfrak{L}$, determine up to canonical equivalence, all possible expressions for the generators of $\mathfrak{L}$ as polynomials in a given number n of pairs of canonical operators. This problem can be considered in the context of either the Poisson bracket Lie algebra of classical mechanics[2-4] or the commutator bracket Lie algebra of quantum mechanics.[5,6] Now although a distinction is not always made between these two approaches, they may indeed admit quite different solutions. This arises because on the polynomials these Lie algebras are not isomorphic.[7] This fact can be ignored in the simplest cases which only involve subalgebras of the polynomials that do admit an isomorphism.[8] Yet it must be taken into consideration in general for it can lead to important differences in the dynamical symmetries of the two mechanical systems.

The solution to the above problem in the classical framework is greatly simplifed[4] through the use of certain existence theorems[9-11] in the theory of differential equations. It is our aim to develop a corresponding formalism which applies in the quantum domain. This is achieved through the study of com-

muting polynomials in the canonical operators. It enables us to extend to quantum mechanics some of the results previously obtained in classical mechanics and to point out some of the important differences that can arise. Though there is some overlap here with results obtained using the quotient division ring,[5] the present analysis is quite distinct from this and generally applies to a rather different domain. Indeed our approach is much closer to that given[4] in the classical problem, and for this reason we start with a discussion of the Poisson bracket.

2. COMMUTING POLYNOMIALS AND THE POISSON BRACKET

Let P denote the set of all polynomials over the complex field $\underline{C}$ in the real variables q, p. This forms a Lie algebra under the Poisson bracket defined by

$$\{f,g\} = \frac{\partial f}{\partial q}\frac{\partial g}{\partial p} - \frac{\partial f}{\partial p}\frac{\partial g}{\partial q}, \tag{2.1}$$

for all $f, g \in P$.

Suppose we are given a pair of elements $a, b \in P$ having vanishing Poisson bracket. Then inspection of (2.1) shows that[10] there exists a differentiable function F satisfying

$$F(a, b) = 0. \tag{2.2}$$

For an application of (2.2) in the realizations of Lie algebras we note the following proposition.

Proposition 2.1: Given $a, b, h \in P$ satisfying

$$\{a, b\} = 0, \tag{2.3a}$$

$$\{a, h\} = sa, \tag{2.3b}$$

$$\{b, h\} = tb, \tag{2.3c}$$

where s, t are positive integers, then to within an arbitrary complex constant

$$a^t = b^s. \tag{2.4}$$

Proof: Equation (2.3a) implies (2.2) which may be substituted into (2.3b), (2.3c) to yield

$$sa\frac{\partial F}{\partial a} + tb\frac{\partial F}{\partial b} = 0. \tag{2.5}$$

Integration of (2.5) gives (2.4) as required.

To see the significance of this result, we note that (2.4) implies the existence of an element $c \in P$ satisfying

$$a^u = c^s, \quad b^u = c^t \tag{2.6}$$

where u is the highest common divisor of s and t.

From (2.6) and (2.3) we obtain

$$\{c, h\} = uc \tag{2.7}$$

and conversely (2.6) and (2.7) imply (2.3). Hence the original problem of obtaining realizations of (2.3) in P has been simplified to that of obtaining realizations of (2.7) in P. Moreover, identifying a, b, c as eigenvectors belonging to the positive roots of the adjoint representation, we see that this analysis has a general context in the theory of semisimple Lie algebras. Indeed it is not hard to extend proposition (2.1) to verify the well-known result that a semisimple Lie algebra of rank ≥ 2 cannot be realized in P.[4] This argument may be further extended to the case of n degrees of freedom and may also be strengthened to apply to the space of infinitely differentiable functions.

We should clearly like to know to what extent the above arguments apply in the quantum framework. For this purpose we develop an analog to (2.2) for this case. This applies to the polynomials; but is not valid on formal power series. We use it to show that proposition (2.1) still holds for quantum canonical operators, though by contrast we find that the factorization of (2.4) to obtain c fails. We then extend these results to n degrees of freedom and apply them to the study of semisimple Lie algebras. For mathematical convenience we omit the imaginary number i from the canonical commutation relations and set Planck's constant equal to one. This does not affect the present analysis, though the necessary adjustments must be made if one wishes to speak of q and p as self-adjoint operators.

3. COMMUTING POLYNOMIALS AND THE COMMUTATOR BRACKET

Let $\mathcal{P}$ denote the set of all polynomials over $\underline{C}$ in the elements q, p and $\underline{1}$ which satisfy

$$qp - pq = \underline{1}. \tag{3.1}$$

Then $\mathcal{P}$ becomes a Lie algebra with respect to the commutator Lie bracket defined by

$$[f,g] = fg - gf, \tag{3.2}$$

for all $f, g \in \mathcal{P}$. We shall use $\mathcal{P}^n$ to denote the corresponding algebra generated by the elements satisfying

$$[q_i, p_j] = \delta_{ij}\underline{1}, \tag{3.3}$$

$i, j = 1, 2, \ldots, n$, where δ_{ij} is the Kronecker delta.

The general commutator in $\mathcal{P}$ may be computed through the formula

$$[q^m, p^n] = \sum_{k=1}^{(m,n)} \binom{m}{k}\binom{n}{k}(-1)^{k-1}k!\, q^{m-k}p^{n-k}, \tag{3.4}$$

in which m and n are arbitrary positive integers and (m, n) denotes the smaller of m and n. We remark that the corresponding expression in P is obtained from (3.4) by omitting the terms in $k > 1$. It is these higher terms which prevent $\mathcal{P}$ and P being isomorphic and cause the differences arising in the symmetries of classical and quantum mechanics.

Given $a \in \mathcal{P}^n$, we shall say that a is nontrivial if it is not equal to a constant multiple of the identity $\underline{1}$. Given nontrivial elements $a_1, a_2, \ldots, a_m \in \mathcal{P}$, we denote by $\mathcal{R}(a_1, a_2, \ldots, a_m)$ the set of all polynomials over $\underline{C}$ in $a_1, a_2, \ldots, a_m, \underline{1}$ ordered so that a_i appears to the left of a_j for all $i < j$. We shall say that $x \in \mathcal{R}(\cdots)$ is *nontrivial* if as a polynomial in the a_i it is not equal to a constant multiple of the identity.

For the case of one degree of freedom the main result of this section (which we have previously stated without proof[8]) may now be given. This is

Theorem 3.1: Given a, b nontrivial commuting elements of $\mathcal{P}$, then there exists a nontrivial element of $\mathcal{R}(a, b)$ which evaluated in $\mathcal{P}$ is identically zero.

Proof: We start by demonstrating some general properties of $\mathcal{P}$. These are contained in the following lemmas.

Lemma 3.1: Given $a, b \in \mathcal{P}$ with $x(a)$ and $y(b)$ nontrivial elements of $\mathcal{R}(a)$ and $\mathcal{R}(b)$, respectively, then $[x(a), y(b)] = 0$ implies $[a, b] = 0$.

Proof: It is sufficient to show that $[x(a), b] = 0$ implies $[a, b] = 0$. Since $\mathcal{R}$ is over the complex field, we may factor x in the form

$$x = \prod_{i=1}^{k} (a + \alpha_i \underline{1}),$$

with $\alpha_i \in \underline{C}$. From this the identity

$$[x(a), b] = \sum_{j=1}^{k} \left(\prod_{i=1}^{j-1} (a + \alpha_i \underline{1}) \right) [a, b] \left(\prod_{i=j}^{k} (a + \alpha_i \underline{1}) \right) \quad (3.5)$$

is readily checked. Let us suppose that a is a polynomial of degree r in p having coefficient in p^r the polynomial $f(q)$ in q. If $[a, b] \neq 0$, we may likewise suppose it to be a polynomial of degree s in p with $g(q)$ the coefficient of p^s. Then from (3.1) and (3.5) we have

$$(ad^{r(k-1)+s}q)[x(a), b] = (r!)^{k-1} s! f^{j-1} g f^{k-j}$$
$$= (r!)^{k-1} s! f^{k-1} g.$$

Since f and g are both polynomials, $[x(a), b] = 0$ implies that either f or g vanish. In either case we have a contradiction; hence $[a, b] = 0$ and the lemma is proved.

Lemma 3.2: Given a, b non-trivial commuting elements of $\mathcal{P}$. Let a be of degree r in q with $x(p)$ the coefficient of q^r and b of degree s in q with $y(p)$ the coefficient of q^s. Then there exists a nonzero complex constant α such that

$$x^s(p) = \alpha y^r(p). \quad (3.6)$$

Proof: Setting to zero the coefficient of q^{r+s-1} in $[a, b]$, we obtain, through (3.1), the differential equation

$$sy(p)\frac{dx(p)}{dp} = rx(p)\frac{dy(p)}{dp}. \quad (3.7)$$

Integration of (3.7) gives (3.6) as required.

Returning to the theorem we now show that there exists a nontrivial element $x \in \mathcal{R}(a, b)$ which is independent of q. To this end, suppose that a is of degree r in q and b is of degree s in q. If either r or s is zero, we need go no further. Otherwise, raise a to the power s and b to the power r. This gives two new elements a', b' of the same degree $t = rs$. We show that $x \in \mathcal{R}(a', b')$. Since $\mathcal{R}(a', b') \subset \mathcal{R}(a, b)$, we may drop the prime without loss of generality.

From Lemma 3.2 we see that any two commuting elements of the same degree in q have linearly dependent leading terms. On the other hand, the monomials in $\mathcal{R}(a, b)$ of degree kt or less in q span a space of dimension $(k + 1)(k + 2)/2$. Choosing $k \geq 2t$, we have $(k + 1)(k + 2)/2 > kt$, and so the required element x can be obtained as a suitable linear combination of these monomials. Should x not be independent of p, then it must be a nontrivial element of $\mathcal{R}(p)$. As a commutes with x, then through Lemma 3.1 it must also commute with p and so must a nontrivial element of $\mathcal{R}(p)$. As a and b commute, the same applies to b and the theorem follows through a repetition of the above argument to the degree of p.

We remark that this result applies to formal power series when one of the canonical operators appears at least linearly and at most to some finite power in both a and b. However, it fails on general power series. For example, from the Weyl form of the canonical commutation relations, it is easily shown that

$$[\exp(i\alpha q), \exp(2\pi p/\alpha)] = 0,$$

for all real α. Obviously there can be no polynomial (or function) of these operators which vanishes. Moreover, they exist in a well-defined sense if q and ip are self-adjoint. This result which does not hold for the Poisson bracket has been used to discuss quantum effects with no classical analog.[12] It also provides a useful set of commuting observables suitable for discussing periodic systems in quantum physics.[13]

As an application of Theorem 3.1 we extend Proposition 2.1 to $\mathcal{P}$. That is, we show:

Theorem 3.2: Given $a, b, h \in \mathcal{P}$ satisfying

$$[a, b] = 0, \quad (3.8a)$$

$$[a, h] = sa, \quad (3.8b)$$

$$[b, h] = tb, \quad (3.8c)$$

with s and t positive integers, then to within an arbitrary complex constant

$$a^t = b^s. \quad (3.9)$$

Proof: As a and b are both clearly nontrivial, Theorem 3.1 applies and we obtain an expression of the form

$$\sum_{k=0}^{m} \sum_{l=0}^{n} \alpha_{kl} a^k b^l = 0, \quad (3.10)$$

where the $\alpha_{kl} \in \underline{C}$ and not all these constants vanish. Commuting h through (3.10) r times and using (3.8b), (3.8c), we obtain

$$\sum_{k=0}^{m} \sum_{l=0}^{n} (sk + tl)^r \alpha_{kl} a^k b^l = 0.$$

This holds for any nonnegative integer r. Hence the double summation in (3.10) must reduce for each nonnegative integer value u of $sk + tl$ to u single summations which individually vanish. As not all the α_{kl} are zero, there is at least one value of u for which the coefficients in the summation do not all vanish. By formal multiplication of any one such expression by a suitable multiple of a^{-1}, we obtain a nontrivial polynomial in $c \equiv a^{-t} b^s$. Factorization of this polynomial over the complex field into terms linear in c gives $c = \gamma \underline{1}$ for some $\gamma \in \underline{C}$. Multiplication through by a^t gives (3.9) as required and the theorem is proved.

In $\mathcal{P}^n$ Theorem 3.1 admits the following generalization.

Theorem 3.3: Given $a_1, a_2, \ldots, a_{n+1}$ nontrivial commuting elements of $\mathcal{P}^n$, there exists a nontrivial element $z \in \mathfrak{R}(a_1, a_2, \ldots, a_{n+1})$ which evaluated in $\mathcal{P}^n$, is identically zero.

Proof: As in Theorem 3.1, we compare, for given degree, the number of distinct monomials in $\mathfrak{R}(a_1, a_2, \ldots, a_{n+1})$ with the number of commuting elements that are linearly independent in their leading term. The key step in this argument is contained in the following lemma.

Lemma 3.3: Let $S_{l,m,n}$ be the linear span of a collection of homogeneous polynomials of degree m in l independent variables $x_1, x_2, \ldots, x_l$ such that for each subset $f_1, f_2, \ldots, f_{n+1} \in S_{l,m,n}$ and at each point $\underline{x}_0$ of the $\underline{x}$-space, the differentials $df_i(\underline{x}_0)$, $i = 1, 2, \ldots, (n+1)$, are linearly dependent. Then $\dim S_{l,m,n} \leq \binom{m+n-1}{m}$ for all positive integer l, m, n.

Proof: The proof is by simultaneous induction on l and n. Since the number of homogeneous polynomials of degree m in l variables is $\binom{m+l-1}{m}$, it certainly holds for all $l \leq n$. In this case the condition on the differentials is trivial. However, for $l > n$, we have from

$$df_i(\underline{x}_0) = \sum_{j=1}^{l} \left(\frac{\partial f_i}{\partial x_j}\right)_{\underline{x}=\underline{x}_0} dx_j,$$

and the independence of the x_j, that any determinant of order $(n+1)$ of the matrix $(\partial f_i/\partial x_j)$ vanishes at each point $\underline{x}_0$ of the $\underline{x}$-space and hence everywhere.

For $n = 1$, this condition becomes

$$\frac{\partial f}{\partial x_i}\frac{\partial g}{\partial x_j} - \frac{\partial f}{\partial x_j}\frac{\partial g}{\partial x_i} = 0, \tag{3.11}$$

for all $f, g \in S_{l,m,1}$ and all $i, j = 1, 2, \ldots, l$. Multiplication of (3.11) by x_i and summation over i gives after a little rearrangement

$$\left(\sum_{i=1}^{l} x_i \frac{\partial \log f}{\partial x_i}\right)\frac{\partial \log g}{\partial x_j} - \left(\sum_{i=1}^{l} x_i \frac{\partial \log g}{\partial x_i}\right)\frac{\partial \log f}{\partial x_j} = 0.$$

Since f, g are homogeneous of degree m, the expressions in round brackets are both equal to m. Integration of the resulting equation over each x_j shows f and g to be linearly dependent and hence that $\dim S_{l,m,1} \leq 1$.

Let us assume that the lemma holds for all l, m, n satisfying $l \leq l_0$, $m \leq m_0$, $n \leq n_0$ and all l, m, n satisfying $l \leq l_0 + 1, m \leq m_0, n \leq n_0 - 1$. We argue by contradiction to show that it holds for $l = l_0 + 1$, $m = m_0, n = n_0$. Let x denote the $(l_0 + 1)$th variable. We show that $\dim S_{l_0+1,m_0,n_0} > \binom{m_0+n_0-1}{m_0}$ implies the existence of a nonzero element $g \in S_{l_0+1,m_0,n_0}$ which is divisible by x. Suppose there are no such elements; then each $f_i \in S_{l_0+1,m_0,n_0}$ may be written in the form

$$f_i = x^{m_0} f_{i,0} + x^{m_0-1} f_{i,1} + \cdots + f_{i,m_0},$$

where the $f_{i,r}$ are homogeneous polynomials of degree r in l_0 variables. Since $(\partial f_i/\partial x_j)_{x=0} = (\partial f_{i,m_0}/\partial x_j)$ for all $x_j \neq x$, the f_{i,m_0} may be regarded as elements of S_{l_0,m_0,n_0}. Then through the induction hypothesis and the assumed dimensionality of S_{l_0+1,m_0,n_0} we have $\dim S_{l_0+1,m_0,n_0} > \dim S_{l_0,m_0,n_0}$. Hence the required element g can be obtained as a suitable linear combination of the f_i.

We now use the existence of element g to show that the number of linearly independent elements contained in S_{l_0+1,m_0,n_0} and divisible by x is strictly greater than $\binom{m_0+n_0-2}{m_0-1}$. Since this is the maximum possible dimension of S_{l_0,m_0-1,n_0}, repetition of the above argument provides an element of S_{l_0+1,m_0,n_0} divisible by x^2. We can then show that these elements form a space of dimension strictly greater than $\binom{m_0+n_0-3}{m_0-2}$ and, continuing in this fashion, we eventually find that the dimension of the space of elements divisible by x^m is strictly greater than one. This contradiction will prove the lemma.

Consider the $(n+1) \times (n+1)$ matrix with entries $(\partial f_{i_r}/\partial x_{j_s})$ with $f_{i_1} = g$,

$$x_{i_1} = x,\ f_{i_r} \in S_{l_0+1,m_0,n_0},\ r = 2, 3, \ldots, (n+1),$$
$$1 \leq j_s \leq l_0,\ s = 2, 3, \ldots, (n+1).$$

Through the hypothesis of the lemma, $\det(\partial f_{i_r}/\partial x_{j_s}) = 0$. By identifying the smallest coefficient of x in this expression and using the fact that g is divisible by x and is a nonzero polynomial, we obtain

$$\det\left(\frac{\partial f_{i_r,m_0}}{\partial x_{j_s}}\right)' = 0, \tag{3.12}$$

where the prime denotes that the first column and first row of the original determinant has been deleted. Through Eq. (3.12), we may regard f_{i_r,m_0} as an element of S_{l_0+1,m_0,n_0-1}. By the induction hypothesis this space is of maximum dimension $\binom{m_0+n_0-2}{m_0}$. Hence the space of all elements of S_{l_0+1,m_0,n_0} divisible by x is strictly greater than

$$\binom{m_0+n_0-1}{m_0} - \binom{m_0+n_0-2}{m_0} = \binom{m_0+n_0-2}{m_0-1},$$

as we wished to show. The lemma is proved.

As in Theorem 3.1, we may assume without loss of generality that the a_i are all of the degree t. Then for given positive integer k the monomials in $\mathfrak{R}(a_1, a_2, \ldots, a_{n+1})$, which as polynomials in $\mathcal{P}^n$ are of degree kt, span a space of dimension $\binom{k+n}{n}$. Now we may compute the leading term of the commutator of any two elements in $\mathcal{P}^n$ by re-expressing the operators q_i, p_j as real variables followed by use of the Poisson bracket. We stress here that the leading term is independent of the ordering in q_i, p_j. It is then an easy consequence of a theorem[11] on commutative function groups that the hypothesis of Lemma 3.3 applies to these monomials and hence (setting $l = 2n, m = kt$) that their leading terms form a space of dimension $\binom{kt+n-1}{kt}$. Thus, to obtain the required element z satisfying the conclusions of the theorem, it suffices to find a positive integer k_0 such that

$$\sum_{k=0}^{k_0}\binom{k+n}{k} = \binom{k_0+n+1}{k_0} > \sum_{m=0}^{k_0 t}\binom{m+n-1}{m} = \binom{k_0 t+n}{k_0 t}.$$

This holds for all $k_0 \geq (n+1)t^n$, and hence the theorem is proved.

Should the a_i not all commute, then Theorem 3.3 fails. However, we note in passing the following result which holds for general elements of $\mathcal{P}^n$.

Theorem 3.4: Given $2n+1$ nontrivial elements $a_1, a_2, \ldots, a_{2n+1} \in \mathcal{P}^n$, there exists a nontrivial element $x \in \mathcal{R}(a_1, a_2, \ldots, a_{2n+1})$ which evaluated in $\mathcal{P}^n$ is identically zero.

Proof: The proof is by monomial counting. Indeed it is not difficult to show that there exists a positive integer k such that the number of distinct monomials in $\mathcal{R}(a_1, a_2, \ldots, a_{2n+1})$ exceeds the total number of ordered monomials in $q_1, q_2, \ldots, q_n, p_1, p_2, \ldots, p_n$ appearing in them. It should be noted that we use (3.3) to effect this letter ordering so that the above result depends on the particular form of these commutation relations and cannot generally be assumed to hold for arbitrary noncommuting polynomials.

Finally we remark that in analogy with the theory of function groups,[9] an argument parallel to that used in Theorem 3.3 establishes the following.

Theorem 3.5: Let m be a positive integer with $m \leq 2n$. Let a_i, $i = 1, 2, \ldots, m$, and b_j, $j = 1, 2, \ldots, 2n - m + 1$, be nontrivial elements of $\mathcal{P}^n$ with $[a_i, b_j] = 0$, for all i, j. Then either $\mathcal{R}(a_1, a_2, \ldots, a_m)$ or $\mathcal{R}(b_1, b_2, \ldots, b_{2n-m+1})$ contain a nontrivial element which evaluated in $\mathcal{P}^n$ is identically zero.

4. REALIZATIONS OF SEMISIMPLE LIE ALGEBRAS

Applying Theorem 3.3, we can now extend some of the results on realizations of semisimple Lie algebras taken under the Poisson bracket to the quantum domain. First we note a basic property of these algebras which is nearly always used in this connection. This is[14]

Theorem 4.1: A semisimple Lie algebras $\mathcal{L}$ of rank r admits a subalgebra $\mathcal{L}^+$ of dimension $2r$ with basis $h_1, h_2, \ldots, h_r, e_1, e_2, \ldots, e_r$ satisfying

$$[h_i, h_j] = 0, \tag{4.1a}$$

$$[e_i, e_j] = 0, \tag{4.1b}$$

$$[h_i, e_j] = \alpha_{ij} e_j \quad \text{(no summation)}. \tag{4.1c}$$

The α_{ij} are rational constants and form a matrix α for which

$$\det\alpha \neq 0. \tag{4.2}$$

From this result we can now prove:

Theorem 4.2: Given $\mathcal{L}$ a semisimple Lie algebra of rank r. Then, if $r > n$, it has no realizations in $\mathcal{P}^n$.

Proof: The proof proceeds by application of Theorem 3.3 to the commutation relations (4.1). This will give $\det\alpha = 0$, which by (4.2) is a contradiction. To this end, we first observe that on account of (4.1c) and (4.2) the e_i must be nontrivial elements of $\mathcal{P}^n$. Thus Theorem 3.3 applies, and we may assert for $r > n$ that there exists a nontrivial element $e \in \mathcal{R}(e_1, e_2, \ldots, e_r)$ satisfying

$$e = \sum_{k_1, k_2, \ldots, k_r = 0}^{m_1, m_2, m_r} \beta_{k_1 k_2 \ldots k_r} e_1^{k_1} e_2^{k_2} \ldots e_r^{k_r} = 0, \tag{4.3}$$

with $\beta_{k_1 k_2 \ldots k_r} \in \underline{C}$. Not all these constants vanish and, moreover, since the e_i are polynomials, at least two must be nonzero. We may also assume without loss of generality that none of the partial sums in (4.3) individually vanish.

For fixed i we commute h_i, using (4.1c), l times through (4.3). This gives

$$\sum \beta_{k_1 k_2 \ldots k_r} \left(\sum_{j=1}^{r} k_j \alpha_{ij} \right)^l e_1^{k_1} e_2^{k_2} \ldots e_r^{k_r} = 0, \tag{4.4}$$

which holds for any nonnegative integer l. Recalling our assumption on partial sums, it follows that the summand $\sum_{j=1}^{n} k_j \alpha_{ij}$ must be independent of those k_j which lead to nonzero $\beta_{k_1 \ldots k_r}$. As at least two of these constants are nonzero, we may find integers k_j', $j = 1, 2 \ldots r$, which are not all zero and such that

$$\sum_{j=1}^{r} k_j' \alpha_{ij} = 0.$$

This must hold for all i; hence $\det\alpha = 0$ and the theorem is proved.

For $r = n$ we cannot generally assume that $\mathcal{L}$ will have a realization in $\mathcal{P}^n$. On the other hand, given such a realization, it is an easy extension of the above argument to show the following.

Theorem 4.3: Given a realization in $\mathcal{P}^n$ of a semisimple Lie algebra $\mathcal{L}$ of rank n, then the invariants of the enveloping algebra $U\mathcal{L}$ are all constant multiples of the identity.

Proof: Select an arbitrary invariant element $I \in U\mathcal{L}$. If I is nontrivial, it follows from Theorem 3.3, (4.1b), and the invariance of I that there exists a nontrivial $e \in \mathcal{R}(e_1, e_2, \ldots, e_n, I)$ which is zero evaluated in $\mathcal{P}^n$. Commuting through the h_i, as in the previous theorem, using (4.1c) and the invariance of I, we can show that $\det\alpha = 0$. This contradicts (4.2) and hence I must be a multiple of the identity, and so the theorem is proved.

Since we have used only the fact that $I \in \mathcal{P}^n$, this result also applies to any invariant element of $\mathcal{L}$ in $\mathcal{P}^n$. We can show for a compact semisimple Lie group that it leads to certain realizations being excluded. In this we shall assume the q_i, $i = 1, 2, \ldots, n$, to be self-adjoint and the p_i, $i = 1, 2, \ldots, n$, to be skew-adjoint: a choice consistent with (3.3). Then

Theorem 4.4: Let G be a compact semisimple Lie group of rank n with Lie algebra $\mathcal{L}$. Then $\mathcal{L}$ has no realizations in $\mathcal{P}^n$ satisfying either one of the following conditions:

(1) The elements of $\mathcal{L}$ are either all self-adjoint or all skew-adjoint.

(2) The elements of $\mathcal{L}$ are all real polynomials.

Proof:

(1) Since G is compact and semisimple, we may pick a basis $x_1, x_2, \ldots, x_n \in \mathcal{L}$ such that

$$I_1 = \sum_{i=1}^{n} x_i^2$$

is an invariant element of $\mathcal{L}$ in $U\mathcal{L}$. Assume for the moment that $n = 3$ and let m be the degree of the highest term appearing in the x_i. Certainly $m \geq 1$, for otherwise $\mathcal{L}$ would be commutative. To highest order, we may write each x_i in the form

$$x_i = \sum_{k=0}^{m} \alpha_{ik} q^{m-k} p^k,$$

where the α_{ik} are complex numbers. When the x_i are self-adjoint, we have through adjointness properties of the canonical operators and (3.3) that

$$\alpha_{ik} = (-1)^k \bar{\alpha}_{ik}$$

for all i, k where the bar denotes complex conjugation.

Since, by Theorem 4.3, I_1 must be a multiple of the identity, we have, on equating to zero the terms of degree $2m$ in I_1, that

$$\sum_{i=1}^{3} \sum_{k=0}^{m} \sum_{k'=0}^{m} \alpha_{ik} \bar{\alpha}_{ik'} (-1)^{k'} x^{k+k'} = 0,$$

which holds as an identity in the arbitrary real variable x. By induction on k we may show from this that $\alpha_{ik} = 0$ for all i, k, which is a contradiction. Similar arguments apply when the x_i are skew-adjoint and for general values of n.

(2) By Theorem 4.3, I_1 must be a multiple of the identity. This clearly cannot hold for real polynomials in $\mathcal{P}^n$ and so the theorem is proved.

It should be apparent from the above proof that (1) and (2) need not be equivalent. For example, by choosing q and ip to be self-adjoint, then

$$x = \tfrac{1}{4} i(q^2 + p^2), \quad y = \tfrac{1}{4} i(q^2 - p^2), z = \tfrac{1}{4}(qp + pq)$$

are all skew-adjoint. They close on the noncompact Lie algebra $so(2, 1)$ rather than the compact form $so(3, R)$. This is still true if we omit the imaginary factor i to make these polynomials real. To obtain the compact form, we must violate both conditions (1) and (2).

The failure to find realizations satisfying (1) above can be understood in terms of the representation theory. Thus (1) by suitable exponentiation leads to a unitary representation $\mathfrak{D}$ of G which by Theorem 4.3 and the remark following it we should expect to be irreducible. (Actually $\mathfrak{D}$ is a factor representation which could conceivably be of infinite multiplicity.)[15] In the case that $\mathfrak{D}$ is irreducible, then, as G is compact, it must also be finite dimensional, a result which conflicts with the infinite dimensionality of representations of the canonical commutation relations.

We close this section with one more result in the same spirit. It is a comment on canonical transformations in $\mathcal{P}^n$.

Theorem 4.5: Given a realization in $\mathcal{P}^n$ of the nilpotent Lie algebra $\mathcal{L}$ with elements $\hat{\underline{1}}, \hat{q}_i, \hat{p}_j$, $i, j = 1, 2 \ldots, n$, satisfying

$$[\hat{q}_i, \hat{p}_j] = \delta_{ij} \hat{\underline{1}}, \tag{4.5}$$

with $\hat{\underline{1}}$ in the center of $\mathcal{L}$. Then, if $\hat{\underline{1}} \neq 0$, it is a constant multiple of the identity.

Proof: If $\hat{\underline{1}} \neq 0$, then from (4.5) the q_i are nontrivial. If $\hat{\underline{1}}$ is nontrivial, then by Theorem 3.3 there exists a nontrivial element $e \in \mathcal{R}(q_1, q_2, \ldots, q_n, \underline{1})$ which vanishes in $\mathcal{P}^n$. By expanding e as in (4.3) and commuting the p_j through this expression using (4.5), it is easy to show that e is trivial in $\mathcal{R}(\ldots)$. This contradiction proves the theorem.

5. FACTORIZATION OF COMMUTING POLYNOMIALS

Though P^n and $\mathcal{P}^n$ are not isomorphic, there are, as we have seen, many results which hold for both. This cannot always be assumed to be true, and it is our present purpose to point out one such difference. This concerns the factorization of commuting elements of $\mathcal{P}$. Its importance derives from Proposition 2.1. Recall that here the factorization of (2.4) yielded the element c which effected the reduction of (2.3) to (2.7). We show by counterexample that the corresponding factorization cannot generally be carried out in $\mathcal{P}$. However, before we do this, we consider a possible strengthening of Theorem 3.1 which would imply this factorization in order to reveal the precise manner in which this failure occurs.

Suppose we could show that, given any two commuting elements $a, b \in \mathcal{P}$, there exists a polynomial $d \in \mathcal{P}$ such that $a, b \in \mathcal{R}(d)$. Then this would imply the required factorization. To show it is false, we attempt the construction and observe how this fails. To this end we first prove:

Theorem 5.1: Let a, b be nontrivial commuting elements of $\mathcal{P}$ with a of degree m and b of degree n. Let r be the highest common factor of m and n. Set $u = m/r, v = n/r$. Then there exists an element $e \in \mathcal{P}$ such that to highest order

$$a = e^u, \quad b = e^v. \tag{5.1}$$

Remark: It is clear that if d were to exist, then e would be its leading term.

Proof: The proof is by construction. Let a', b' be the leading terms of a, b, respectively. We may write them in the form

$$a' = \sum_{k=0}^{m} \alpha_k q^k p^{m-k}, \quad b' = \sum_{l=0}^{n} \beta_l q^l p^{n-l},$$

with $\alpha_k, \beta_l \in C$. We now compute the leading term in the commutant of a', b'.

Since a, b commute, this must vanish and so we obtain

$$\sum_{k=0}^{m} \sum_{l=0}^{n} \alpha_k \beta_l (km - ln) q^{k+l-1} p^{m+n-l-k-1} = 0,$$

which holds as an identity in the canonical operators q and p. This may be conveniently replaced by the following identity in the arbitrary real variable x, namely

$$\sum_{k=0}^{m} \sum_{l=0}^{n} \alpha_k \beta_l (km - ln) x^{k+l-1} = 0. \tag{5.2}$$

Define the polynomials

$$y_1 = \sum_{k=0}^{m} \alpha_k x^k, \quad y_2 = \sum_{k=0}^{n} \beta_l x^l,$$

and substitute y_1, y_2 and their derivatives y_1', y_2'

into (5.2). The resulting differential equation gives on integration

$$y_1^n = y_2^m, \tag{5.3}$$

to within an arbitrary multiplicative complex constant. Since we have now eliminated the noncommutativity from the problem, Eq. (5.3) may be factored to yield a polynomial y_3 satisfying

$$y_1 = y_3^u, \quad y_2 = y_3^v.$$

We may write y_3 in the form

$$y_3 = \sum_{k=0}^{r} \gamma_k x^k,$$

with $\gamma_k \in \underline{C}$. It is clear that if we set

$$e = \sum_{k=0}^{r} \gamma_k q^k p^{r-k},$$

then this element has the required property and the theorem is proved.

From the fact that we have used only the leading term in the expansion of the commutator $[a, b]$, the above theorem has the following immediate corollary:

Corollary 5.1: Given $a, b, c \in \mathcal{P}$ with $[a, b] = c$. Let m, n, p be the degrees of a, b, c, respectively, and define u, v as before. Then, if $p < m + n - 2$, there exists an element $e \in \mathcal{P}$ satisfying (5.1) to highest order.

This result proves useful in the study of realizations of Lie algebras in terms of high-order polynomials.

Returning to the construction of the element d, we set u, v of Theorem 5.1 equal to 3, 2, respectively, and write

$$a = e^3 + (3/4)(ef + fe), \quad b = e^2 + f, \tag{5.4}$$

with $f \in \mathcal{P}$. With this choice, which is not of course the most general one, the higher-order terms in the commutator $[a, b]$ cancel, and we obtain

$$[a, b] = [e, \{[e, [e, f]] + 3f^2\}] = 0. \tag{5.5}$$

Equation (5.5) will certainly hold if the term in curly brackets is made to vanish. This can be effected by setting

$$e = q^2 p, \quad f = -2q^2.$$

Moreover, with this choice, an elementary, though tedious calculation, which we omit, shows that

$$a^2 = b^3. \tag{5.6}$$

This provides our counterexample. Indeed, noting that e and f are not themselves polynomials in a common element and that deg $f <$ deg e, we see that the required element d cannot exist. By contrast the classical analog of the expression $[a, b]$ does not contain the double commutator $[e, [e, f]]$ appearing in (5.5). Consequently, in this case, e and f commute and are themselves functionally dependent. A corresponding role is played by the double commutator in the verification of (5.6), which is not mirrored by the classical expressions for a and b. We may also anticipate a similar behavior present in realizations of semisimple Lie algebras using high-order polynomials, though this is more difficult to exhibit. As such it would provide realizations with the property that the generators have different Lie algebraic relations with respect to the Poisson bracket. The possibility of this circumstance is a pointer to nonclassical behavior and merits further investigation.

Finally we remark that (5.6) factorizes trivially in the quotient division ring, a property which distinguishes it from $\mathcal{P}$.

ACKNOWLEDGMENTS

This work was supported by a Royal Society-Israel Academy Research Fellowship held at Tel-Aviv University. I should like to thank Professor Y. Ne'eman and the other members of the Physics Department for their hospitality and interest.

[1] M. Pauri and G. M. Prosperi, J. Math. Phys. 7, 366 (1966); D. M. Fradkin, Progr. Theoret. Phys. 37, 798 (1967); N. Mukunda, Phys. Rev. 155, 1383 (1967).
[2] F. Diumio and M. Pauri, Nuovo Cimento 51A, 1141 (1967).
[3] N. Mukunda, J. Math. Phys. 8, 1069 (1967); P. Chand, C. F. Mehta, N. Mukunda, and E. C. G. Sudarshan, J. Math. Phys. 8, 2048 (1967); J. Rosen, Nuovo Cimento 49A, 614 (1967).
[4] A. Simoni, B. Vitale and F. Zaccaria, Nuovo Cimento 51A, 448 (1967); F. Bunchaft, *ibid.* 57A, 689 (1968); P. B. Guest, *ibid.* 61A, 593 (1969).
[5] A. Simoni and F. Zaccaria, Nuovo Cimento 59A, 280 (1969).
[6] Y. Chow, J. Math. Phys. 10, 975 (1969).
[7] J. R. Shewell, Am. J. Phys. 27, 16 (1959) and references therein.
[8] A. Joseph, Commun. Math. Phys. 17, 210 (1970).
[9] L. P. Eisenhart, *Continuous Groups of Transformations* (Dover, New York, 1961), Chap. VI, pp. 281–92.
[10] R. P. Gillepsie, *Partial Differential Equations* (Interscience, New York, 1951), Chap. II, §10, pp. 43–46.
[11] R. Abraham, *Foundations of Mechanics* (Benjamin, New York, 1967), Chap. III, p. 112, Theorem 16.29.
[12] Y. Aharanov, H. Pendleton, and A. Petersen, Intern. J. Theoret. Phys. 2, 213 (1969).
[13] J. Zac, Phys. Rev. 168, 686 (1968); 177, 1151 (1969).
[14] B. Kostant, Amer. J. Math. 81, 973 (1959). Theorem 8.1 contains this result.
[15] I am indebted to Professor R. Haag for this remark.

Stochastic Master Equations: A Perturbative Approach*

Ioannis M. Besieris
Department of Mathematics, New York University, University Heights, Bronx, New York 10453
(Received 11 October 1971)

Linear stochastic master equations for wave propagation in a continuous random medium are derived along the lines of the resolvent theory used in nonequilibrium statistical mechanics. Equations for the mean and fluctuating fields are subsequently obtained by operating directly on the stochastic master equations with statistical projection operators. The findings are compared with the results determined using the method of renormalization and the method of smoothing.

1. INTRODUCTION

The study of wave propagation in a randomly inhomogeneous medium leads, in general, to a family of linear partial differential equations

$$L(\omega)u(\omega) = f(\omega). \tag{1.1}$$

Here, $L(\omega)$ is a stochastic operator depending on a parameter $\omega \in \Omega$, Ω being a probability measure space. In addition, $f(\omega)$ is a random source distribution and $u(\omega)$, the random field quantity, is an element of an infinitely dimensional vector space $\mathcal{K}$. $u(\omega)$ and $f(\omega)$ can be either scalar or vector quantities.[1]

The operator L is now split into two parts as follows:

$$L = L_0 - L_1. \tag{1.2}$$

L_0 and L_1 are linear operators in $\mathcal{K}$ corresponding, respectively, to "free" propagation and "interaction" propagation.

The field function u is, in turn, decomposed abstractly into two mutually independent terms:

$$u = Vu + Cu \tag{1.3}$$

by means of the formal introduction of the two operators V and C.[2] Vu is called the *mean* or *coherent* component and Cu is the *fluctuating* or *incoherent* component of the field function u.[3] The uniqueness of the decomposition (1.3) as well as the mutual independence of the two components are ensured by prescribing the properties

$$V + C = I, \quad V^2 = V, \quad C^2 = C, \quad VC = 0, \quad CV = 0, \tag{1.4}$$

where I is the identity operator. By virtue of these relations, V and C are called projection operators.

The interconnection between the decompositions (1.2) and (1.3) is contained in the commutation relations[4]:

$$[L_0, V]_- = 0, \quad [L_0, C]_- = 0 \tag{1.5}$$

which constitute a mathematical statement of the fact that the fluctuating component in (1.3) is due to the interaction part of the operator in (1.2) alone, not to the free propagation part of it. Therefore, L_0 must commute with V, and, thus, also with $C = I - V$. From (1.5) and (1.4) follow the relationships

$$CL_0V = 0, \quad VL_0C = 0. \tag{1.6}$$

A specific realization of the projection operators V and C which will be used in the ensuing work is the following:

$$V: u \to \mathcal{E}\{u\}, \quad C: u \to \delta u. \tag{1.7}$$

The ensemble average of the random field function $u(\omega)$ is given by

$$\mathcal{E}\{u(\omega)\} = \int_\Omega u(\omega)p(\omega)d\omega, \tag{1.8}$$

with $p(\omega)$ denoting the probability density in the measure space Ω.

The mean part of the field is identified with the ensemble average, and the fluctuating part is defined as

$$\delta u = u - \mathcal{E}\{u\}. \tag{1.9}$$

From this definition it follows that $\mathcal{E}\{\delta u\} = 0$.

Thus, any field $u \in \mathcal{K}$ can be written as a sum of the mean field $\mathcal{E}\{u\}$ and a fluctuating field δu.[5] Within the framework of the specific realization (1.7), the commutation relations (1.5) signify that L_0 is a nonrandom (deterministic) operator, and L_1 is a generally noncentered random operator.

2. SMALL PERTURBATION THEORY

There exists a well-known solution of the equation

$$L_0u - L_1u = f \tag{2.1}$$

in the form of a power series in the operator L_0^{-1}; it is obtained from the equation

$$u = L_0^{-1}(1 - L_1L_0^{-1})^{-1}f \tag{2.2}$$

with the aid of the binomial expansion theorem[6]:

$$u = \sum_{n=0}^{\infty} L_0^{-1}(L_1L_0^{-1})^n f. \tag{2.3}$$

This is the frequently used perturbation solution of (2.1). The expression for the coherent field $\mathcal{E}\{u\}$ may be found by applying the projection operator V (statistical average) on both sides of (2.3). Similarly, for the fluctuating field δu, one operates with the projection operator C. It is known, however, that the individual terms of the perturbational series for $\mathcal{E}\{u\}$ are secular, i.e., they increase without bound for large distances. To circumvent these restrictions one resorts to judicious partial summations of the infinite perturbational series solution. Selective partial summations of the infinite perturbational series have been effected most systematically by the method of *renormalization* introduced in the study of wave propagation in continuous random media by Tatarskii and Gertsenshtein,[7] and the method of *smoothing* expounded by Frisch.[8] One should also consider the equivalent (but more specialized) techniques developed by Meecham,[9] Bourret,[10] and Keller.[11]

In the following we shall develop another such technique along the lines of the resolvent theory used in nonequilibrium statistical mechanics (cf. Ref. 2) and compare the findings with those obtained using the aforementioned methods.

3. STOCHASTIC MASTER EQUATIONS

Equation (2.3) is rewritten by inserting between each factor L_1 and L_0^{-1} within the parenthesis the identity operator in the form

$$L_1 L_0^{-1} = L_1(V + C)L_0^{-1}. \tag{3.1}$$

With (3.1) taken into consideration, the sum (2.3) is subsequently rearranged by cutting every product at the point where the projection operator V appears, and grouping together all possible products between successive V factors. Finally, one sums over all possible number of V factors. The result is established to be

$$u = \sum_{n=0}^{\infty} L_0^{-1}(\mathfrak{M}VL_0^{-1})^n(1 + \mathfrak{M}CL_0^{-1})f, \tag{3.2}$$

with the operator $\mathfrak{M}$ defined as

$$\mathfrak{M} = \sum_{p=0}^{\infty} L_1(CL_0^{-1}L_1)^p. \tag{3.3}$$

This operator plays a significant role in the development of the theory.

Before proceeding further, it will be assumed that f is a deterministic source. Then, making use of the commutation relation $[L_0^{-1}, C]_- = 0$ and the fact that $CV = 0$ [cf. Eq. (1.4)], one finds that $\mathfrak{M}CL_0^{-1}f = 0$. Hence, (3.2) simplifies to

$$u = \sum_{n=0}^{\infty} L_0^{-1}(\mathfrak{M}VL_0^{-1})^n f. \tag{3.4}$$

This expression may be rewritten as

$$L_0 u - \mathfrak{M}Vu = f. \tag{3.5}$$

The relations (3.4) and (3.5) will be referred to as the *stochastic master equations*.[12] The first is an integral expression while the second is an integro-differential equation.

One can obtain a relation for the mean field by operating on (3.5) with the projection operator V:

$$\mathcal{E}\{u\} = L_0^{-1}f + L_0^{-1}M\mathcal{E}\{u\}. \tag{3.6}$$

By analogy to a similar equation in quantum electrodynamics, (3.6) is called the *Dyson–Schwinger equation* with

$$M = V\mathfrak{M}V = \sum_{p=1}^{\infty} VL_1(L_0^{-1}CL_1)^p V \tag{3.7}$$

the *mass operator*. (For the sake of simplicity, we have imposed the restriction that L_1 is a centered random operator. This condition is stated mathematically as $VL_1V = 0$.) Equation (3.6), with the mass operator given by (3.7), is identically the result reached by either the method of renormalization or the method of smoothing.

The *first-order smoothing approximation*[13] is determined by retaining only the first term in the series expansion for the mass operator, and introducing it into the Dyson–Schwinger equation, viz.,

$$M \simeq VL_1L_0^{-1}CL_1V, \tag{3.8a}$$

$$\mathcal{E}\{u\} = L_0^{-1}f + L_0^{-1}VL_1L_0^{-1}CL_1V\mathcal{E}\{u\}. \tag{3.8b}$$

It should be noted that the exact Dyson–Schwinger equation (3.6) corresponds to the formal summation

$$\mathcal{E}\{u\} = \sum_{n=0}^{\infty} (L_0^{-1}VM)^n L_0^{-1}f. \tag{3.9}$$

The same result can be found by applying the projection operator V on the integral expression (3.4). Also, (3.8b) corresponds to the formal summation

$$\mathcal{E}\{u\} = \sum_{n=0}^{\infty} (L_0^{-1}VL_1L_0^{-1}L_1V)^n L_0^{-1}f. \tag{3.10}$$

which, in turn, can be obtained from (3.9) by retaining the first nonvanishing term in the mass operator.

Finally, the fluctuating field is given in terms of the mean field, as seen by operating on (3.5) with the projection operator C:

$$\delta u = L_0^{-1}C\mathfrak{M}\mathcal{E}\{u\} = \sum_{p=1}^{\infty} (L_0^{-1}CL_1)^p \mathcal{E}\{u\}. \tag{3.11}$$

This is exactly the result obtained from the method of renormalization or the method of smoothing.

4. CONCLUDING REMARKS

In the method of renormalization and the method of smoothing one derives first a closed system of equations for the mean field and the fluctuating field. This system is then solved by the method of successive substitutions. Essentially, one iterates the field equation in the space $C\mathcal{K}$ to obtain the fluctuating field in terms of the mean field, and hence an equation for the sole mean field.

In contradistinction, in this exposition, one iterates the field equation in the entire space $\mathcal{K}$ [cf. Eq. (3.2)], and the mean and fluctuating components of the field are obtained by projecting the result on $V\mathcal{K}$ [cf. Eq. (3.9)] and $C\mathcal{K}$ [cf. Eq. (3.11)], respectively. Finite-order iterations are equivalent to the summation of infinite subseries, e.g., Eq. (3.10).

In closing, it should be pointed out that lifting the restrictions that the forcing function f be deterministic and L_1 a centered random operator introduces no essential difficulties.

---

* The research for this paper was supported by the Air Force Office of Scientific Research under Grant No. AFOSR-71-2107.

1 The parameter ω will usually be suppressed for convenience.

2 R. Balescu, Physica **38**, 98 (1968); **42**, 464 (1969).

3 *Vu* and *Cu* in Balescu's application to nonequilibrium statistical mechanics refer to the *vacuum* and *correlation* components of the field u, respectively; hence the notation for the operators V and C.

4 $[A, B]_- \equiv AB - BA$.

5 In addition to the inherent mathematical convenience, the resolution of the field into a coherent and incoherent part is in accordance with conventional measurement procedures.

6 It has been assumed here that the norm of the operator $L_0^{-1}L_1$ is less than unity ($\| L_0^{-1}L \| < 1$) so that the binomial theorem can be used.

7 V. I. Tatarskii and M. E. Gertsenshtein, Zh. Eksp. Teor. Fiz. **44**, 676 (1963) [Sov. Phys. JETP **17**, 458 (1963)].

8 U. Frisch, in *Probabilistic Methods in Applied Mathematics*, Vol. 1, edited by A. T. Bharucha-Reid (Academic, New York, 1968).

9 W. C. Meecham, Space Technology Laboratories Rept. BSD-TR-61-36, Los Angeles, California, 1961 (unpublished).

10 R. C. Bourret, Nuovo Cimento **26**, 1 (1962); Can. J. Phys. **40**, 782 (1962).

11 J. B. Keller, Proc. Symp. Appl. Math. **13**, 227 (1962); **16**, 145 (1964).

[12] This term is used here by virtue of the analogy of wave propagation in continuous random media and nonequilibrium statistical mechanics. As it has been pointed out by Frisch (cf. Ref. 8) this analogy arises principally as a consequence of the mathematical similarity of the linear random wave equation (1.1) and the linear Liouville equation.

[13] This terminology is due to Frisch (cf. Ref. 8). In diagrammatic form, this approximation was first introduced by Bourret (cf. Ref. 10) under the name of *one-fiction approximation*. Tatarskii and Gertsenshtein (cf. Ref. 7) refer to it as the *bilocal approximation*.

On Wave Propagation In Inhomogeneous Media

K. M. Case
The Rockefeller University, New York, New York 10021
(Received 17 September 1971)

Familiar relations between phase and group velocity for wave propagation in homogeneous media are generalized to the inhomogeneous case. The constant velocity "c" is merely replaced by an appropriate weighted average. The key tool lies in the stationary property of the frequency as a functional of the wavefunction.

1. INTRODUCTION

In many areas one encounters the phenomenon of "waveguide propagation". By this we will mean waves traveling in a medium such that the medium is homogeneous in the direction of propagation but inhomogeneous in the perpendicular direction.

In simple cases of homogeneous media it is well known that there are relations between the group velocity (v_g) and the phase velocity (v_p). Thus, for example,

$$v_p \geq v_g, \tag{1}$$

and

$$v_p v_g = c^2, \tag{2}$$

where c is the velocity occurring in the wave equation. Here we will show that there are simple generalizations of Eqs. (1) and (2) which hold for "waveguide propagation." While there are many variations of the waveguide problem,[1] we will for purposes of clarity restrict attention to only one such—and, indeed, one of the simplest.

2. FORMULATION

Consider the following generalization of the simple wave equation

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = 0, \tag{3}$$

where now c is independent of z but may be a general real function of the perpendicular coordinate $\mathbf{r}_p$. We look for solutions

$$\Phi = e^{i(kz-\omega t)}\Psi(\mathbf{r}_p). \tag{4}$$

Here $\Psi(\mathbf{r}_p)$ is to be such that Eq. (3) and prescribed boundary conditions are satisfied.

Thus

$$(\nabla_p^2 - k^2)\Psi = -\omega^2\Psi. \tag{5}$$

We note these conditions are equivalent to the requirement that the expression

$$\omega^2 = \frac{\int\{|\nabla_p\Psi|^2 + k^2|\Psi|^2\}d^2r_p}{\int|\Psi|^2/c^2(r_p)d^2r_p} \tag{6}$$

is stationary with respect to variations of Ψ.

3. RESULTS AND PROOFS

Let

$$\overline{1/c^2} = \frac{\int|\Psi|^2/c^2(r_p)d^2r_p}{\int|\Psi|^2d^2r_p} \tag{7}$$

then with

$$v_p = \frac{\omega}{k} \qquad v_g = \frac{d\omega}{dk}$$

we have

$$v_p v_g = 1/\overline{1/c^2} \tag{8}$$

and

$$v_p \geq \sqrt{1/\overline{1/c^2}}, \tag{9}$$

$$v_g \leq \sqrt{1/\overline{1/c^2}} \tag{10}$$

Proofs: Since the expression in Eq. (6) is stationary with respect to small variations in Ψ, we may find $d\omega/dk$ by differentiating Eq. (6) only where k occurs explicitly. Thus

$$2\omega\frac{d\omega}{dk} = 2k\frac{\int|\Psi|^2d^2r_p}{\int|\Psi|^2/c^2(r_p)d^2r_p} = 2k/\overline{1/c^2} \tag{11}$$

Therefore, dividing by $2k$, we obtain Eq. (8).

Further, since $|\nabla_p\Psi|^2$ is positive definite, we see that

$$\omega^2 \geq \frac{k^2\int|\Psi|^2d^2r_p}{\int|\Psi|^2d^2r_p/c^2(r_p)},$$

from which Eq. (9) follows. Finally, combining Eqs. (8) and (9), we get Eq. (10).

4. CONCLUSIONS

It has been shown that the relations between phase and group velocities are generalizable in inhomogeneous situations. The proof is in some sense simpler than the conventional one of using explicit expressions to verify the relations. The key property is the stationary property of ω^2 as a functional of the wavefunction.

[1] See, for example, L. M. Brekhovskikh, *Waves in Layered Media* (Academic, New York, 1960).

[12] This term is used here by virtue of the analogy of wave propagation in continuous random media and nonequilibrium statistical mechanics. As it has been pointed out by Frisch (cf. Ref. 8) this analogy arises principally as a consequence of the mathematical similarity of the linear random wave equation (1.1) and the linear Liouville equation.

[13] This terminology is due to Frisch (cf. Ref. 8). In diagrammatic form, this approximation was first introduced by Bourret (cf. Ref. 10) under the name of *one-fiction approximation*. Tatarskii and Gertsenshtein (cf. Ref. 7) refer to it as the *bilocal approximation*.

On Wave Propagation In Inhomogeneous Media

K. M. Case
The Rockefeller University, New York, New York 10021
(Received 17 September 1971)

Familiar relations between phase and group velocity for wave propagation in homogeneous media are generalized to the inhomogeneous case. The constant velocity "c" is merely replaced by an appropriate weighted average. The key tool lies in the stationary property of the frequency as a functional of the wavefunction.

1. INTRODUCTION

In many areas one encounters the phenomenon of "waveguide propagation". By this we will mean waves traveling in a medium such that the medium is homogeneous in the direction of propagation but inhomogeneous in the perpendicular direction.

In simple cases of homogeneous media it is well known that there are relations between the group velocity (v_g) and the phase velocity (v_p). Thus, for example,

$$v_p \geq v_g, \tag{1}$$

and

$$v_p v_g = c^2, \tag{2}$$

where c is the velocity occurring in the wave equation. Here we will show that there are simple generalizations of Eqs. (1) and (2) which hold for "waveguide propagation." While there are many variations of the waveguide problem,[1] we will for purposes of clarity restrict attention to only one such—and, indeed, one of the simplest.

2. FORMULATION

Consider the following generalization of the simple wave equation

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = 0, \tag{3}$$

where now c is independent of z but may be a general real function of the perpendicular coordinate $\mathbf{r}_p$. We look for solutions

$$\Phi = e^{i(kz-\omega t)}\Psi(\mathbf{r}_p). \tag{4}$$

Here $\Psi(\mathbf{r}_p)$ is to be such that Eq. (3) and prescribed boundary conditions are satisfied.

Thus

$$(\nabla_p^2 - k^2)\Psi = -\omega^2\Psi. \tag{5}$$

We note these conditions are equivalent to the requirement that the expression

$$\omega^2 = \frac{\int\{|\nabla_p\Psi|^2 + k^2|\Psi|^2\}d^2r_p}{\int|\Psi|^2/c^2(r_p)d^2r_p} \tag{6}$$

is stationary with respect to variations of Ψ.

3. RESULTS AND PROOFS

Let

$$\overline{1/c^2} = \frac{\int|\Psi|^2/c^2(r_p)d^2r_p}{\int|\Psi|^2d^2r_p} \tag{7}$$

then with

$$v_p = \frac{\omega}{k} \qquad v_g = \frac{d\omega}{dk}$$

we have

$$v_p v_g = 1/\overline{1/c^2} \tag{8}$$

and

$$v_p \geq \sqrt{1/\overline{1/c^2}}, \tag{9}$$

$$v_g \leq \sqrt{1/\overline{1/c^2}} \tag{10}$$

Proofs: Since the expression in Eq. (6) is stationary with respect to small variations in Ψ, we may find $d\omega/dk$ by differentiating Eq. (6) only where k occurs explicitly. Thus

$$2\omega\frac{d\omega}{dk} = 2k\frac{\int|\Psi|^2d^2r_p}{\int|\Psi|^2/c^2(r_p)d^2r_p} = 2k/\overline{1/c^2} \tag{11}$$

Therefore, dividing by $2k$, we obtain Eq. (8).

Further, since $|\nabla_p\Psi|^2$ is positive definite, we see that

$$\omega^2 \geq \frac{k^2\int|\Psi|^2d^2r_p}{\int|\Psi|^2d^2r_p/c^2(r_p)},$$

from which Eq. (9) follows. Finally, combining Eqs. (8) and (9), we get Eq. (10).

4. CONCLUSIONS

It has been shown that the relations between phase and group velocities are generalizable in inhomogeneous situations. The proof is in some sense simpler than the conventional one of using explicit expressions to verify the relations. The key property is the stationary property of ω^2 as a functional of the wavefunction.

[1] See, for example, L. M. Brekhovskikh, *Waves in Layered Media* (Academic, New York, 1960).

On the Interaction of the Electromagnetic Field with Heat Conducting Deformable Insulators

H. F. Tiersten and C. F. Tsai
Mechanics Division, Rensselaer Polytechnic Institute, Troy, New York 12181
(Received 16 July 1971; Revised Manuscript Received 18 October 1971)

The differential equations and boundary conditions describing the behavior of a finitely deformable, polarizable, and magnetizable, heat conducting continuum in interaction with the electromagnetic field are derived by means of a systematic application of the laws of continuum physics to a well-defined macroscopic model. The model consists of an electronic charge and spin continuum coupled to a lattice continuum, which in itself consists of two interpenetrating ionic continua, which can displace with respect to each other to produce ionic polarization. Since spin angular momentum and electronic and ionic linear momentum are taken into account, magnetic spin resonance and both ionic and electronic polarization resonances are included in the treatment. Magnetic interaction terms are obtained by regarding magnetization as a consequence of point circulating current densities. When material resonances are suppressed, a simpler model is applicable and a not only smaller but somewhat different system of equations turns out to be convenient.

1. INTRODUCTION

In recent years a number of workers have obtained reasonably consistent descriptions of the interactions of the electric, magnetic, and electromagnetic fields with deformable continua, beginning with the work of Toupin[1] in 1956, in which he derived a properly invariant description of static electroelasticity from a variational principle. Subsequently, Eringen[2] obtained a similar consistent description of static electroelasticity from a somewhat different variational principle. Some time later Brown,[3] and Tiersten[4,5] presented essentially equivalent rotationally invariant descriptions of magnetoelasticity. The former author[3] employed a variational principle to treat the static case, while the latter[4] introduced a continuum model and employed the notion of the quasistatic magnetic field to treat the dynamic case in the presence of heat conduction, linear mechanical viscosity, and magnetic dissipation. Shortly thereafter Eastman[6] reported on a wave velocity experiment, using a static biasing magnetic field, in agreement with the linear limit of the rotationally invariant nonlinear description of magnetoelasticity and at variance with the linear descriptions of Kittel,[7] Schlömann,[8] and Akhiezer, Bariakhtar, and Peletminskii,[9] which are based on the infinitesimal magnetostrictive theory of Becker and Döring.[10] Recently Tiersten[11] introduced a continuum model of polarization and employed the notion of the quasi-static electric field[12] to obtain a description of dynamic thermoelectroelasticity. In the static case in the absence of heat conduction, these equations are equivalent to those of Toupin[1] and Eringen.[2]

In 1963 Toupin[13] presented a theory of the electrodynamics of finitely deformable, polarizable continua. This description encompasses electromagnetic propagation and reduces to his earlier theory in the static case. Toupin obtains his equations by postulating certain electromagnetic-mechanical interaction terms without defining a model in any specific detail. In 1965 Dixon and Eringen[14] employed a particle model and averaging techniques to obtain equations for the electrodynamics of deformable continua. When the unusually general case treated by Dixon and Eringen[14] is reduced to that treated by Toupin,[13] the resulting equations appear to differ in certain respects. Penfield and Haus[15] have discussed the interaction of the electromagnetic field with deformable continua in a variety of circumstances. Throughout their treatment, they employ a procedure they call the Principle of Virtual Power in order to obtain momentum interaction terms, and they systematically employ the concept of magnetic poles and use the formalism of Chu,[16] which they compare to other formalisms. It is difficult to compare their equations with those of others in seemingly equivalent situations because they do not present the complete constitutive theory. Recently, Nelson and Lax,[17] in a treatment of the acoustical scattering of optical waves, obtain a description of the interaction of the electromagnetic field with a finitely deformable, polarizable continuum by means of a variational principle. The resulting equations, although similar to, appear to differ in certain respects from those of Toupin[13] and Dixon and Eringen.[14] However, more careful examination may reveal that they are actually equivalent in more ways than are readily apparent.[18] None of the aforementioned electrodynamic works is Lorentz invariant, but that is not a severe limitation because material velocities encountered in practice are considerably less than the speed of light. Nevertheless, Grot and Eringen[19] obtained a Lorentz invariant description using an appropriately simplified version of the model of Dixon and Eringen[14] and Bragg[20] presented a Lorentz invariant description of the electrodynamic theory of Toupin.[13] In the absence of intrinsic magnetic moment, the low velocity limit of the work of Grot and Eringen[19] has been shown[21] to be equivalent to that of Toupin.[13] However, the flexibility of the model as related to its potential use in diverse physical situations and the range of applicability of the resulting equations are of greater interest to us at present than whether or not a particular set of equations is Lorentz invariant, although Lorentz invariance per se is certainly desirable.

In this paper the differential equations and boundary conditions describing the behavior of a finitely deformable, polarizable, and magnetizable, heat conducting continuum in interaction with the electromagnetic field are derived by means of a systematic application of the laws of continuum physics to a well-defined macroscopic model consisting of appropriately defined interpenetrating continua. Magnetic spin resonance and both ionic and electronic polarization resonances are included in the treatment. In essence, this work couples previous work in magnetoelasticity[4] with an extended version of recent work in thermoelectroelasticity.[11] The resulting description is not Lorentz invariant, but, as already noted, that is not a severe limitation because macroscopic material velocities are considerably less than the speed of light. However, it is to be noted that the description should be accurate to terms linear in the ratio of the material velocities to the speed of light, and, consequently, should be capable of accurately describing very small velocity effects. Moreover, as we indicate later on, there are reasons to believe that

it will not be difficult to find the equivalent Lorentz invariant description in simplified cases.

The model, from which the description is obtained, consists of an electronic charge and spin continuum coupled to a lattice continuum. The lattice continuum is somewhat more complicated than any considered heretofore,[4,11] in that it consists of two interpenetrating ionic continua, which can displace with respect to each other and, thus, produce ionic polarization. The electronic continuum is a combination of one used previously in magnetoelasticity[4] and one employed recently in electroelasticity,[11] in that both the charge and angular momentum of the electronic continuum are taken into account. However, the definition of the spin continuum employed here is more fundamental than the one used previously,[4] in that this one consists of a circulating current density, which, in the appropriate limit, accounts for the magnetization. The more fundamental approach employed here results in a different and more satisfying description of the magnetic energetics. The difference, although important in principle, is not significant in practice whenever quasi-magnetostatics is applicable. The identified continua interact by means of defined local electric and magnetic material fields, which cause balancing forces and couples to be exerted between the continua.

Since the fundamental electrical and magnetic constituents of the matter are taken to consist of charge and current as opposed to polarization and magnetization, the electromagnetic field interacts with the material continuum in accordance with the Lorentz force. However, the rates of supply of linear momentum and energy from the electromagnetic field to the magnetized and polarized continuum are determined from the Lorentz prescription by means of appropriate continuum definitions of magnetization and polarization rather than by some particle averaging techniques, as has been done in the past. In so doing we are able to treat magnetization in addition to polarization, and in a straightforward analogous manner; and further, we are able to treat magnetism as a dynamic current-induced phenomenon, which we find desirable since magnetic monopoles have never been observed experimentally. Lorentz, in his theory of electrons,[22] does not attempt to treat magnetization explicitly, nor do Toupin[13] or Nelson and Lax.[17] Dixon and Eringen[14] include magnetization and define a model in some detail, but their model is a particle model and does not consist of interpenetrating continua and is not defined in as much specific detail as the one employed here.

The application of the appropriate equations of balance of mass and momentum to the respective continua yields the material equations of motion, which, with the electromagnetic field equations, constitute, as usual, an underdetermined system. The application of the equation of the conservation of energy to the combined material continuum results in the first law of thermodynamics which, with the aid of the second law of thermodynamics[23–25] and the principle of material objectivity,[26,27] enables the determination of the constitutive equations of our nonlinear theory. These constitutive equations along with the aforementioned equations of motion and electromagnetism and the thermodynamic dissipation equation result in a properly determined system, which can readily be reduced to 16 equations in 16 dependent variables. This system of equations in addition to encompassing such phenomena as magnetic spin resonance including the exchange interaction and ionic and electronic polarization resonances, also accounts for frequency dispersion caused by ionic polarization. This occurs because the material stored energy function turns out to depend on the ionic polarization gradient, among other variables, on account of the definition of the ionic continuum. Mindlin[28] has considered the dependence of the stored energy function on the polarization gradient in the static linear case by means of a variational technique. The relation of Mindlin's theory of the polarization gradient to the small displacement equations of lattice dynamics has been discussed by Askar, Lee, and Cakmak.[29]

When the interaction problem under consideration is such that ionic and electronic polarization resonances can be left out of account and the dependence of the stored energy function on the ionic polarization gradient can be ignored, the distinction between ionic and electronic polarization may be omitted and the model of polarization may be replaced by the simpler electronic-lattice continuum model employed previously.[11] If, in addition to the abovementioned simplifications, magnetic spin resonance and the exchange interaction are left out of account, the resulting system of nonlinear equations can readily be reduced to eight equations in eight dependent variables in place of the aforementioned 16 equations in 16 dependent variables. Under appropriate circumstances, intermediate systems of equations can be obtained. These reductions are greatly facilitated by the use of a Legendre transformation in the first law of thermodynamics, which in each instance results in a valuable change of constitutive variables.

In order to complete the system of equations, jump (or boundary) conditions across moving, not necessarily material, surfaces of discontinuity are determined from the appropriate integral forms of the field equations, which are taken to be valid even when the differential forms from which they were obtained are not. In addition to the integral forms of the balance equations, important integral forms of the momentum and energy relations for the electromagnetic field equations in deformable continua are obtained from the aforementioned fundamental expressions for the rates of supply of momentum and energy from the electromagnetic field to the matter with the aid of the low velocity limit of the relativistic transformations of the electromagnetic field vectors. The integral form of the momentum relations results in expressions for the Maxwell tensor and the electromagnetic momentum that reduce to those of Livens[30] when the matter is at rest. The integral form of the energy relation reduces to Poynting's theorem, which is interpreted in a somewhat more general way than heretofore, when the matter is at rest. The integral form of the momentum relations is essential in the determination of boundary conditions, which are required if boundary value problems are to be properly formulated and solved. The integral form of the energy relation is required in principle, provides interesting insight and can be used to obtain approximate information about jumps across surfaces of discontinuity when a solution is not available.

2. THE INTERACTING CONTINUA

As stated in the Introduction, the macroscopic model consists essentially of an electronic charge and spin continuum coupled to a lattice continuum, which, in itself, consists of two interpenetrating ionic continua. Actually there are two electronic continua, one for each ionic continuum. However, as we shall see, a sufficient number of assumptions are made so that the two electronic continua can be regarded essentially as one. Initially, all continua occupy the same region of space and, hence, have the same material coordinates X_L. The motion of a point of the lattice continuum, which is at the center of mass of the ionic continua, is described by the mapping[31]

$$y_i = y_i(X_L, t), \qquad \mathbf{y} = \mathbf{y}(\mathbf{X}, t), \tag{2.1}$$

which is one-to-one and differentiable as often as required. In (2.1), the y_i denote the spatial coordinates and X_L, the material coordinates and t denotes time. We consistently use the convention that capital indices denote the Cartesian components of $\mathbf{X}$ and lower case indices, the Cartesian components of $\mathbf{y}$. ($\mathbf{X}$ and $\mathbf{y}$ denote the initial position of all material points and the center of mass of the ionic continua, respectively.) Both dyadic and Cartesian tensor notations are used interchangeably. A comma followed by an index denotes partial differentiation with respect to a coordinate, i.e.,

$$y_{i,L} = \frac{\partial y_i}{\partial X_L}, \qquad X_{K,j} = \frac{\partial X_K}{\partial y_j}, \tag{2.2}$$

and the summation convention for repeated tensor indices is employed. The lattice continuum has a positive charge density σ^l and the (total) electronic continuum, a negative charge density σ^e. The plus (minus) ionic continuum has a positive net charge density $\sigma^+(\sigma^-)$, which consists of a positive lattice charge density $\sigma^{l+}(\sigma^{l-})$ and a negative electronic charge density $\sigma^{e+}(\sigma^{e-})$, and we may write

$$\sigma^+ = \sigma^{l+} + \sigma^{e+}, \tag{2.3}$$

$$\sigma^- = \sigma^{l-} + \sigma^{e-}, \tag{2.4}$$

and

$$\sigma^l = \sigma^{l+} + \sigma^{l-}, \tag{2.5}$$

$$\sigma^e = \sigma^{e+} + \sigma^{e-}, \tag{2.6}$$

so that we have

$$\sigma^+ + \sigma^- = \sigma^l + \sigma^e. \tag{2.7}$$

Clearly, in addition to possessing charge, each portion of each ionic continuum possesses mass, and we have the analogous mass relationships

$$\rho^+ = \rho^{l+} + \rho^{e+}, \tag{2.8}$$

$$\rho^- = \rho^{l-} + \rho^{e-}, \tag{2.9}$$

$$\rho^l = \rho^{l+} + \rho^{l-}, \tag{2.10}$$

$$\rho^e = \rho^{e+} + \rho^{e-}, \tag{2.11}$$

where the ρ with the appropriate superscript refers to the mass density of the continuum associated with that superscript, and all ρ and σ in (2.3)–(2.11) have the same material coordinate $\mathbf{X}$. Clearly, from (2.8)–(2.11), we have

$$\rho = \rho^+ + \rho^- = \rho^l + \rho^e, \tag{2.12}$$

where ρ is the (total) mass density of the combined continuum.

TABLE I. Charge densities of continua.

	Lattice	Electronic	Total
+ Ionic	σ^{l+}	σ^{e+}	σ^+
− Ionic	σ^{l-}	σ^{e-}	σ^-
Total	σ^l	σ^e	0

In a (finite) motion, the ionic continua are permitted to displace with respect to the center of mass of the ionic continua by *infinitesimal* displacement fields $\mathbf{w}^+$ and $\mathbf{w}^-$ and each electronic continuum is permitted to displace with respect to its ionic continuum by *additional infinitesimal* displacement fields $\boldsymbol{\eta}^+$ and $\boldsymbol{\eta}^-$. However, since we are interested in considering only *one* electronic polarization resonance, we arbitrarily assume that $\boldsymbol{\eta}^+ = \boldsymbol{\eta}^- = \boldsymbol{\eta}$. This is one of the assumptions which enables us to discuss *one* electronic continuum without regard to the ionic continuum with which the specific electronic continuum is associated. A schematic diagram indicating the motion of the model appears in Fig. 1. The infinitesimal displacement fields $\mathbf{w}^+$, $\mathbf{w}^-$, and $\boldsymbol{\eta}$ are regarded as functions of $\mathbf{y}$ and t and are constrained to satisfy

$$w^+_{k,k} = 0, \qquad w^-_{k,k} = 0, \qquad \eta_{k,k} = 0, \tag{2.13}$$

in order to assure that elements of the different continua, with the same material coordinates, have equal volumes at all times. Since the net charge density at any material point vanishes initially, i.e., in the reference configuration, and (2.13) are satisfied identically, we have, by virtue of the conservation of charge,

$$\sigma^+ + \sigma^- = 0, \tag{2.14}$$

$$\sigma^l + \sigma^e = 0 \tag{2.15}$$

for all times. The charge relations contained in (2.3)–(2.7), (2.14), and (2.15) are shown in Table I.

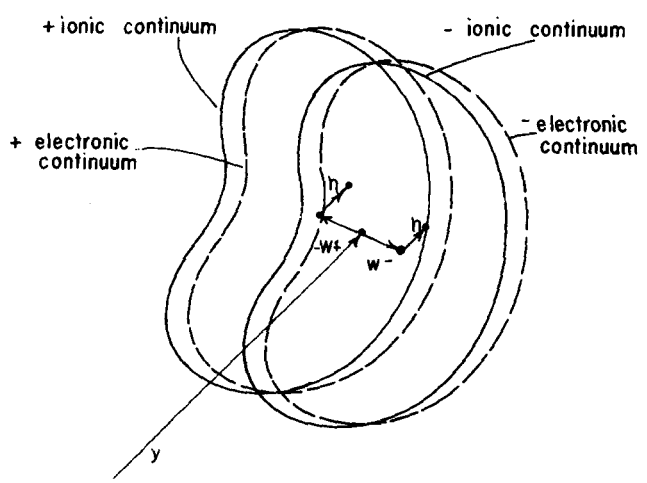


FIG. 1. Schematic diagram showing the relative displacements of the interacting continua.

On account of the conservation of mass and charge of each continuum and (2.13), the ratios of mass density to charge density for the respective continua are

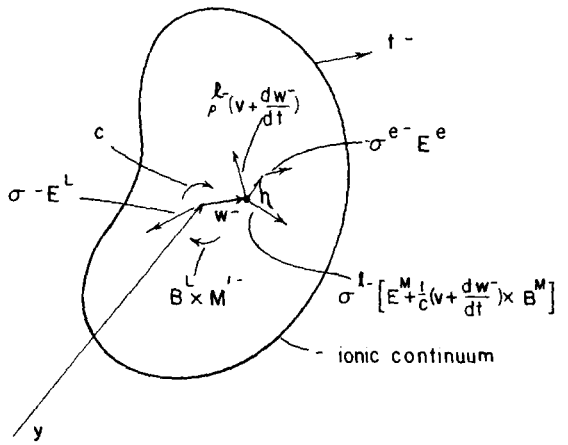


FIG. 2. Schematic diagram showing the linear momentum and force and couple vectors acting in the negative ionic continuum. A similar diagram can be drawn for the positive ionic continuum.

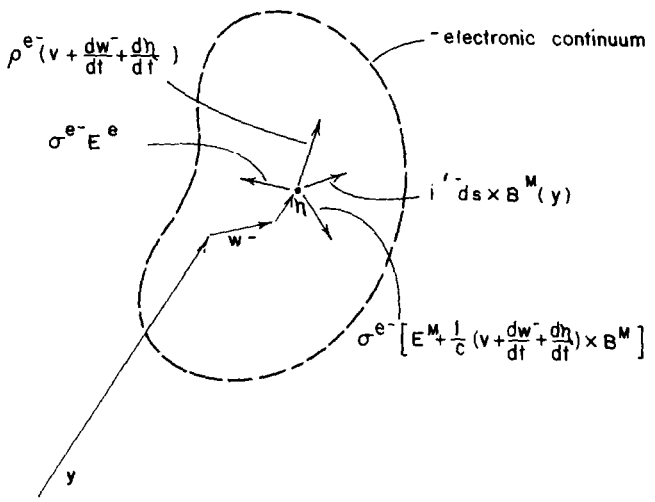


FIG. 3. Schematic diagram showing the linear momentum and force vectors acting in the negative electronic continuum. A similar diagram can be drawn for the positive electronic continuum.

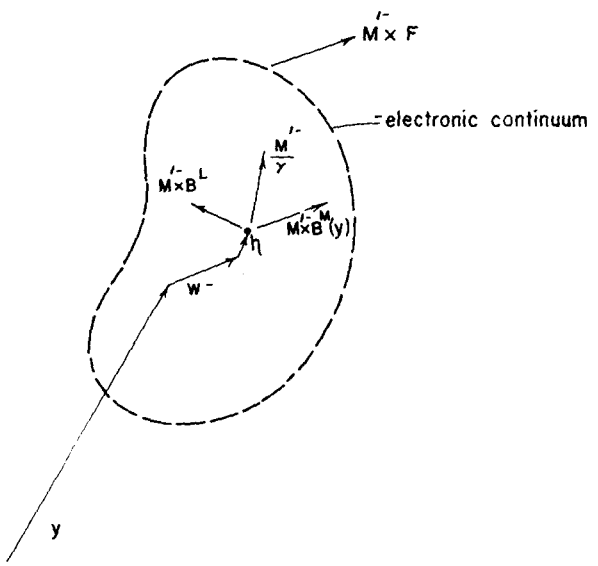


FIG. 4. Schematic diagram showing the angular momentum and couple vectors acting in the negative electronic continuum. A similar diagram can be drawn for the positive electronic continuum.

fixed constant numbers, and we may write

$$\rho^+ = r^+\sigma^+, \quad \rho^- = r^-\sigma^-, \quad \rho^e = r^e\sigma^e, \tag{2.16}$$

where r^+ and r^- are given constants for a specific material and r^e is the ratio of the mass to the charge of an electron. Although other r coefficients can readily be defined, r^+, r^-, and r^e turn out to be the only ones needed because we introduce the additional material constants

$$m = \rho^-/\rho^+, \quad s = \rho^e/\rho^l, \tag{2.17}$$

$$e = \sigma^{e-}/\sigma^{e+}, \quad r^m = |\mathbf{M}'^-|/|\mathbf{M}'^+|, \tag{2.18}$$

which are more convenient. In (2.18), $\mathbf{M}'^+$ and $\mathbf{M}'^-$ denote the magnetizations associated with the respective ionic continua. From (2.14), (2.16), and (2.17), we have

$$r^- = -mr^+. \tag{2.19}$$

Since a point $\mathbf{y}$ of the lattice continuum is defined as the center of mass of the interpenetrating ionic continua, we may write

$$\int_{V^+}(\mathbf{y} + \mathbf{w}^+)\rho^+ dV + \int_{V^-}(\mathbf{y} + \mathbf{w}^-)\rho^- dV = \int_V(\rho^+ + \rho^-)\mathbf{y}dV, \tag{2.20}$$

and since, by virtue of (2.13), $V^+ = V^- = V$, we have

$$\rho^+\mathbf{w}^+ + \rho^-\mathbf{w}^- = 0. \tag{2.21}$$

In addition, because of the conservation of mass of each ionic continuum, we further obtain

$$\rho^+\frac{d\mathbf{w}^+}{dt} + \rho^-\frac{d\mathbf{w}^-}{dt} = 0, \tag{2.22}$$

where d/dt denotes the material derivative[32] following $\mathbf{y}$.

The electronic continuum associated with each ionic continuum, in addition to possessing the negative charge density σ^e and appropriate linear momentum, possesses at each point a circulating current density $\oint_C i'd\mathbf{s}$, which in the appropriate limit accounts for the magnetization $\mathbf{M}'$, and angular momentum density $\mathbf{M}'/\gamma$ (γ the gyromagnetic ratio and a negative number), where $\mathbf{M}'$ is the magnetization referred to the instantaneous local rest system of inertia[33] of that point $\mathbf{y}$ of the deformable continuum. Each electronic continuum interacts with its ionic continuum by means of defined local electric material fields $\mathbf{E}^{e\pm}$, which cause equal and opposite forces $\sigma^{e\pm}\mathbf{E}^{e\pm}$ and $-\sigma^{e\pm}\mathbf{E}^{e\pm}$ to be exerted between the respective electronic and ionic continua that act through the respective points $(\mathbf{y} + \mathbf{w}^\pm + \eta)$, and defined local magnetic material fields $\mathbf{B}^{L\pm}$, which cause equal and opposite couples $\mathbf{M}'^\pm \times \mathbf{B}^{L\pm}$ and $\mathbf{B}^{L\pm} \times \mathbf{M}'^\pm$ to be exerted between the respective electronic spin and ionic continua. Each electronic spin continuum interacts with neighboring elements of the same spin continuum by means of a magnetic exchange field[4] $\mathbf{F}^\pm$, which acts across the surface of separation to produce a couple per unit surface area $\mathbf{M}'^\pm \times \mathbf{F}^\pm$. However, since we are interested in treating only *one* electronic polarization resonance and *one* magnetic spin resonance, we arbitrarily assume that $\mathbf{E}^{e+} = \mathbf{E}^{e-} = \mathbf{E}^e$,

$\mathbf{B}^{L+} = \mathbf{B}^{L-} = \mathbf{B}^{L}$, and $\mathbf{F}^{+} = \mathbf{F}^{-} = \mathbf{F}$. By virtue of these assumptions and the previous assumption on η, we can usually discuss the two distinct electronic continua as if they were one. The ionic continua interact with each other by means of a defined local electric material field $\mathbf{E}^{L}$, which causes equal and opposite forces $\sigma^{+}\mathbf{E}^{L}$ and $\sigma^{-}\mathbf{E}^{L}$ to be exerted between the ionic continua at the position $\mathbf{y}$, and defined equal and opposite local material couples $\mathbf{c}$. Each ionic continuum interacts with neighboring elements of the same ionic continuum by means of a traction force per unit area $\mathbf{t}^{\pm}$ acting across the surface of separation. As stated in the Introduction, the Maxwell electric field $\mathbf{E}^{M}$ and magnetic induction field $\mathbf{B}^{M}$ exert the usual Lorentz force on all elements of charge and current density. Schematic diagrams illustrating the abovementioned interactions in the model are shown in Figs. 2–4.

3. THE EQUATIONS OF BALANCE

In view of the discussion in Sec. 2, the equations of the conservation of mass for the different continua may be written in the form

$$\frac{d}{dt}\int_V \rho^{l+}\,dV = 0, \quad \frac{d}{dt}\int_V \rho^{e+}\,dV = 0, \tag{3.1}$$

$$\frac{d}{dt}\int_V \rho^{l-}\,dV = 0, \quad \frac{d}{dt}\int_V \rho^{e-}\,dV = 0, \tag{3.2}$$

where V is an arbitrary element of volume for which each of the continua has the same material coordinates. From (2.8), (2.9) or (2.10), (2.11), and (2.12), (3.1), and (3.2), we obtain the equation of the conservation of mass for the combined continuum in the form

$$\frac{d}{dt}\int_V \rho\,dV = 0. \tag{3.3}$$

The equations of the conservation of linear and angular momentum for the electronic charge and spin continuum associated with the positive ionic continuum are, respectively,

$$\int_V \sigma^{e+}\left[\mathbf{E}^{M}(\mathbf{y}+\mathbf{w}^{+}+\eta) + \mathbf{E}^{e} + \frac{1}{C}\left(\mathbf{v}+\frac{d\mathbf{w}^{+}}{dt} + \frac{d\eta}{dt}\right)\times \mathbf{B}^{M}(\mathbf{y}+\mathbf{w}^{+}+\eta)\right]dV + \int_V \oint_{C'} i'^{+}d\mathbf{s}\times \mathbf{B}^{M}(\mathbf{y})\,dV$$
$$= \frac{d}{dt}\int_V \rho^{e+}\left(\mathbf{v}+\frac{d\mathbf{w}^{+}}{dt}+\frac{d\eta}{dt}\right)dV, \tag{3.4}$$

$$\int_V (\mathbf{y}+\mathbf{w}^{+}+\eta)\times\left\{\sigma^{e+}\left[\mathbf{E}^{M}(\mathbf{y}+\mathbf{w}^{+}+\eta)+\mathbf{E}^{e} + \frac{1}{C}\left(\mathbf{v}+\frac{d\mathbf{w}^{+}}{dt}+\frac{d\eta}{dt}\right)\times \mathbf{B}^{M}(\mathbf{y}+\mathbf{w}^{+}+\eta)\right] + \oint_{C'} i'^{+}\,d\mathbf{s}\times \mathbf{B}^{M}(\mathbf{y})\right\}dV + \int_S \mathbf{M}'^{+}\times \mathbf{F}\,dS$$
$$+ \int_V \mathbf{M}'^{+}\times(\mathbf{B}^{M}(\mathbf{y}) + \mathbf{B}^{L})\,dV = \frac{d}{dt}\int_V (\mathbf{y}+\mathbf{w}^{+}+\eta)\times\rho^{e+}\left(\mathbf{v}+\frac{d\mathbf{w}^{+}}{dt}+\frac{d\eta}{dt}\right)dV + \frac{d}{dt}\int_V \frac{\mathbf{M}'^{+}}{\gamma}\,dV, \tag{3.5}$$

where $\mathbf{v} = d\mathbf{y}/dt$, i'^{+} is in magnetic units, C is the speed of light, C' is an arbitrary vanishingly small circulating current loop taken to be stationary with respect to the local rest system of inertia Gaussian electromagnetic units are employed and we have used the relation[34]

$$\lim_{\substack{\mathbf{r}\to 0 \\ i'^{+}\to\infty \\ C'\to 0 \text{ in a plane}}} \oint_{C'} \mathbf{r}\times(i'^{+}d\mathbf{s}\times\mathbf{B}^{M}) = \lim_{\substack{S'_0\to 0 \\ i'^{+}\to\infty \\ \mathbf{n}\text{ fixed}}} i'^{+}\int_{S'_0}\mathbf{n}\times\mathbf{B}^{M}\,dS = \mathbf{M}'^{+}\times\mathbf{B}^{M}, \tag{3.6}$$

where

$$\mathbf{M}'^{+} = \lim_{\substack{i'^{+}\to\infty \\ S'_0\to 0 \\ \mathbf{n}\text{ fixed}}} i'^{+}\mathbf{n}S'_0. \tag{3.7}$$

The equations of the conservation of linear and angular momentum for the lattice portion of the positive ionic continuum are, respectively,

$$\int_S \mathbf{t}^{+}dS + \int_V\left[\sigma^{l+}\mathbf{E}^{M}(\mathbf{y}+\mathbf{w}^{+}) - \sigma^{e+}\mathbf{E}^{e} + \sigma^{+}\mathbf{E}^{L} + \frac{\sigma^{l+}}{C}\left(\mathbf{v}+\frac{d\mathbf{w}^{+}}{dt}\right)\times\mathbf{B}^{M}(\mathbf{y}+\mathbf{w}^{+})\right]dV = \frac{d}{dt}\int_V \rho^{l+}\left(\mathbf{v}+\frac{d\mathbf{w}^{+}}{dt}\right)dV, \tag{3.8}$$

$$\int_S(\mathbf{y}+\mathbf{w}^{+})\times\mathbf{t}^{+}dS + \int_V(\mathbf{y}+\mathbf{w}^{+})\times\left[\sigma^{l+}\mathbf{E}^{M}(\mathbf{y}+\mathbf{w}^{+}) - \sigma^{e+}\mathbf{E}^{e} + \frac{\sigma^{l+}}{C}\left(\mathbf{v}+\frac{d\mathbf{w}^{+}}{dt}\right)\times\mathbf{B}^{M}(\mathbf{y}+\mathbf{w}^{+})\right]dV$$
$$+ \int_V[\mathbf{y}\times\sigma + \mathbf{E}^{L} + \eta\times(-\sigma^{e+}\mathbf{E}^{e})\,\mathbf{C} + \mathbf{B}^{L}\times\mathbf{M}'^{+}]dV = \frac{d}{dt}\int_V(\mathbf{y}+\mathbf{w}^{+})\times\rho^{l+}\left(\mathbf{v}+\frac{d\mathbf{w}^{+}}{dt}\right)dV. \tag{3.9}$$

The equivalent equations of the conservation of linear and angular momentum for the electronic and lattice portions of the negative ionic continuum take the forms

$$\int_V \sigma^{e-}\left[\mathbf{E}^{M}(\mathbf{y}+\mathbf{w}^{-}+\eta) + \mathbf{E}^{e} + \frac{1}{C}\left(\mathbf{v}+\frac{d\mathbf{w}^{-}}{dt} + \frac{d\eta}{dt}\right)\times\mathbf{B}^{M}(\mathbf{y}+\mathbf{w}^{-}+\eta)\right]dV + \int_V\oint_{C'} i'^{-}d\mathbf{s}\times\mathbf{B}^{M}(\mathbf{y})\,dV$$
$$= \frac{d}{dt}\int_V \rho^{e-}\left(\mathbf{v}+\frac{d\mathbf{w}^{-}}{dt}+\frac{d\eta}{dt}\right)dV, \tag{3.10}$$

$$\int_V(\mathbf{y}+\mathbf{w}^{-}+\eta)\times\left\{\sigma^{e-}\left[\mathbf{E}^{M}(\mathbf{y}+\mathbf{w}^{-}+\eta)+\mathbf{E}^{e} + \frac{1}{C}\left(\mathbf{v}+\frac{d\mathbf{w}^{-}}{dt}+\frac{d\eta}{dt}\right)\times\mathbf{B}^{M}(\mathbf{y}+\mathbf{w}^{-}+\eta)\right] + \oint_{C'} i'^{-}d\mathbf{s}\times\mathbf{B}^{M}(\mathbf{y})\right\}dV + \int_S \mathbf{M}'^{-}\times\mathbf{F}\,dS$$
$$+ \int \mathbf{M}'^{-}\times(\mathbf{B}^{M}(\mathbf{y}) + \mathbf{B}^{L})\,dV = \frac{d}{dt}\int_V(\mathbf{y}+\mathbf{w}^{-}+\eta)\times\rho^{e-}\left(\mathbf{v}+\frac{d\mathbf{w}^{-}}{dt}+\frac{d\eta}{dt}\right)dV + \frac{d}{dt}\int_V\frac{\mathbf{M}'^{+}}{\gamma}dV, \tag{3.11}$$

$$\int_S \mathbf{t}^{-}dS + \int_V\left[\sigma^{l-}\mathbf{E}^{M}(\mathbf{y}+\mathbf{w}^{-}) - \sigma^{e-}\mathbf{E}^{e} + \sigma^{-}\mathbf{E}^{L} + \frac{\sigma^{l-}}{C}\left(\mathbf{v}+\frac{d\mathbf{w}^{-}}{dt}\right)\times\mathbf{B}^{M}(\mathbf{y}+\mathbf{w}^{-})\right]dV = \frac{d}{dt}\int_V\rho^{l-}\left(\mathbf{v}+\frac{d\mathbf{w}^{-}}{dt}\right)dV, \tag{3.12}$$

$$\int_S(\mathbf{y}+\mathbf{w}^{-})\times\mathbf{t}^{-}dS + \int_V(\mathbf{y}+\mathbf{w}^{-})\times\left[\sigma^{l-}\mathbf{E}^{M}(\mathbf{y}+\mathbf{w}^{-}) - \sigma^{e-}\mathbf{E}^{e} + \frac{\sigma^{l-}}{C}\left(\mathbf{v}+\frac{d\mathbf{w}^{-}}{dt}\right)\times\mathbf{B}^{M}(\mathbf{y}+\mathbf{w}^{-})\right]dV + \int_V[\mathbf{y}\times\sigma^{-}\mathbf{E}^{L}$$

$$+\eta \times (-\sigma^{e^-}\mathbf{E}^e) - \mathbf{C} + \mathbf{B}^L \times \mathbf{M}'^-]dV = \frac{d}{dt}\int_V (\mathbf{y} + \mathbf{w}^-) \times \rho^{l^-}\left(\mathbf{v} + \frac{d\mathbf{w}^-}{dt}\right) dV. \tag{3.13}$$

Since $\mathbf{w}^+$, $\mathbf{w}^-$, and η are all infinitesimal displacement fields, we expand $\mathbf{E}^M$ and $\mathbf{B}^M$ at $(\mathbf{y} + \mathbf{w}^\pm + \eta)$ in a Taylor series about y and retain the first term to obtain

$$\mathbf{E}^M(\mathbf{y} + \xi) = \mathbf{E}^M(\mathbf{y}) + \xi \cdot \nabla \mathbf{E}^M(\mathbf{y}), \tag{3.14}$$

$$\mathbf{B}^M(\mathbf{y} + \xi) = \mathbf{B}^M(\mathbf{y}) + \xi \cdot \nabla \mathbf{B}^M(\mathbf{y}), \tag{3.15}$$

where $\nabla = \mathbf{e}_i \partial/\partial y_i$ and $\mathbf{e}_i$ is a unit base vector in the ith Cartesian direction. However, since the displacement fields $\mathbf{w}^+$, $\mathbf{w}^-$, and η constitute an integral part of the models of the ionic and electronic polarization densities, but bear no relation to the model of the magnetization densities, which consists of the circulating current terms $i'^\pm d\mathbf{s}$, the expansion (3.15) is employed in charge density terms only and not in current density terms. This is tantamount to assuming that $i'^\pm d\mathbf{s}$ experiences the $\mathbf{B}^M$ at y rather than at $(\mathbf{y} + \mathbf{w}^\pm + \eta)$. This fact has already been taken into consideration in writing Eqs. (3.4), (3.5), (3.10), and (3.11).

Substituting (3.14) and (3.15) into (3.4) and (3.10), taking the material time derivative, employing the second of (3.1) and the second of (3.2), neglecting terms containing products of η and/or $\mathbf{w}^\pm$ and utilizing the fact that V is arbitrary, we obtain, respectively,

$$\begin{aligned}&\sigma^{e^+}\mathbf{E}^M + \sigma^{e^+}\mathbf{w}^+ \cdot \nabla \mathbf{E}^M + \sigma^{e^+}\eta \cdot \nabla \mathbf{E}^M + \frac{\sigma^{e^+}}{C}\mathbf{v} \times \mathbf{B}^M \\ &\quad + \frac{\sigma^{e^+}}{C}\frac{d\mathbf{w}^+}{dt} \times \mathbf{B}^M + \frac{\sigma^{e^+}}{C}\frac{d\eta}{dt} \times \mathbf{B}^M + \frac{\sigma^{e^+}}{C}\mathbf{v} \times (\mathbf{w}^+ \cdot \nabla \mathbf{B}^M) \\ &\quad + \frac{\sigma^{e^+}}{C}\mathbf{v} \times (\eta \cdot \nabla \mathbf{B}^M) + \sigma^{e^+}\mathbf{E}^e + \mathbf{M}'^+ \cdot \mathbf{B}^M\nabla \\ &= \rho^{e^+}\frac{d\mathbf{v}}{dt} + \rho^{e^+}\frac{d^2\mathbf{w}^+}{dt^2} + \rho^{e^+}\frac{d^2\eta}{dt^2},\end{aligned} \tag{3.16}$$

$$\begin{aligned}&\sigma^{e^-}\mathbf{E}^M + \sigma^{e^-}\mathbf{w}^- \cdot \nabla \mathbf{E}^M + \sigma^{e^-}\eta \cdot \nabla \mathbf{E}^M + \frac{\sigma^{e^-}}{C}\mathbf{v} \times \mathbf{B}^M \\ &\quad + \frac{\sigma^{e^-}}{C}\frac{d\mathbf{w}^-}{dt} \times \mathbf{B}^M + \frac{\sigma^{e^-}}{C}\frac{d\eta}{dt} \times \mathbf{B}^M + \frac{\sigma^{e^-}}{C}\mathbf{v} \times (\mathbf{w}^- \cdot \nabla \mathbf{B}^M) \\ &\quad + \frac{\sigma^{e^-}}{C}\mathbf{v} \times (\eta \cdot \nabla \mathbf{B}^M) + \sigma^{e^-}\mathbf{E}^e + \mathbf{M}'^- \cdot \mathbf{B}^M\nabla \\ &= \rho^{e^-}\frac{d\mathbf{v}}{dt} + \rho^{e^-}\frac{d^2\mathbf{w}^-}{dt^2} + \rho^{e^-}\frac{d^2\eta}{dt^2},\end{aligned} \tag{3.17}$$

where we have utilized the relation[34]

$$\lim_{\substack{i'\to\infty \\ C'\to 0 \text{ in a plane}}} \oint_{C'} i'^\pm d\mathbf{s} \times \mathbf{B}^M = \lim_{\substack{i'\to\infty \\ S_0'\to 0 \\ \mathbf{n} \text{ fixed}}} \int_{S_0'} i'^\pm \mathbf{n} \cdot (\mathbf{B}^M\nabla) dS = \mathbf{M}'^\pm \cdot \mathbf{B}^M\nabla \tag{3.18}$$

and employed (3.7). Adding (3.5) and (3.11), and utilizing the second of (3.1) and (3.2), (3.4), (3.10), and the relation

$$\mathbf{M}' = \mathbf{M}'^+ + \mathbf{M}'^-, \tag{3.19}$$

we obtain

$$\int_S \mathbf{M}' \times \mathbf{F} dS + \int_V \mathbf{M}' \times (\mathbf{B}^M + \mathbf{B}^L) dV = \frac{d}{dt}\int_V \frac{\mathbf{M}'}{\gamma} dV, \tag{3.20}$$

which is the integral form of the conservation of angular momentum for the electronic spin continuum.[35] Application of (3.20) to an elementary tetrahedron in the usual manner yields the definition of the magnetic exchange tensor **A**:

$$\mathbf{F} = -\mathbf{n} \cdot \mathbf{A}, \tag{3.21}$$

where **n** is the outwardly directed unit normal and, by virtue of the definition of **F**, we have, without any loss of generality,

$$\mathbf{A} \cdot \mathbf{M}' = 0, \tag{3.22}$$

which reduces from nine to six the number of possible components of **A**. Since the material is magnetically saturated and mass is conserved, the magnitude of the magnetic moment per unit mass μ', defined by

$$\mu' = \mathbf{M}'/\rho, \tag{3.23}$$

is conserved, and we have

$$\mu' \cdot \mu' = \mu_s'^2, \tag{3.24}$$

and μ_s' is constant in a homogeneous material. Substituting from (3.21) and (3.23) into (3.20), taking the material time derivative, utilizing (3.3), the divergence theorem and the arbitrariness of V, we obtain

$$\mu' \times (\mathbf{B}^M - \nabla \cdot \mathbf{A} - (1/\rho)\nabla\rho \cdot \mathbf{A} + \mathbf{B}^L) = \frac{1}{\gamma}\frac{d\mu'}{dt}, \tag{3.25}$$

where we have introduced the condition[36]

$$A_{lk}\mu'_{j,l} = A_{lj}\mu'_{k,l}, \tag{3.26}$$

which is required on account of (3.24). Equation (3.25) is the magnetodynamic equation of motion of the spin system. Adding (3.16) and (3.17), and substituting from (2.6), (2.11), and (3.19), we obtain

$$\begin{aligned}&\sigma^e\mathbf{E}^M + (\sigma^{e^+}\mathbf{w}^+ + \sigma^{e^-}\mathbf{w}^-) \cdot \nabla \mathbf{E}^M + \sigma^e\eta \cdot \nabla \mathbf{E}^M \\ &\quad + \frac{\sigma^e}{C}\mathbf{v} \times \mathbf{B}^M + \frac{1}{C}\left(\sigma^{e^+}\frac{d\mathbf{w}^+}{dt} + \sigma^{e^-}\frac{d\mathbf{w}^-}{dt}\right) \times \mathbf{B}^M \\ &\quad + \frac{\sigma^e}{C}\frac{d\eta}{dt} \times \mathbf{B}^M + \frac{\mathbf{v}}{C} \times [(\sigma^{e^+}\mathbf{w}^+ + \sigma^{e^-}\mathbf{w}^-) \cdot \nabla \mathbf{B}^M] \\ &\quad + \frac{\mathbf{v}}{C} \times (\sigma^e\eta \cdot \nabla \mathbf{B}^M) + \sigma^e\mathbf{E}^e + \mathbf{M}' \cdot \mathbf{B}^M\nabla = \rho^e\frac{d\mathbf{v}}{dt} \\ &\quad + \rho^{e^+}\frac{d^2\mathbf{w}^+}{dt^2} + \rho^{e^-}\frac{d^2\mathbf{w}^-}{dt^2} + \rho^e\frac{d^2\eta}{dt^2},\end{aligned} \tag{3.27}$$

which is one form of the equation of motion of *all* the electronic charge, i.e., of *the* electronic charge continuum.

Application of (3.8) and (3.12) to an elementary tetrahedron in the usual manner yields the definition of the respective stress tensors of the positive and negative ionic continua, thus,

$$\mathbf{t}^+ = \mathbf{n} \cdot \tau^+, \quad \mathbf{t}^- = \mathbf{n} \cdot \tau^-. \tag{3.28}$$

Substituting from the first of (3.28) into (3.8) and

from the second of (3.28) into (3.12), taking the material time derivatives, using the first of (3.1) and (3.2), (3.14), (3.15), the divergence theorem and the arbitrariness of V, we obtain

$$\nabla\cdot\tau^{+}+\sigma^{l+}\mathbf{E}^{M}+\sigma^{l+}\mathbf{w}^{+}\cdot\nabla\mathbf{E}^{M}-\sigma^{e+}\mathbf{E}^{e}+\sigma^{+}\mathbf{E}^{L}+\frac{\sigma^{l+}}{C}\mathbf{v}\times\mathbf{B}^{M}+\frac{\sigma^{l+}}{C}\mathbf{v}\times(\mathbf{w}^{+}\cdot\nabla\mathbf{B}^{M})+\frac{\sigma^{l+}}{C}\frac{d\mathbf{w}^{+}}{dt}\times\mathbf{B}^{M}=\rho^{l+}\frac{d\mathbf{v}}{dt}+\rho^{l+}\frac{d^{2}\mathbf{w}^{+}}{dt^{2}},\tag{3.29}$$

$$\nabla\cdot\tau^{-}+\sigma^{l-}\mathbf{E}^{M}+\sigma^{l-}\mathbf{w}^{-}\cdot\nabla\mathbf{E}^{M}-\sigma^{e-}\mathbf{E}^{e}+\sigma^{-}\mathbf{E}^{L}+\frac{\sigma^{l-}}{C}\mathbf{v}\times\mathbf{B}^{M}+\frac{\sigma^{l-}}{C}\mathbf{v}\times(\mathbf{w}^{-}\cdot\nabla\mathbf{B}^{M})+\frac{\sigma^{l-}}{C}\frac{d\mathbf{w}^{-}}{dt}\times\mathbf{B}^{M}=\rho^{l-}\frac{d\mathbf{v}}{dt}+\rho^{l-}\frac{d^{2}\mathbf{w}^{-}}{dt^{2}},\tag{3.30}$$

which are the stress equations of motion of the lattice portions of the positive and negative ionic continua, respectively. Substituting from the first of (3.28) into (3.9) and from the second of (3.28) into (3.13), taking the material time derivatives, using the first of (3.1) and (3.2), the divergence theorem (3.29) and (3.30) and the arbitrariness of V, and neglecting products of $\mathbf{w}^{+}$, $\mathbf{w}^{-}$ and/or η in terms in which the Taylor expansions (3.14) and (3.15) were employed, but not in terms in which they were not, we obtain, respectively,

$$\mathbf{e}_{l}e_{lij}\tau^{+}_{ij}+\mathbf{C}+\mathbf{B}^{L}\times\mathbf{M}'^{+}+\mathbf{e}_{l}e_{lkj}(w^{+}_{k}\tau^{+}_{ij})_{,i}+\sigma^{l+}\mathbf{w}^{+}\times\mathbf{E}^{M}+\frac{\sigma^{l+}}{C}\mathbf{w}^{+}\times(\mathbf{v}\times\mathbf{B}^{M})+\mathbf{w}^{+}\times\left(\frac{\sigma^{l+}}{C}\frac{d\mathbf{w}^{+}}{dt}\times\mathbf{B}^{M}\right)-(\eta+\mathbf{w}^{+})\times\sigma^{e+}\mathbf{E}^{e}-\mathbf{w}^{+}\times\rho^{l+}\frac{d\mathbf{v}}{dt}-\mathbf{w}^{+}\rho^{l+}\frac{d^{2}\mathbf{w}^{+}}{dt^{2}}=0,\tag{3.31}$$

$$\mathbf{e}_{l}e_{lij}\tau^{-}_{ij}-\mathbf{C}+\mathbf{B}^{L}\times\mathbf{M}'^{-}+\mathbf{e}_{l}e_{lkj}(w^{-}_{k}\tau^{-}_{ij})_{,i}+\sigma^{l-}\mathbf{w}^{-}\times\mathbf{E}^{M}+\frac{\sigma^{l-}}{C}\mathbf{w}^{-}\times(\mathbf{v}\times\mathbf{B}^{M})+\mathbf{w}^{-}\times\left(\frac{\sigma^{l-}}{C}\frac{d\mathbf{w}^{-}}{dt}\times\mathbf{B}^{M}\right)-(\eta+\mathbf{w}^{-})\times\sigma^{e-}\mathbf{E}^{e}-\mathbf{w}^{-}\times\rho^{l-}\frac{d\mathbf{v}}{dt}-\mathbf{w}^{-}\times\rho^{l-}\frac{d^{2}\mathbf{w}^{-}}{dt^{2}}=0.\tag{3.32}$$

The procedure used in neglecting products of $\mathbf{w}^{+}$, $\mathbf{w}^{-}$ and/or η implicitly assumes that the products of the magnitudes of any of the infinitesimal displacement fields with the magnitudes of the gradients of the electric and magnetic fields $\mathbf{E}^{M}$ and $\mathbf{B}^{M}$ are sufficiently small compared to the fields themselves to justify the operations. Now, substituting from (3.16) into (3.31) and from (3.17) into (3.32), employing (2.3) and (2.4) and neglecting products of $\mathbf{w}^{+}$, $\mathbf{w}^{-}$, and η resulting from the use of (3.14) and (3.15), we obtain

$$\mathbf{e}_{l}e_{lij}\tau^{+}_{ij}+\mathbf{e}_{l}e_{lkj}(w^{+}_{k}\tau^{+}_{ij})_{,i}+\mathbf{C}+\mathbf{B}^{L}\times\mathbf{M}'^{+}+\sigma^{+}\mathbf{w}^{+}\times\mathbf{E}^{M}+\frac{\sigma^{+}}{C}\mathbf{w}^{+}\times(\mathbf{v}\times\mathbf{B}^{M})+\mathbf{w}^{+}\times\left(\frac{\sigma^{+}}{C}\frac{d\mathbf{w}^{+}}{dt}\times\mathbf{B}^{M}\right)+\eta\times\sigma^{e+}\mathbf{E}^{M}+\eta\times\frac{\sigma^{e+}}{C}(\mathbf{v}\times\mathbf{B}^{M})+\eta\times\frac{\sigma^{e+}}{C}\left(\frac{d\mathbf{w}^{+}}{dt}\times\mathbf{B}^{M}\right)+\eta\times\frac{\sigma^{e+}}{C}\left(\frac{d\eta}{dt}\times\mathbf{B}^{M}\right)+\mathbf{w}^{+}\times\frac{\sigma^{e+}}{C}\left(\frac{d\eta}{dt}\times\mathbf{B}^{M}\right)-\mathbf{w}^{+}\times\rho^{+}\frac{d\mathbf{v}}{dt}-\mathbf{w}^{+}\times\rho^{+}\frac{d^{2}\mathbf{w}^{+}}{dt^{2}}-\mathbf{w}^{+}\times\rho^{e+}\frac{d^{2}\eta}{dt^{2}}-\eta\times\rho^{e+}\frac{d\mathbf{v}}{dt}-\eta\times\rho^{e+}\frac{d^{2}\mathbf{w}^{+}}{dt^{2}}-\eta\times\rho^{e+}\frac{d^{2}\eta}{dt^{2}}=0,\tag{3.33}$$

$$\mathbf{e}_{l}e_{lij}\tau^{-}_{ij}+\mathbf{e}_{l}e_{lkj}(w^{-}_{k}\tau^{-}_{ij})_{,i}-\mathbf{C}+\mathbf{B}^{L}\times\mathbf{M}'^{-}+\sigma^{-}\mathbf{w}^{-}\times\mathbf{E}^{M}+\frac{\sigma^{-}}{C}\mathbf{w}^{-}\times(\mathbf{v}\times\mathbf{B}^{M})+\mathbf{w}^{-}\times\left(\frac{\sigma^{-}}{C}\frac{d\mathbf{w}^{-}}{dt}\times\mathbf{B}^{M}\right)+\eta\times\sigma^{e-}\mathbf{E}^{M}+\eta\times\frac{\sigma^{e-}}{C}(\mathbf{v}\times\mathbf{B}^{M})+\eta\times\frac{\sigma^{e-}}{C}\left(\frac{d\mathbf{w}^{-}}{dt}\times\mathbf{B}^{M}\right)+\eta\times\frac{\sigma^{e-}}{C}\left(\frac{d\eta}{dt}\times\mathbf{B}^{M}\right)+\mathbf{w}^{-}\times\frac{\sigma^{e-}}{C}\left(\frac{d\eta}{dt}\times\mathbf{B}^{M}\right)-\mathbf{w}^{-}\times\rho^{-}\frac{d\mathbf{v}}{dt}-\mathbf{w}^{-}\times\rho^{-}\frac{d^{2}\mathbf{w}^{-}}{dt^{2}}-\mathbf{w}^{-}\times\rho^{e-}\frac{d^{2}\eta}{dt^{2}}-\eta\times\rho^{e-}\frac{d\mathbf{v}}{dt}-\eta\times\rho^{e-}\frac{d^{2}\mathbf{w}^{-}}{dt^{2}}-\eta\times\rho^{e-}\frac{d^{2}\eta}{dt^{2}}=0,\tag{3.34}$$

which constitute one form of the equations of the conservation of angular momentum of the positive and negative ionic continua, respectively, including the electronic continuum associated with each ionic continuum.

At this point we introduce the following definitions,

$$\mathbf{P}^{e}=\sigma^{e}\eta=\rho\pi^{e},\tag{3.35}$$

$$\mathbf{P}^{I}=\sigma^{+}\mathbf{w}^{+}+\sigma^{-}\mathbf{w}^{-}=\rho\pi^{I},\tag{3.36}$$

where $\mathbf{P}^{e}$ and $\mathbf{P}^{I}$, π^{e} and π^{I} are the electronic and ionic polarizations per unit volume and per unit mass, respectively. In view of (2.14), (2.17), (2.21), and (3.36), we have

$$\mathbf{P}^{I}=\sigma^{-}\mathbf{w}^{-}(1+m)=\sigma^{+}\mathbf{w}^{+}(m+1)/m\tag{3.37}$$

and from (3.35)–(3.37), (3.1)–(3.3), (2.16), and (2.22), we have

$$\rho\frac{d\pi^{e}}{dt}=\sigma^{e}\frac{d\eta}{dt},\tag{3.38}$$

$$\rho\frac{d\pi^{I}}{dt}=\sigma^{+}\frac{d\mathbf{w}^{+}}{dt}+\sigma^{-}\frac{d\mathbf{w}^{-}}{dt}=(1+m)\sigma^{-}\frac{d\mathbf{w}^{-}}{dt}=\left(\frac{m+1}{m}\right)\sigma^{+}\frac{d\mathbf{w}^{+}}{dt}.\tag{3.39}$$

From (3.1)–(3.3) and (2.16), we find

$$\frac{1}{\rho}\frac{d\rho}{dt}=\frac{1}{\sigma^{e}}\frac{d\sigma^{e}}{dt}=\frac{1}{\sigma^{+}}\frac{d\sigma^{+}}{dt}=\frac{1}{\sigma^{-}}\frac{d\sigma^{-}}{dt};$$

hence, we may write

$$\rho\frac{d^{2}\pi^{e}}{dt^{2}}=\sigma^{e}\frac{d^{2}\eta}{dt^{2}},\tag{3.40}$$

$$\rho\frac{d^{2}\pi^{I}}{dt^{2}}=\sigma^{+}\frac{d^{2}\mathbf{w}^{+}}{dt^{2}}+\sigma^{-}\frac{d^{2}\mathbf{w}^{-}}{dt^{2}}=(1+m)\sigma^{-}\frac{d^{2}\mathbf{w}^{-}}{dt^{2}}=\left(\frac{m+1}{m}\right)\sigma^{+}\frac{d^{2}\mathbf{w}^{+}}{dt^{2}}.\tag{3.41}$$

Substituting from (3.35), (3.36), and (3.38)–(3.41) into (3.27) and utilizing (2.6), (2.12), (2.16)–(2.19),

(2.21), (2.22), and possibly some other relations in Sec. 2 and (3.23), we obtain

$$\mathbf{E}^M + \left(\frac{m-e}{1+e}\right) r^+ \boldsymbol{\pi}^I \cdot \nabla \mathbf{E}^M + \frac{r^e}{\beta} \boldsymbol{\pi}^e \cdot \nabla \mathbf{E}^M + \frac{\mathbf{v}}{C} \times \mathbf{B}^M + \left(\frac{m-e}{1+e}\right) \frac{r^+}{C} \frac{d\boldsymbol{\pi}^I}{dt} \times \mathbf{B}^M + \frac{r^e}{\beta C} \frac{d\boldsymbol{\pi}^e}{dt} \times \mathbf{B}^M + \left(\frac{m-e}{1+e}\right) \frac{r^+}{C} \mathbf{v} \times (\boldsymbol{\pi}^I \cdot \nabla \mathbf{B}^M) + \frac{r^e}{\beta C} \mathbf{v} \times (\boldsymbol{\pi}^e \cdot \nabla \mathbf{B}^M) + \mathbf{E}^e + \frac{r^e}{\beta} \boldsymbol{\mu}' \cdot \mathbf{B}^M \nabla = r^e \frac{d\mathbf{v}}{dt} + r^+ r^e \left(\frac{m-e}{1+e}\right) \frac{d^2\boldsymbol{\pi}^I}{dt^2} + \frac{(r^e)^2}{\beta} \frac{d^2\boldsymbol{\pi}^e}{dt^2} \tag{3.42}$$

which can properly be called the equation of motion of the electronic polarization, and where

$$\beta = s/(1+s). \tag{3.43}$$

Adding (3.29) and (3.30), substituting from (3.16) and (3.17) and utilizing (2.3), (2.4), (2.8), (2.9), (2.12), (2.14), (2.16), (2.22), (3.35), (3.36), and (3.38)–(3.40), we obtain

$$\nabla \cdot \boldsymbol{\tau} + \mathbf{P} \cdot \nabla \mathbf{E}^M + \frac{\mathbf{v}}{C} \times (\mathbf{P} \cdot \nabla \mathbf{B}^M) + \frac{\rho}{C} \frac{d\boldsymbol{\pi}}{dt} \times \mathbf{B}^M + \mathbf{M}' \cdot \mathbf{B}^M \nabla = \rho \frac{d\mathbf{v}}{dt} + r^e \rho \frac{d^2\boldsymbol{\pi}^e}{dt^2}, \tag{3.44}$$

which are the stress equations of motion of the combined continuum, consisting of the positive and negative ionic continua including the electronic continuum associated with each, and where $\rho \det y_{i,M} = \rho_0$, the initial mass density and

$$\boldsymbol{\tau} = \boldsymbol{\tau}^+ + \boldsymbol{\tau}^-, \quad \mathbf{P} = \mathbf{P}^I + \mathbf{P}^e, \quad \boldsymbol{\pi} = \boldsymbol{\pi}^I + \boldsymbol{\pi}^e, \tag{3.45}$$

where $\boldsymbol{\tau}$ is the total mechanical stress tensor and $\mathbf{P}$ and $\boldsymbol{\pi}$ are the ordinary total polarizations per unit volume and per unit mass, respectively. Now, subtracting $1/m$ of (3.30) from (3.29), substituting from (3.16) and (3.17) and employing (2.3), (2.4), (2.6), (2.8), (2.9), (2.14), (2.16)–(2.19), (2.21), (2.22), (3.23), and (3.35)–(3.41), we obtain

$$m r^+ \nabla \cdot \boldsymbol{\Delta} + \rho \mathbf{E}^M + r^+(m-1) \mathbf{P}^I \cdot \nabla \mathbf{E}^M + \rho \mathbf{E}^L + \frac{\rho}{C} \mathbf{v} \times \mathbf{B}^M + \frac{r^+(m-1)}{C} \mathbf{v} \times (\mathbf{P}^I \cdot \nabla \mathbf{B}^M) + \frac{r^+(m-1)}{C} \rho \frac{d\boldsymbol{\pi}^I}{dt} \times \mathbf{B}^M + r^+ \left(\frac{m-e}{1+e}\right) \mathbf{P}^e \cdot \nabla \mathbf{E}^M + r^+ \left(\frac{m-e}{1+e}\right) \frac{\rho}{C} \frac{d\boldsymbol{\pi}^e}{dt} \times \mathbf{B}^M + \frac{r^+}{C} \left(\frac{m-e}{1+e}\right) \mathbf{v} \times (\mathbf{P}^e \cdot \nabla \mathbf{B}^M) + \frac{r^+(m-r^m)}{(1+r^m)} \mathbf{M}' \cdot \mathbf{B}^M \nabla = m(r^+)^2 \rho \frac{d^2\boldsymbol{\pi}^I}{dt^2} + r^e r^+ \left(\frac{m-e}{1+e}\right) \rho \frac{d^2\boldsymbol{\pi}^e}{dt^2}, \tag{3.46}$$

where

$$\boldsymbol{\Delta} = \boldsymbol{\tau}^+ - \frac{1}{m} \boldsymbol{\tau}^-, \tag{3.47}$$

and $\boldsymbol{\Delta}$ is called the ionic polarization stress tensor. Equation (3.46) can properly be called the equation of motion of the ionic polarization. Adding (3.33) and (3.34) and employing (2.6), (2.14), (2.16)–(2.19), (2.21), (2.22), (3.19), (3.35)–(3.41), (3.45), and (3.47), we obtain

$$\mathbf{e}_l e_{lij} \tau_{ij} + \mathbf{e}_l e_{lkj} m r^+ (\pi_k^I \Delta_{ij})_{,i} + \mathbf{B}^L \times \mathbf{M}' + \mathbf{P} \times \left(\mathbf{E}^M + \frac{\mathbf{v}}{C} \times \mathbf{B}^M\right) + \frac{r^+(m-1)}{C} \rho \boldsymbol{\pi}^I \times \left(\frac{d\boldsymbol{\pi}^I}{dt} \times \mathbf{B}^M\right) + \frac{r^+(m-e)}{C(1+e)} \rho \left[\boldsymbol{\pi}^e \times \left(\frac{d\boldsymbol{\pi}^I}{dt} \times \mathbf{B}^M\right) + \boldsymbol{\pi}^I \times \left(\frac{d\boldsymbol{\pi}^e}{dt} \times \mathbf{B}^M\right)\right] + \frac{r^e}{C\beta} \rho \boldsymbol{\pi}^e \times \left(\frac{d\boldsymbol{\pi}^e}{dt} \times \mathbf{B}^M\right) - m(r^+)^2 \rho \boldsymbol{\pi}^I \times \frac{d^2\boldsymbol{\pi}^I}{dt^2} - r^e \rho \boldsymbol{\pi}^e \times \frac{d\mathbf{v}}{dt} - r^+ r^e \left(\frac{m-e}{1+e}\right) \rho \left(\boldsymbol{\pi}^I \times \frac{d^2\boldsymbol{\pi}^e}{dt^2} + \boldsymbol{\pi}^e \times \frac{d^2\boldsymbol{\pi}^I}{dt^2}\right) - \frac{(r^e)^2}{\beta} \rho \boldsymbol{\pi}^e \times \frac{d^2\boldsymbol{\pi}^e}{dt^2} = 0, \tag{3.48}$$

which is the equation of the conservation of angular momentum for the combined continuum. As can readily be seen, this has been an involved derivation of a rather complicated relation. However, in Sec. 6 we shall show that this rather complicated relation is a straightforward consequence of the principle of material objectivity, which in this instance is satisfied if the stored energy function is invariant in a rigid rotation. As a consequence, the equation of the conservation of angular momentum is not explicitly required because it is satisfied identically by virtue of the abovementioned rotational invariance condition. Moreover, since we are not interested in evaluating the couple of interaction $\mathbf{C}$ between the ionic continua, which couple can readily be determined *a posteriori*, the difference equation of the conservation of angular momenta of the ionic continua is not of interest and will not be presented.

Since the electromagnetic field vectors $\mathbf{E}^M, \mathbf{B}^M, \mathbf{P}$, and $\mathbf{M}'$ appear in the pertinent equations of balance (3.25), (3.42), (3.44), and (3.46), the Maxwell electromagnetic field equations naturally must be included as part of the theory. Consequently, we briefly discuss the electromagnetic field equations for continua in the next section.

4. THE ELECTROMAGNETIC FIELD

As we have already noted, the field vectors $\mathbf{E}^M, \mathbf{B}^M$, $\mathbf{P}$, and $\mathbf{M}'$ must satisfy the Maxwell field equations for electrical insulators, which in Gaussian units take the form[37]

$$C \nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t}, \tag{4.1}$$

$$C \nabla \times \mathbf{E}^M = -\frac{\partial \mathbf{B}^M}{\partial t}, \tag{4.2}$$

where C is the speed of light,

$$\mathbf{D} = \mathbf{E}^M + 4\pi \mathbf{P}, \quad \mathbf{H} = \mathbf{B}^M - 4\pi \mathbf{M} \tag{4.3}$$

and

$$\mathbf{M} = \mathbf{M}' - \frac{1}{C} \mathbf{v} \times \mathbf{P}', \quad \mathbf{P} = \mathbf{P}' + \frac{1}{C} \mathbf{v} \times \mathbf{M}', \tag{4.4}$$

are the low velocity limit of the relativistic transformations[38] from one inertial coordinate system to another. In (4.4) $\mathbf{P}'$ and $\mathbf{M}'$ are the polarization and magnetization, respectively, in the instantaneous local rest system of inertia for the point $\mathbf{y}(\mathbf{X}, t)$ moving with velocity $\mathbf{v}$ relative to *our* rest system of

inertia, and **P** and **M** are the polarization and magnetization, respectively, in *our* rest system of inertia. In addition to (4.1) and (4.2), the auxiliary Maxwell equations

$$\nabla \cdot \mathbf{B}^M = 0, \quad \nabla \cdot \mathbf{D} = 0, \tag{4.5}$$

are satisfied identically. Equations (4.1), (4.2), and (4.5), respectively, may properly be regarded as consequences of the integral forms[39]

$$\oint_c \mathbf{H} \cdot d\mathbf{y} = \frac{1}{C}\frac{\partial}{\partial t}\int_s \mathbf{n} \cdot \mathbf{D} ds, \tag{4.6}$$

$$\oint_c \mathbf{E}^M \cdot d\mathbf{y} = -\frac{1}{C}\frac{\partial}{\partial t}\int_s \mathbf{n} \cdot \mathbf{B}^M ds, \tag{4.7}$$

$$\int_s \mathbf{n} \cdot \mathbf{B}^M dS = 0, \quad \int_S \mathbf{n} \cdot \mathbf{D} dS = 0, \tag{4.8}$$

from which (4.1), (4.2), and (4.5) can be obtained when the field vectors in (4.6)–(4.8) are properly differentiable. However, (4.6)–(4.8) are taken to be valid even when the field vectors are not differentiable and, consequently, (4.1), (4.2), and (4.5) are not valid. In (4.6)–(4.8), c denotes a closed circuit surrounding an open area s and S denotes a closed surface surrounding a volume V, all of which are stationary with respect to *our* inertial reference system. Clearly, Eqs. (4.6)–(4.8) can be used to determine jump conditions on the field vectors across moving surfaces of discontinuity in addition to determining (4.1), (4.2), and (4.5) under appropriate circumstances. Detailed use is made of (4.6)–(4.8) across moving surfaces of discontinuity in Sec. 7.

In view of Eq. (3.44), which is the conservation of linear momentum for the combined material continuum, there is an electromagnetic force relation that can be derived with the aid of (4.1)–(4.5) and used to obtain an integral form from (3.44), which can be used to determine jump conditions on traction across moving surfaces of discontinuity. We now proceed to derive this integral form. To this end we consider the electromagnetic body force term in (3.44), which resulted from our model and in indicial notation takes the form

$$f_j = P_i E^M_{j,i} + M'_i B^M_{i,j} + \frac{1}{C} e_{jkl} v_k P_i B^M_{l,i} + \frac{1}{C} e_{jkl} \rho \frac{d\pi_k}{dt} B^M_l, \tag{4.9}$$

which with the aid of the relations

$$P_i = \rho \pi_i, \quad \frac{d\rho}{dt} = -\rho v_{k,k}, \tag{4.10}$$

$$\frac{dP_i}{dt} = \frac{\partial P_i}{\partial t} + v_k P_{i,k} \tag{4.11}$$

can be written in the form

$$f_j = P_i E^M_{j,i} + M'_i B^M_{i,j} + \frac{1}{C} e_{jkl} v_k P_i B^M_{l,i} + \frac{1}{C} e_{jil} \frac{\partial P_i}{\partial t} B^M_l + \frac{1}{C} e_{jkl} (P_k v_i)_{,i} B^M_l. \tag{4.12}$$

From (4.1), (4.3), and (4.4), introducing indicial notation, we obtain

$$\frac{\partial P_i}{\partial t} = \frac{C}{4\pi} e_{irk} (B^M_k - 4\pi M'_k)_{,r} + (v_i P'_r - v_r P'_i)_{,r} - \frac{1}{4\pi}\frac{\partial E^M_i}{\partial t}, \tag{4.13}$$

which, with the aid of (4.2), enables us to write

$$\frac{1}{C} e_{jil} \frac{\partial P_i}{\partial t} B^M_l = \frac{1}{4\pi}[(B^M_j - 4\pi M'_j)_{,l} - (B^M_l - 4\pi M'_l)_{,j}] B^M_l + \frac{1}{C} e_{jil} (v_i P'_r - v_r P'_i)_{,r} B^M_l - \frac{\partial}{\partial t}\left(\frac{e_{jil}}{4\pi C} E^M_i B^M_l\right) - \frac{E^M_i}{4\pi}(E^M_{i,j} - E^M_{j,i}). \tag{4.14}$$

Substituting from (4.14) into (4.12) and employing (4.3), (4.4), and the second equation of (4.5) and neglecting terms in $v_k v_j / C^2$, we obtain

$$f_j = \frac{1}{4\pi}[4\pi P_i E'^M_j + E^M_i E^M_j + B^M_i B^M_j - 4\pi B^M_i M'_j - \tfrac{1}{2}(E^M_k E^M_k + B^M_k B^M_k - 8\pi M'_k B^M_k)\delta_{ij}]_{,i} - \frac{\partial}{\partial t}\left(\frac{e_{jil}}{4\pi C} E^M_i B^M_l\right), \tag{4.15}$$

where

$$E'^M_j = E^M_j + \frac{1}{C} e_{jkl} v_k B^M_l, \tag{4.16}$$

and $(e_{jil} E^M_i B^M_l / 4\pi C)$ is Livens'[40] expression for the linear momentum of the electromagnetic field, and

$$T^{EM}_{ij} = \frac{1}{4\pi}[4\pi P_i E'^M_j + E^M_i E^M_j + B^M_i B^M_j - 4\pi B^M_i M'_j - \tfrac{1}{2}(E^M_k E^M_k + B^M_k B^M_k - 8\pi M'_k B^M_k)\delta_{ij}], \tag{4.17}$$

is the Maxwell electromagnetic stress tensor for our polarizable and magnetizable deformable dielectric continuum.

When the dielectric is rigid and at rest with respect to our inertial reference system, $\mathbf{v} \equiv 0$, $\mathbf{E}^M \equiv \mathbf{E}'^M$, and $\mathbf{M}' \equiv \mathbf{M}$, and the force relation (4.15), and the expression for the Maxwell stress tensor (4.17), reduce to those of Livens.[40] Consequently, these electromagnetic force relations for a *deformable* continuum can be considered to be a generalization of the force relations of Livens[40] for the *rigid* continuum. It should be carefully noted that these electromagnetic force relations are valid only to terms linear in $\mathbf{v}/C$ because terms quadratic and higher in $\mathbf{v}/C$ have been neglected in the derivation of these relations.

Substituting from (4.9), (4.15), and (4.17) into (3.44), integrating over an arbitrary material region and employing the divergence theorem, the transport theorem[41] and (3.3), we obtain the integral form

$$\int_S \mathbf{n} \cdot (\tau + \mathbf{T}^{EM} + \mathbf{vg}) dS = \frac{d}{dt}\int_V \left(\rho \mathbf{v} + r^e \rho \frac{d\pi^e}{dt} + \mathbf{g}\right) dV, \tag{4.18}$$

where

$$\mathbf{g} = \mathbf{E}^M \times \mathbf{B}^M / 4\pi C \tag{4.19}$$

is the electromagnetic momentum in Gaussian units. When the dependent variables are properly differentiable, the differential form (3.44) can readily be obtained from (4.18). However, the integral form (4.18) is taken to be valid even when the field variables are not differentiable and the differential form (3.44) cannot be obtained, such as across surfaces of discontinuity. Equation (4.18) is applied across moving surfaces of discontinuity in Sec. 7 in the determination of jump conditions on traction. When all the appropriate substitutions are made, Eqs. (3.25), (3.42), (3.44), (3.46), (4.1), and (4.2) constitute an

underdetermined system and constitutive equations are required in order to obtain a determinate system, as usual with continuum descriptions. To this end we consider the conservation of energy for the material continuum in the next section.

5. THERMODYNAMIC CONSIDERATIONS

The conservation of energy for the combined material continuum, which consists of the positive and negative ionic continua including the electronic continuum associated with each, can be written in the form

$$\frac{d}{dt}\int_V (T+\rho\epsilon)\,dV = \int_S \left[\mathbf{t}^+\cdot\left(\mathbf{v}+\frac{d\mathbf{w}^+}{dt}\right)+\mathbf{t}^-\cdot\left(\mathbf{v}+\frac{d\mathbf{w}^-}{dt}\right) + \mathbf{F}\cdot\rho\frac{d\boldsymbol{\mu}'}{dt}-\mathbf{n}\cdot\mathbf{q}\right]dS + \int_V \Sigma dV, \tag{5.1}$$

where T is the kinetic energy per unit volume, ϵ is the internal stored energy per unit mass, $\mathbf{t}^\pm\cdot(\mathbf{v}+d\mathbf{w}^\pm/dt)$ denote the rates of working per unit area of the mechanical surface tractions acting in the positive and negative ionic continua, respectively, $\mathbf{F}\cdot\rho d\boldsymbol{\mu}'/dt$ is the rate of working per unit area of the surface exchange torques,[42] $\mathbf{n}\cdot\mathbf{q}$ is the rate of efflux of heat per unit area and Σ is the rate of supply of energy to the entire material continuum from the electromagnetic field. In order to obtain expressions for T and Σ, we must return to our model of the combined material continuum.

From the model of the continuum it is clear that the kinetic energy per unit volume is of the form

$$\begin{aligned}T = \tfrac{1}{2}\Big[&\rho^{l+}\left(\mathbf{v}+\frac{d\mathbf{w}^+}{dt}\right)\cdot\left(\mathbf{v}+\frac{d\mathbf{w}^+}{dt}\right)\\ &+\rho^{e+}\left(\mathbf{v}+\frac{d\mathbf{w}^+}{dt}+\frac{d\boldsymbol{\eta}}{dt}\right)\cdot\left(\mathbf{v}+\frac{d\mathbf{w}^+}{dt}+\frac{d\boldsymbol{\eta}}{dt}\right)\\ &+\rho^{l-}\left(\mathbf{v}+\frac{d\mathbf{w}^-}{dt}\right)\cdot\left(\mathbf{v}+\frac{d\mathbf{w}^-}{dt}\right)\\ &+\rho^{e-}\left(\mathbf{v}+\frac{d\mathbf{w}^-}{dt}+\frac{d\boldsymbol{\eta}}{dt}\right)\cdot\left(\mathbf{v}+\frac{d\mathbf{w}^-}{dt}+\frac{d\boldsymbol{\eta}}{dt}\right)\Big],\end{aligned} \tag{5.2}$$

and the kinetic energy associated with the spin angular momentum has been omitted since its material time derivative vanishes on account of (3.3) and (3.24). Expanding terms in (5.2) and employing (2.8)–(2.12), (2.14), (2.16)–(2.19), (2.22), (3.38), (3.39), and (3.43), we obtain

$$T = \frac{\rho}{2}\left[\mathbf{v}\cdot\mathbf{v}+m(r^+)^2\frac{d\boldsymbol{\pi}^I}{dt}\cdot\frac{d\boldsymbol{\pi}^I}{dt}+\frac{(r^e)^2}{\beta}\frac{d\boldsymbol{\pi}^e}{dt}\cdot\frac{d\boldsymbol{\pi}^e}{dt} + 2r^e\mathbf{v}\cdot\frac{d\boldsymbol{\pi}^e}{dt}+2r^e r^+\left(\frac{m-e}{1+e}\right)\frac{d\boldsymbol{\pi}^e}{dt}\cdot\frac{d\boldsymbol{\pi}^I}{dt}\right]. \tag{5.3}$$

From the fundamental charge and spin model of the continua, the rate R^M at which work is done on the matter by the Maxwell electric field $\mathbf{E}^M$ is of the form

$$\begin{aligned}R^M = &\sigma^{l+}\left(\mathbf{v}+\frac{d\mathbf{w}^+}{dt}\right)\cdot\left(\mathbf{E}^M+\mathbf{w}^+\cdot\nabla\mathbf{E}^M\right)\\ &+\sigma^{e+}\left(\mathbf{v}+\frac{d\mathbf{w}^+}{dt}+\frac{d\boldsymbol{\eta}}{dt}\right)\cdot\left(\mathbf{E}^M+\mathbf{w}^+\cdot\nabla\mathbf{E}^M+\boldsymbol{\eta}\cdot\nabla\mathbf{E}^M\right)\\ &+\sigma^{l-}\left(\mathbf{v}+\frac{d\mathbf{w}^-}{dt}\right)\cdot\left(\mathbf{E}^M+\mathbf{w}^-\cdot\nabla\mathbf{E}^M\right)\\ &+\sigma^{e-}\left(\mathbf{v}+\frac{d\mathbf{w}^-}{dt}+\frac{d\boldsymbol{\eta}}{dt}\right)\cdot\left(\mathbf{E}^M+\mathbf{w}^-\cdot\nabla\mathbf{E}^M+\boldsymbol{\eta}\cdot\nabla\mathbf{E}^M\right)\\ &+C\oint_{C'} i'd\mathbf{s}\cdot\mathbf{E}^M,\end{aligned} \tag{5.4}$$

and the rate at which work is done by the magnetic induction field $\mathbf{B}^M$ is zero in general because fundamentally the force exerted on any moving charge element by $\mathbf{B}^M$ is always normal to the instantaneous total velocity of that charge element. Consequently, the rate at which work is done on the matter by $\mathbf{E}^M$ is identically the rate of supply of energy to the matter from the electromagnetic field, and we have

$$\Sigma \equiv R^M. \tag{5.5}$$

The constant C appears in the last term of (5.4) because i' is in magnetic units. The last term in (5.4), which represents the rate at which work is done by $\mathbf{E}^M$ on the circulating conduction current loops $\oint_{C'} i'd\mathbf{s}$, may, with the aid of (4.2), be written in the form

$$C\oint_{C'} i'd\mathbf{s}\cdot\mathbf{E}^M = C\int_{S'} i'\mathbf{n}\cdot\nabla\times\mathbf{E}^M dS = -\int_{S'} i'\mathbf{n}\cdot\frac{\partial\mathbf{B}^M}{\partial t}dS,$$

which, with (3.7) in the limit, yields

$$\lim_{\substack{i'\to\infty\\ C'\to 0 \text{ in a plane}}} C\oint_{C'} i'd\mathbf{s}\cdot\mathbf{E}^M = -\lim_{\substack{i'\to\infty\\ S_0'\to 0\\ \mathbf{n}\text{ fixed}}} i'\int_{S_0'}\mathbf{n}\cdot\frac{\partial\mathbf{B}^M}{\partial t}dS = -\mathbf{M}'\cdot\frac{\partial\mathbf{B}^M}{\partial t}. \tag{5.6}$$

In the magnetic terms in (5.1) and (5.4) the notation for the + and − continua has been dispensed with and the sum of both used directly. Substituting from (2.3), (2.4), (2.6), (3.35), (3.36), (3.38), (3.39), the second and third equation of (3.45), and (5.6) into (5.4), and neglecting products of $\mathbf{w}^\pm$ and $\boldsymbol{\eta}$, we obtain

$$\Sigma = \mathbf{E}^M\cdot\rho\frac{d\boldsymbol{\pi}}{dt}+\mathbf{P}\cdot\nabla\mathbf{E}^M\cdot\mathbf{v}-\mathbf{M}'\cdot\frac{\partial\mathbf{B}^M}{\partial t}. \tag{5.7}$$

Substituting from (2.16), (2.17), (2.22), (3.39), (5.3), and (5.7) into (5.1), we obtain

$$\begin{aligned}\frac{d}{dt}\int_V\rho\Big[&\tfrac{1}{2}\mathbf{v}\cdot\mathbf{v}+\frac{m}{2}(r^+)^2\frac{d\boldsymbol{\pi}^I}{dt}\cdot\frac{d\boldsymbol{\pi}^I}{dt}+\frac{(r^e)^2}{2\beta}\frac{d\boldsymbol{\pi}^e}{dt}\cdot\frac{d\boldsymbol{\pi}^e}{dt}\\ &+r^e\mathbf{v}\cdot\frac{d\boldsymbol{\pi}^e}{dt}+r^e r^+\frac{m-e}{1+e}\frac{d\boldsymbol{\pi}^e}{dt}\cdot\frac{d\boldsymbol{\pi}^I}{dt}+\epsilon\Big]dV\\ &=\int_S\left(\mathbf{t}\cdot\mathbf{v}+mr^+\mathbf{d}\cdot\frac{d\boldsymbol{\pi}^I}{dt}+\mathbf{F}\cdot\rho\frac{d\boldsymbol{\mu}'}{dt}-\mathbf{n}\cdot\mathbf{q}\right)dS\\ &+\int_V\left(\mathbf{E}^M\cdot\rho\frac{d\boldsymbol{\pi}}{dt}+\mathbf{P}\cdot\nabla\mathbf{E}^M\cdot\mathbf{v}-\mathbf{M}'\cdot\frac{\partial\mathbf{B}^M}{\partial t}\right)dV,\end{aligned} \tag{5.8}$$

where

$$\mathbf{t}=\mathbf{t}^+ + \mathbf{t}^-, \qquad \mathbf{d}=\mathbf{t}^+-\frac{1}{m}\mathbf{t}^-, \tag{5.9}$$

and from (3.28), (3.45), and (3.47), we have

$$\mathbf{t}=\mathbf{n}\cdot\boldsymbol{\tau}, \qquad \mathbf{d}=\mathbf{n}\cdot\boldsymbol{\Delta}, \tag{5.10}$$

and $\mathbf{t}$ is the mechanical traction vector of the ionic material continuum and $\mathbf{d}$ may be thought of as an ionic polarization traction vector acting across neighboring surfaces of the ionic continuum. Taking the material time derivative in (5.8) and using (3.3), substituting from (3.21), (3.45), and (5.10), employing the divergence theorem (3.24), (3.25), (3.42), (3.44),

(3.46), and the arbitrariness of V, we obtain

$$\rho\frac{d\epsilon}{dt}=\tau_{ij}v_{j,i}-\rho\mu_i'B^M_{i,j}v_j+m\gamma^+\Delta_{ij}\left(\frac{d\pi_j^I}{dt}\right)_{,i}-A_{ij}\rho\left(\frac{d\mu_j'}{dt}\right)_{,i}-\mathcal{E}_j^I\rho\frac{d\pi_j^I}{dt}-\mathcal{E}_j^e\rho\frac{d\pi_j^e}{dt}-(B_j^L+B_j^M)\rho\frac{d\mu_j'}{dt}-\rho\mu_i'\frac{\partial B_i^M}{\partial t}-q_{i,i},\tag{5.11}$$

where we have introduced indicial notation and

$$\mathcal{E}_j^I=\gamma^+(m-1)\pi_k^IE^M_{j,k}+E_j^L+\frac{\gamma^+}{C}(m-1)e_{jkl}v_k\pi_m^IB^M_{l,m}+\gamma^+\left(\frac{m-e}{1+e}\right)\pi_k^eE^M_{j,k}+\frac{\gamma^+}{C}\left(\frac{m-e}{1+e}\right)e_{jkl}v_k\pi_m^eB^M_{l,m}+\frac{\gamma^+(m-\gamma^m)}{(1+\gamma^m)}\mu_k'B^M_{k,j},\tag{5.12}$$

$$\mathcal{E}_j^e=\gamma^+\left(\frac{m-e}{1+e}\right)\pi_k^IE^M_{j,k}+\pi_k^eE^M_{j,k}+\frac{\gamma^e}{\beta C}e_{jkl}v_k\pi_m^eB^M_{l,m}+\frac{\gamma^+}{C}\frac{m-e}{1+e}e_{jkl}v_k\pi_mB^M_{l,m}+E_j^e+\frac{\gamma^e}{\beta}\mu_k'B^M_{k,j}.\tag{5.13}$$

Employing (4.11), with P_i replaced by B_i^M, and defining

$$\chi=\epsilon+B_i^M\mu_i',\tag{5.14}$$

$$\mathfrak{D}_{ij}=m\gamma^+\Delta_{ij},\tag{5.15}$$

we may write (5.11) in the form

$$\rho\frac{d\chi}{dt}=\tau_{ij}v_{j,i}-B_i^L\rho\frac{d\mu_i'}{dt}-\mathcal{E}_i^I\rho\frac{d\pi_i^I}{dt}-\mathcal{E}_i^e\rho\frac{d\pi_i^e}{dt}-A_{ij}\rho\left(\frac{d\mu_j'}{dt}\right)_{,i}+\mathfrak{D}_{ij}\left(\frac{d\pi_j^I}{dt}\right)_{,i}-q_{i,i},\tag{5.16}$$

which is the first law of thermodynamics for our combined continuum.

Since we are considering a heat conducting, polarizable and magnetizable, deformable ionic continuum without viscous type dissipation and with a first law of thermodynamics of the form shown in (5.16), the mathematical expression of the second law of thermodynamics may be written in the form[43–45]

$$\rho\frac{d\chi}{dt}-\tau_{ij}v_{j,i}+B_i^L\rho\frac{d\mu_i'}{dt}+\mathcal{E}_i^I\rho\frac{d\pi_i^I}{dt}+\mathcal{E}_i^e\rho\frac{d\pi_i^e}{dt}+A_{ij}\rho\left(\frac{d\mu_j'}{dt}\right)_{,i}-\mathfrak{D}_{ij}\left(\frac{d\pi_j^I}{dt}\right)_{,i}=\rho\theta\frac{d\eta}{dt},\tag{5.17}$$

where θ is the positive absolute temperature and η is the entropy per unit mass. From (5.16) and (5.17), we have the dissipation equation

$$-q_{i,i}=\rho\theta\frac{d\eta}{dt},\tag{5.18}$$

and the entropy inequality may be written in the form

$$\rho\frac{d\eta}{dt}+\left(\frac{q_i}{\theta}\right)_{,i}=-\frac{q_i\theta_{,i}}{\theta^2}=\rho\Gamma\geq 0,\tag{5.19}$$

where Γ is the (positive) rate of entropy production. At this point it should be noted that this continuum theory can readily be *generalized*[46,47] to account for arbitrary functional (viscous) constitutive response in the manner set forth in a previous paper.[11]

Before proceeding to a determination of the constitutive equations, we wish to write the conservation of energy (5.8), in a particularly interesting integral form in which no volume source terms appear. To this end we consider the expression (5.7) for Σ, which, with the aid of (4.10) and (4.11), can be written in the form

$$\Sigma=E_i^M\frac{\partial P_i}{\partial t}+(v_kP_i)_{,k}E_i^M+P_iE^M_{k,i}v_k-M_i'\frac{\partial B_i^M}{\partial t}.\tag{5.20}$$

Substituting from (4.2) and (4.4) into (5.20) and neglecting terms in v_jv_k/C^2, we obtain

$$\Sigma=E_i^M\frac{\partial P_i}{\partial t}-M_i\frac{\partial B_i^M}{\partial t}+(v_kP_iE_i^M)_{,k}.\tag{5.21}$$

From (4.1) and (4.2) in the usual way,[48] with the aid of (4.3), we obtain

$$-\left(\frac{C}{4\pi}e_{ijk}E_j^MH_k\right)_{,i}=\frac{\partial U^F}{\partial t}+E_i^M\frac{\partial P_i}{\partial t}-M_i\frac{\partial B_i^M}{\partial t},\tag{5.22}$$

where

$$U^F=\frac{1}{8\pi}(E_k^ME_k^M+B_k^MB_k^M)\tag{5.23}$$

may be interpreted as the free-space electromagnetic field energy. Equation (5.22) is a particularly interesting and useful differential form of Poynting's theorem for nonconducting continua.[49] Since the form in (5.22) depends only on the validity of Maxwell's equations [Eqs. (4.1) and (4.2)] and the relations (4.3), and is independent of any particular constitutive assumption, it is always valid. Substituting from (5.3), (5.21), and (5.22) into (5.1), and employing the divergence theorem and the transport theorem,[41] we obtain

$$\frac{d}{dt}\int_V\left[\frac{\rho}{2}\left(v_kv_k+m(\gamma^+)^2\frac{d\pi_k^I}{dt}\frac{d\pi_k^I}{dt}+\frac{(\gamma^e)^2}{\beta}\frac{d\pi_k^e}{dt}\frac{d\pi_k^e}{dt}+2\gamma^ev_k\frac{d\pi_k^e}{dt}+2\gamma^e\gamma^+\frac{(m-e)}{1+e}\frac{d\pi_k^e}{dt}\frac{d\pi_k^I}{dt}\right)+\rho\epsilon+U^F\right]dV=\int_S\left[t_jv_j+m\gamma^+d_j\frac{d\pi_j^I}{dt}+F_j\rho\frac{d\mu_j'}{dt}-n_iq_i-n_i\frac{C}{4\pi}e_{ijk}E_j^MH_k+n_kv_kP_iE_i^M+n_kv_kU^F\right]dS,\tag{5.24}$$

which is the particularly interesting integral form of the equation of the conservation of energy we have been after. As with all such forms, Eq. (5.24) is taken to be valid across moving surfaces of discontinuity even when (5.1) is not meaningful, and (5.24) reduces to (5.1) [or (5.8)] when $\mathbf{P}$, $\mathbf{M}$, $\mathbf{E}^M$, and $\mathbf{B}^M$ are appropriately differentiable. Clearly, Eq. (5.24) enables the determination of jump conditions on the energy. Moreover, Eq. (5.24), which is a consequence of our model, says that the material time rate of change of kinetic plus stored internal plus electromagnetic field energy is equal to the rate at which work is done by the mechanical surface tractions, magnetic exchange torques and ionic polarization tractions acting across S less the flux of thermal and electromagnetic Poynting energy outward across S plus a convective flux of electromagnetic field energy and electric field-electric polarization interaction energy.

6. CONSTITUTIVE EQUATIONS

Since we are concerned with thermodynamic processes for which both the state function equation (5.17) and the dissipation equation (5.18) are valid, we may determine the heat conducting constitutive equation

from (5.19) and the remaining constitutive equations from (5.17), which, by virtue of the relations

$$v_{j,i} = X_{M,i}\frac{d}{dt}(y_{j,M}), \quad \left(\frac{d\mu_j'}{dt}\right)_{,i} = X_{M,i}\frac{d}{dt}(\mu_{j,M}'), \quad \left(\frac{d\pi_j^I}{dt}\right)_{,i} = X_{M,i}\frac{d}{dt}(\pi_{j,M}^I),$$

may be written in the form

$$\rho\frac{d\chi}{dt} = \tau_{ij}X_{M,i}\frac{d}{dt}(y_{j,M}) - \rho B_j^L\frac{d\mu_j'}{dt} - \rho\mathcal{E}_j^I\frac{d\pi_j^I}{dt} - \rho\mathcal{E}_j^e\frac{d\pi_j^e}{dt} - \rho A_{ij}X_{M,i}\frac{d}{dt}(\mu_{j,M}') + \mathfrak{D}_{ij}X_{M,i}\frac{d}{dt}(\pi_{j,M}^I) + \rho\theta\frac{d\eta}{dt}. \tag{6.1}$$

Since the entropy inequality is of the form shown in (5.19), it turns out to be convenient to define the thermodynamic function ψ by the Legendre transformation

$$\psi = \chi - \eta\theta. \tag{6.2}$$

The substitution of the material time derivative of (6.2) into (6.1) yields

$$\rho\frac{d\psi}{dt} = \tau_{ij}X_{M,i}\frac{d}{dt}(y_{j,M}) - \rho B_j^L\frac{d\mu_j'}{dt} - \rho\mathcal{E}_j^I\frac{d\pi_j^I}{dt} - \rho\mathcal{E}_j^e\frac{d\pi_j^e}{dt} - \rho A_{ij}X_{M,i}\frac{d}{dt}(\mu_{j,M}') + \mathfrak{D}_{ij}X_{M,i}\frac{d}{dt}(\pi_{j,M}^I) - \rho\eta\frac{d\theta}{dt}. \tag{6.3}$$

Motivated by (6.3), we assume

$$\psi = \psi(y_{j,M}; \mu_j'; \pi_j^I; \pi_j^e; \mu_{j,M}'; \pi_{j,M}^I; \theta), \tag{6.4}$$

whence

$$\frac{d\psi}{dt} = \frac{\partial\psi}{\partial(y_{j,M})}\frac{d}{dt}(y_{j,M}) + \frac{\partial\psi}{\partial\mu_j'}\frac{d\mu_j'}{dt} + \frac{\partial\psi}{\partial\pi_j^I}\frac{d\pi_j^I}{dt} + \frac{\partial\psi}{\partial\pi_j^e}\frac{d\pi_j^e}{dt} + \frac{\partial\psi}{\partial(\mu_{j,M}')}\frac{d}{dt}(\mu_{j,M}') + \frac{\partial\psi}{\partial(\pi_{j,M}^I)}\frac{d}{dt}(\pi_{j,M}^I) + \frac{\partial\psi}{\partial\theta}\frac{d\theta}{dt}. \tag{6.5}$$

At this point we must recall that 12 of the 37 time derivatives appearing on the rhs of (6.3) and (6.5) are not independent, but on account of (3.24) are connected by the four relations

$$\mu_j'\frac{d\mu_j'}{dt} = 0, \quad \mu_j'\frac{d}{dt}(\mu_{j,M}') + \mu_{j,M}'\frac{d\mu_j'}{dt} = 0. \tag{6.6}$$

Consequently, we must introduce four Lagrangian undetermined multipliers λ and L_M, then multiply $(6.6)_1$ by λ and $(6.6)_2$ by L_M and add the sum to the rhs of (6.3) while substituting from (6.5) to obtain

$$\left(\tau_{ij}X_{M,i} - \rho\frac{\partial\psi}{\partial(y_{j,M})}\right)\frac{d}{dt}(y_{j,M}) - \rho\left(B_j^L - \lambda\mu_j' - L_M\mu_{j,M}' + \frac{\partial\psi}{\partial\mu_j'}\right)\frac{d\mu_j'}{dt} - \rho\left(\mathcal{E}_j^I + \frac{\partial\psi}{\partial\pi_j^I}\right)\frac{d\pi_j^I}{dt} - \rho\left(\mathcal{E}_j^e + \frac{\partial\psi}{\partial\pi_j^e}\right)\frac{d\pi_j^e}{dt} - \rho\left(A_{ij}X_{M,i} - L_M\mu_j' + \frac{\partial\psi}{\partial(\mu_{j,M}')}\right)\frac{d}{dt}(\mu_{j,M}') + \left(\mathfrak{D}_{ij}X_{M,i} - \rho\frac{\partial\psi}{\partial(\pi_{j,M}^I)}\right)\frac{d}{dt}(\pi_{j,M}^I) - \rho\left(\eta + \frac{\partial\psi}{\partial\theta}\right)\frac{d\theta}{dt} = 0. \tag{6.7}$$

Since we have introduced the proper number of undetermined multipliers in the usual Lagrangian manner, we may treat all 37 time derivatives as if they were independent; and since (6.7) holds for arbitrary $d(y_{j,M})/dt, d\mu_j'/dt, d\pi_j^I/dt, d\pi_j^e/dt, d(\mu_{j,M}')/dt, d(\pi_{j,M}^I)/dt,$ and $d\theta/dt$, we have

$$X_{M,i}\tau_{ij} = \rho\frac{\partial\psi}{\partial(y_{j,M})}, \tag{6.8}$$

$$B_j^L = -\frac{\partial\psi}{\partial\mu_j'} + L_M\mu_{j,M}' + \lambda\mu_j', \tag{6.9}$$

$$\mathcal{E}_j^I = -\frac{\partial\psi}{\partial\pi_j^I}, \tag{6.10}$$

$$\mathcal{E}_j^e = -\frac{\partial\psi}{\partial\pi_j^e}, \tag{6.11}$$

$$X_{M,i}A_{ij} = -\frac{\partial\psi}{\partial(\mu_{j,M}')} + L_M\mu_j' \tag{6.12}$$

$$X_{M,i}\mathfrak{D}_{ij} = \rho\frac{\partial\psi}{\partial(\pi_{j,M}^I)}, \tag{6.13}$$

$$\eta = -\frac{\partial\psi}{\partial\theta}. \tag{6.14}$$

Note that by virtue of the definition of B_j^L, without loss of generality, we may take

$$\mu_j'B_j^L = 0. \tag{6.15}$$

The Lagrangian multipliers may be determined by substituting from (6.9) and (6.12), respectively, into (6.15) and (3.22), with the results

$$\lambda = \frac{1}{(\mu_s')^2}\frac{\partial\psi}{\partial\mu_k'}\mu_k', \quad L_M = \frac{1}{(\mu_s')^2}\frac{\partial\psi}{\partial(\mu_{k,M}')}\mu_k'. \tag{6.16}$$

Substituting from (6.16) into (6.9) and (6.12) and solving (6.8), (6.12), and (6.13), respectively, for τ, A, and $\mathfrak{D}$, we find

$$\tau_{ij} = \rho y_{i,M}\frac{\partial\psi}{\partial(y_{j,M})}, \tag{6.17}$$

$$B_j^L = -\frac{\partial\psi}{\partial\mu_j'} + \frac{1}{(\mu_s')^2}\left(\frac{\partial\psi}{\partial\mu_k'}\mu_k'\mu_j' + \frac{\partial\psi}{\partial(\mu_{k,M}')}\mu_k'\mu_{j,M}'\right), \tag{6.18}$$

$$A_{ij} = -y_{i,M}\left(\frac{\partial\psi}{\partial(\mu_{j,M}')} - \frac{1}{(\mu_s')^2}\frac{\partial\psi}{\partial(\mu_{k,M}')}\mu_k'\mu_j'\right), \tag{6.19}$$

$$\mathfrak{D}_{ij} = \rho y_{i,M}\frac{\partial\psi}{\partial(\pi_{j,M}^I)}. \tag{6.20}$$

Clearly, ψ cannot be an *arbitrary* function of the variables shown in (6.4) because, in order to satisfy the principle of material objectivity,[26,27] ϵ and, hence, χ and ψ must be scalar invariants under rigid rotations[50] of the deformed, polarized and magnetized body, and any arbitrary function of the 37 assumed variables (12 vectors and a scalar at the point y_k) will not be so invariant. However, there is a theorem on rotationally invariant functions of several vectors due to Cauchy,[51] which says that ψ may be an arbitrary single-valued function of the scalar products of the vectors and the determinants of their components taken three at a time. Application of this theorem shows that ψ is expressible as an arbitrary function of 78 scalar products and 194 determinants and θ for a total of 273 quantities. However, the 273 quantities

are not all functionally independent and it can be shown, by using procedures similar to those employed in Sec. 6 of Ref. 4, that the 273 variables are expressible in terms of the 34 arguments

$$C_{KL} = y_{i,K}y_{i,L}, \quad N_L = y_{k,L}\mu'_k, \quad W^I_L = y_{k,L}\pi^I_k, \quad (6.21)$$
$$W^e_L = y_{k,L}\pi^e_k, \quad K_{LM} = \mu'_{k,L}y_{k,M}, \quad Q_{LM} = y_{k,L}\pi^I_{k,M}, \quad \theta.$$

Thus, we find that ψ is invariant in a rigid rotation if it is a single-valued function of the 34 arguments listed in (6.21). Now, as in Ref. 4, we must recognize that any single-valued function ψ of the 34 arguments listed in (6.21) will not necessarily satisfy (3.26) by virtue of (6.19). Equations (3.26) comprise a system of three independent differential equations in the 34 variables listed in (6.21), which must be satisfied by ψ. Consequently, ψ must reduce to an arbitrary function of any $34 - 3 = 31$ functionally independent solutions of (3.26), which must be composed of $\mathbf{C}, \mathbf{N}, \mathbf{W}^I, \mathbf{W}^e, \mathbf{K}, \mathbf{Q}$, and θ. Clearly, $\mathbf{C}, \mathbf{N}, \mathbf{W}^I, \mathbf{W}^e, \mathbf{Q}$, and θ constitute 25 such solutions and six additional solutions are given by

$$K_{LK}C^{-1}_{KN}K_{MN} \equiv \mu'_{i,L}\mu'_{i,M} = G_{LM}, \quad (6.22)$$

as may be verified by following the procedure employed in Sec. 6 of Ref. 4. Thus we find that ψ may be reduced to the form

$$\psi = \psi(E_{KL}, N_L, W^I_L, W^e_L, G_{LM}, Q_{LM}, \theta), \quad (6.23)$$

in place of the form shown in (6.4), and where we have taken the liberty of replacing Green's deformation tensor C_{KL}, which does not vanish in the undeformed state, by the equivalent material strain tensor E_{KL}, which does vanish in the undeformed state, and is related to C_{KL} by

$$E_{KL} = \tfrac{1}{2}(C_{KL} - \delta_{KL}). \quad (6.24)$$

It is interesting to note that $\mathbf{G}$ is invariant in a rigid rotation of the entire spin continuum with respect to the lattice continuum. Thus it is clear that (3.26) has served to make the exchange energy invariant in a rigid rotation of the entire spin system as it is in the quantum mechanical description.[52]

From (6.10), (6.11), (6.14), and (6.17)–(6.24), we obtain

$$\tau_{ij} = \rho y_{i,L}y_{j,M}\frac{\partial\psi}{\partial E_{LM}} + \rho y_{i,L}\frac{\partial\psi}{\partial N_L}\mu'_j + \rho y_{i,L}\frac{\partial\psi}{\partial W^I_L}\pi^I_j + \rho y_{i,L}\frac{\partial\psi}{\partial W^e_L}\pi^e_j + \rho y_{i,L}\frac{\partial\psi}{\partial Q_{LM}}\pi^I_{j,M}, \quad (6.25)$$

$$B^L_j = -y_{j,L}\frac{\partial\psi}{\partial N_L} + \frac{1}{(\mu'_s)^2}y_{k,L}\mu'_k\mu'_j\frac{\partial\psi}{\partial N_L}, \quad (6.26)$$

$$\mathcal{E}^I_j = -y_{j,L}\frac{\partial\psi}{\partial W^I_L}, \quad (6.27)$$

$$\mathcal{E}^e_j = -y_{j,L}\frac{\partial\psi}{\partial W^e_L}, \quad (6.28)$$

$$A_{ij} = -2y_{i,M}\mu'_{j,L}\frac{\partial\psi}{\partial G_{ML}}, \quad (6.29)$$

$$\mathfrak{D}_{ij} = \rho y_{i,M}y_{j,L}\frac{\partial\psi}{\partial Q_{LM}}, \quad (6.30)$$

$$\eta = -\frac{\partial\psi}{\partial\theta}, \quad (6.31)$$

where we have introduced the conventions $\partial\psi/\partial E_{LM} = \partial\psi/\partial E_{ML}$ and $\partial\psi/\partial G_{LM} = \partial\psi/\partial G_{ML}$ and it is to be assumed that $\partial E_{KL}/\partial E_{LK} = 0$ and $\partial G_{KL}/\partial G_{LK} = 0$ in differentiating ψ and we have found that

$$L_M = \frac{1}{(\mu'_s)^2}\frac{\partial\psi}{\partial\mu'_{k,M}}\mu'_k = \frac{2}{(\mu'_s)^2}\frac{\partial\psi}{\partial G_{RM}}\mu'_{k,R}\mu'_k = 0. \quad (6.32)$$

At this point it should be noted that since we have constitutive equations for $\mathcal{E}^I_j$ and $\mathcal{E}^e_j$, which are composed of the E^L_j and E^e_j, respectively, plus other terms, the equations of motion (3.42) and (3.46) of the electronic and ionic polarizations, respectively, may be written in the more convenient forms

$$\mathcal{E}^e_j + E^M_j + \frac{1}{C}e_{jkl}v_kB^M_l + \frac{r^+}{C}\left(\frac{m-e}{1+e}\right)e_{jkl}\frac{d\pi^I_k}{dt}B^M_l + \frac{r^e}{\beta C}e_{jkl}\frac{d\pi^e_k}{dt}B^M_l = r^e\frac{dv_j}{dt} + r^er^+\left(\frac{m-e}{1+e}\right)\frac{d^2\pi^I_j}{dt^2} + \frac{(r^e)^2}{\beta}\frac{d^2\pi^e_j}{dt^2}, \quad (6.33)$$

$$\mathfrak{D}_{ij,i} + \rho\mathcal{E}^I_j + \rho E^M_j + \frac{\rho}{C}e_{jkl}v_kB^M_l + \frac{r^+(m-1)}{C}\rho e_{jkl}\frac{d\pi^I_k}{dt}B^M_l + \frac{r^+}{C}\left(\frac{m-e}{1+e}\right)e_{jkl}\rho\frac{d\pi^e_k}{dt}B^M_l = m(r^+)^2\rho\frac{d^2\pi^I_j}{dt^2} + r^er^+\left(\frac{m-e}{1+e}\right)\rho\frac{d^2\pi^e_j}{dt^2}. \quad (6.34)$$

Substituting from (6.26)–(6.28) and (6.30) into (6.25) and employing the chain rule of differentiation, we obtain

$$\tau_{ij} = \rho y_{i,L}y_{j,M}\frac{\partial\psi}{\partial E_{LM}} - B^L_i\rho\mu'_j + \frac{\rho}{(\mu'_s)^2}y_{k,L}\frac{\partial\psi}{\partial N_L}\mu'_k\mu'_i\mu'_j - \mathcal{E}^I_i\rho\pi^I_j - \mathcal{E}^e_i\rho\pi^e_j + \mathfrak{D}_{ki}\pi^I_{j,k}. \quad (6.35)$$

Taking the axial vector of the antisymmetric part of τ_{ij} in (6.35), substituting from (6.33) and (6.34) and employing (3.23), (3.35), (3.36), (3.45), and (5.15), we obtain

$$e_{mij}\tau_{ij} = e_{mij}(mr^+\Delta_{ki}\pi^I_j)_{,k} - e_{mij}B^L_iM'_j - e_{mij}P_iE^M_j - e_{mij}P_ie_{jkl}\frac{v_k}{C}B^M_l - \frac{r^+(m-1)}{C}e_{mij}\rho\pi_ie_{jkl}\frac{d\pi^I_k}{dt}B^M_l - \frac{r^+(m-e)}{C(1+e)}\rho e_{mij}\left(\pi^e_ie_{jkl}\frac{d\pi^I_k}{dt}B^M_l + \pi^I_ie_{jkl}\frac{d\pi^e_k}{dt}B^M_l\right) - \frac{r^e}{C\beta}\rho e_{mij}\pi^e_ie_{jkl}\frac{d\pi^e_k}{dt}B^M_l + m(r^+)^2\rho e_{mij}\pi^I_i\frac{d^2\pi^I_j}{dt^2} + r^+r^e\left(\frac{m-e}{1+e}\right)\rho\left(e_{mij}\pi^I_i\frac{d^2\pi^e_j}{dt^2} + e_{mij}\pi^e_i\frac{d^2\pi^I_j}{dt^2}\right) + r^e\rho e_{mij}\pi^e_i\frac{dv_j}{dt} + \frac{(r^e)^2}{\beta}\rho e_{mij}\pi^e_i\frac{d^2\pi^e_j}{dt^2}, \quad (6.36)$$

which is identical with the component form of the vector form given in (3.48), which was obtained from the conservation of angular momentum for the combined continuum. Thus, even in this rather complex

situation, the antisymmetric portion of the mechanical stress tensor is derivable from a thermodynamic state function and has just the value required by the conservation of angular momentum.

This brings us to a consideration of the heat conducting constitutive equation, which proceeds from the entropy inequality (5.19), which indicates that we must have

$$q_i = q_i(\theta, _k), \tag{6.37}$$

since $\theta > 0$; but since the other constitutive equations (6.25)–(6.31) depend on the $y_{j,L}, \mu_i', \pi_i^I, \pi_i^e, \mu_{i,L}', \pi_{i,L}^I$, and θ, there is no logical reason to exclude them from this one.[53] Thus, because of the chain rule of differentiation, we may write

$$q_i = q_i(\theta, _M; y_{j,M}; \mu_j'; \pi_j^I; \pi_j^e; \mu_{j,M}'; \pi_{j,M}^I; \theta), \tag{6.38}$$

for the general functional dependence of the heat flux vector in this case of the interaction of the electromagnetic field with a heat conducting, polarizable and magnetizable deformable continuum. Now, q_i cannot be an arbitrary function of the variables shown in (6.38) because an arbitrary function of the 40 variables (12 vectors and four scalars) will not satisfy the principle of material objectivity,[26, 27] which requires the constitutive equations to be independent of the frame of reference of the observer. However, if q_i is expressed in the form

$$q_i = y_{i,K} L_K(\theta, _M; y_{j,M}; \mu_j'; \pi_j^I; \pi_j^e; \mu_{j,M}'; \pi_{j,M}^I; \theta), \tag{6.39}$$

then it can be shown readily, using established methods,[46,54,55] that the principle of material objectivity is satisfied if L_K is a vector invariant in a rigid motion. Then the previous application of Cauchy's theorem on invariant functions of vectors shows that the required invariance of L_K is assured if L_K is of the form[56]

$$L_K = L_K(\theta, _M; E_{L M}; N_L; W_L^I; W_L^e; G_{L M}; Q_{L M}; \theta), \tag{6.40}$$

where the functional dependence on the 34 variables shown in (6.40) may be arbitrary. However, L_K can have no term[46,55] independent of $\theta_{,M}$ and the dominant term must be odd in $\theta_{,M}$ because of (5.19). On the other hand on account of (6.1), χ cannot depend on $\theta_{,M}$. Thus, the constitutive equation for the heat flux vector in the general case is given by

$$q_i = y_{i,K} L_K, \tag{6.41}$$

with L_K as given in (6.40).

Equations (6.25)–(6.31) and (6.41) determine the constitutive equations for our continuum. Thus, all that remains in the determination of explicit constitutive equations is the selection of specific forms for χ and L_K. Once the constitutive equations have been determined, we have a determinate theory, which by appropriate substitution can readily be reduced to 18 equations in the 18 dependent variables $y_j, \theta, \pi_j^I, \pi_j^e, B_i^M, E_i^M$ and two of the three μ_i'. The 18 equations are the three each of (3.44), (6.33), (6.34), (4.1), and (4.2), two of the three of (3.25) and (5.18). Clearly, the system can be reduced further to 16 equations in 16 dependent variables with the aid of the electromagnetic potentials.[57] In order to have a complete field theory, the boundary (or jump) conditions at moving surfaces of discontinuity have to be adjoined to the aforementioned system of equations. This is done in the next section.

7. THE BOUNDARY CONDITIONS

In this section we determine the boundary conditions which must be adjoined to the system of differential equations, as noted at the end of Sec. 6, in order to formulate boundary-value problems. These boundary (or jump) conditions are determined by applying the integral forms of the pertinent field equations to appropriate limiting regions surrounding the moving (not necessarily material) surface of discontinuity[58] with normal velocity u_n, and assuming that certain variables remain bounded. The pertinent integral forms are (3.3), (3.20), (4.6)–(4.8), (4.18), the integral form of (5.19), which takes the form

$$\frac{d}{dt}\int_V \rho\eta dV + \int_S \frac{n_i q_i}{\theta} dS = \int_V \frac{-q_i\theta_{,i}}{\theta^2} dV \geq 0, \tag{7.1}$$

and an integral form associated with either (6.34) or (preferably) (3.46), neither of which can be used directly to find an integral form for the determination of jump conditions without making some sort of physical assumption about the manner in which $\mathbf{E}^L$ becomes unbounded in the vicinity of the surface of discontinuity. Different assumptions concerning this unboundedness of $\mathbf{E}^L$ result in different jump conditions on $n_i\mathfrak{D}_{ij}$. The most plausible form of these jump conditions can be determined from a consideration of the physical model in the limit required at a surface of discontinuity, while having recourse to the method of derivation of the equations from the model. Specifically, we first observe that the stress equations of motion (3.44), which ultimately resulted in the equivalent integral form (4.18), were determined by summing the force equations for the two ionic continua. We then note that the equation of motion of the ionic polarization was determined by taking the difference of the force equations for the two ionic continua according to the prescription $mr^+[(+) - (-)/m]$. We further observe that the electromagnetic force terms in (3.46), which become unbounded across surfaces of discontinuity, are of the same form, term by term, as those which become unbounded in the sum force equation (3.44), and resulted in the Maxwell stress tensor and the electromagnetic momentum terms in (4.18). In addition, we note that the expression for the Maxwell stress tensor for charge, current, and free-space regions, i.e., when no polarization or magnetization is present, is exactly the same[59] as in (4.17) in the absence of $\mathbf{P}$ and $\mathbf{M}'$, and the expression for the electromagnetic momentum is exactly the same as in (4.19). On account of the foregoing reasoning, we postulate that the integral form associated with the differential form (3.46), which is valid in the vicinity of a surface of discontinuity is

$$\begin{aligned} mr^+ &\int_S n_i\Delta_{ij} dS + \int_V \rho E_j^{SI} dV \\ &+ r^+ \int_S n_i[T_{ij}^{SI} + (m-1)v_i g_j] dS \\ &= r^+ \frac{d}{dt}\int \rho \left[\left(mr^+ \frac{d\pi_i^I}{dt} + r^e\left(\frac{m-e}{1+e}\right)\frac{d\pi_j^e}{dt}\right)\right. \\ &\left. + (m-1)g_j\right] dV, \end{aligned} \tag{7.2}$$

where

$$T_{ij}^{SI} =$$

$$\frac{1}{4\pi}\left[4\pi\mathcal{P}_i^I E_j^M + (m-1)(E_i^M E_j^M + B_i^M B_j^M) - 4\pi B_i^M \mathfrak{M}_j'^I - \frac{(m-1)}{2}(E_k^M E_k^M + B_k^M B_k^M - \frac{8\pi\mathfrak{M}_k'^I B_k^M}{(m-1)})\delta_{ij}\right] \quad (7.3)$$

and

$$\mathcal{P}_i^I = (m-1)P_i^I + [(m-e)/(1+e)]P_i^e, \quad (7.4)$$

$$\mathfrak{M}_i' = [(m-r^m)/(1+r^m)]M_i', \quad (7.5)$$

and $\mathbf{g}$ is given in (4.19), and $\mathbf{E}^{SI}$ is the portion of the volumetric interaction force density, exerted between the ionic continua, that remains bounded in the vicinity of the surface of discontinuity. Note that this procedure, which is based on the aforementioned physical argument, in essence makes a specific assumption about the manner in which $\mathbf{E}^L$ becomes unbounded in the vicinity of a surface of discontinuity. At this point it should be noted that similar considerations apply in the case of the integral form associated with (6.33) or (preferably) (3.42), but since there is no surface interaction between neighboring elements of the electronic charge continuum, the jump condition associated with this integral form is not needed in the formulation of boundary-value problems. However, appropriate consideration of this integral form indicates the existence of a surface force of interaction between the electronic charge continuum and the lattice continuum, which can be determined *a posteriori*. On the basis of our earlier physical argument, the relevant integral form is taken to be

$$\int_V \rho E_j^{Se} dV + r^e \int_S n_i \left(T_{ij}^{Se} + \frac{v_i g_j}{\beta}\right) dS = r^e \frac{d}{dt}\int_V \left\{\rho\left[v_j + r^+\left(\frac{m-e}{1+e}\right)\frac{d\pi_j^I}{dt} + \frac{r^e}{\beta}\frac{d\pi_j^e}{dt}\right] + \frac{g_j}{\beta}\right\} dV, \quad (7.6)$$

where

$$T_{ij}^{Se} = \frac{1}{4\pi}\left(4\pi\mathcal{P}_i^e E_j^M + \frac{1}{\beta}(E_i^M E_j^M + B_i^M B_j^M - 4\pi B_i^M M_j') - \frac{1}{2\beta}(E_k^M E_k^M + B_k^M B_k^M - 8\pi M_k' B_k^M)\delta_{ij}\right), \quad (7.7)$$

and

$$\mathcal{P}_i^e = (r^e/\beta)P_i^e + r^+[(m-e)/(1+e)]P_i^I, \quad (7.8)$$

and $\mathbf{E}^{Se}$ ($\neq \mathbf{E}^e$) becomes unbounded in the vicinity of the surface of discontinuity and $\mathbf{g}$ is given in (4.19).

For all integral forms considered, except (4.6) and (4.7), a volumetric region is taken in the usual way,[58] and it is assumed that all pertinent variables remain bounded. The jump conditions obtained from the respective integral forms consisting of (3.3), (3.20), (4.8), (4.18), (7.1), and (7.2) are

$$u_n[\rho] - n_i[v_i\rho] = 0, \quad (7.9)$$

$$n_i[A_{ik}e_{jkl}\rho\mu_l'] + u_n[\rho\mu_j'/\gamma] - n_i[v_i\rho\mu_j'/\gamma] = 0, \quad (7.10)$$

$$n_i[B_i^M] = 0, \quad n_i[D_i] = 0, \quad (7.11)$$

$$n_i[\tau_{ij} + T_{ij}^{EM}] + u_n[\rho(v_j + r^e\, d\pi_j^e/dt) + g_j] - n_i[v_i\rho(v_j + r^e\, d\pi_j^e/dt)] = 0, \quad (7.12)$$

$$n_i[q_i/\theta] - u_n[\rho\eta] + n_i[v_i\rho\eta] \geq 0, \quad (7.13)$$

$$n_i[\mathfrak{D}_{ij} + r^+T_{ij}^{SI}] + u_n\left[m(r^+)^2\frac{d\pi_j^I}{dt} + r^e\left(\frac{m-e}{1+e}\right)\frac{d\pi_j^e}{dt} + r^+(m-1)g_j\right] - n_i\left[v_i\left(m(r^+)^2\frac{d\pi_j^I}{dt} + r^e\left(\frac{m-e}{1+e}\right)\frac{d\pi_j^e}{dt}\right)\right] = 0, \quad (7.14)$$

where we have introduced the conventional notation $[C_i]$ for $C_i^+ - C_i^-$ and n_i denotes the components of the unit normal directed from the $-$ to the $+$ side of the surface of discontinuity. The jump conditions on $\mathbf{H}$ and $\mathbf{E}^M$, respectively, are determined from (4.6) and (4.7) by considering the circulation around a limiting open surface intersecting the moving surface of discontinuity in the usual way,[60] and are given by

$$n_j e_{ijk}[H_k] + (u_n/C)[D_i] = 0, \quad (7.15)$$

$$n_j e_{ijk}[E_k^M] - (u_n/C)[B_i^M] = 0. \quad (7.16)$$

If the surface of discontinuity is material

$$u_n = n_i v_i^+ = n_i v_i^-, \quad (7.17)$$

and (7.9) evaporates and (7.10) and (7.12)–(7.14), respectively, reduce to

$$n_i[A_{ik}e_{jkl}\rho\mu_l'] = 0, \quad (7.18)$$

$$n_i[\tau_{ij} + T_{ij}^{EM} + v_i g_j] = 0, \quad (7.19)$$

$$n_i[q_i/\theta] \geq 0, \quad (7.20)$$

$$n_i[\mathfrak{D}_{ij} + r^+T_{ij}^{SI} + r^+(m-1)v_i g_j] = 0, \quad (7.21)$$

and in (7.15) and (7.16), $u_n = n_i v_i$. Moreover, if θ is continuous, i.e.,

$$[\theta] = 0, \quad (7.22)$$

across the surface of discontinuity, Γ is bounded and, from (5.19), in place of (7.20), we have

$$n_i[q_i] = 0. \quad (7.23)$$

This latter situation, consisting of the jump conditions (7.11), (7.15), (7.16), (7.18), (7.19), and (7.21)–(7.23), is the most common, and if the body does not abut another solid body but abuts, say, air instead, the boundary conditions are fully defined by the noted equations. However, if a body does abut another solid body and the full field equations have to be satisfied in each region, additional conditions on any two of the $[\mu_i']$, $[y_i]$, and $[\pi_i^I]$ have to be satisfied at the surface of discontinuity. The condition on $[y_i]$ usually is

$$[y_i] = 0, \quad (7.24)$$

and the conditions on $[\mu_i']$ and $[\pi_i^I]$ we take as

$$\left[(\mu_s')^{-2}\frac{e_{ijk}\mu_j' d\mu_k'}{dt}\right] = 0, \quad [\pi_i^I] = 0. \quad (7.25)$$

Frequently, the thermal conditions are such that we may eliminate either (7.22) or (7.23). Clearly, all boundary expressions, which are not prescribed, may

be expressed in terms of the same 18 field variables as the 18 equations mentioned at the end of Sec. 6 by making the appropriate straightforward substitutions.

Thus, at this point we have obtained the nonlinear differential equations and boundary conditions describing the behavior of the interaction of the electromagnetic field with polarizable and magnetizable, heat conducting, deformable solid continua. The description consists of the aforementioned 18 equations and 24 boundary conditions, all expressed in terms of the 18 field variables y_i, two of the three μ_i', π_i^I, π_i^e, E_i^M, B_i^M, and θ in each region. All that remains in the determination of explicit equations is the selection of specific forms for ψ and L_K. A sensible polynomial approximation for ψ might be of the form

$$\psi = \frac{1}{2\rho_0} C_{KLMN}E_{KL}E_{MN} + \frac{\rho_0}{2}\chi^m_{KL}N_KN_L + \frac{\rho_0}{2}\chi^I_{KL}W^I_KW^I_L + \frac{\rho_0}{2}\chi^e_{KL}W^e_KW^e_L + \rho_0\zeta^I_{KL}N_KW^I_L + \rho_0\zeta^e_{KL}N_KW^e_L + \rho_0 h_{KL}W^I_KW^e_L + \frac{\rho_0}{2}\alpha_{KL}G_{KL} + \beta_{KLMN}Q_{KL}Q_{MN} + \frac{1}{2}C\theta^2 + \epsilon^m_{KLM}E_{LM}N_K + \epsilon^I_{KLM}E_{LM}W^I_K + \epsilon^e_{KLM}E_{LM}W^e_K + \eta_{KLMN}E_{KL}Q_{MN} + \frac{1}{\rho_0}\nu_{KL}E_{KL}\theta + \lambda^m_KN_K\theta + \lambda^I_KW^I_K\theta + \lambda^e_KW^e_K\theta + \rho_0 b^{mm}_{KLMN}E_{KL}N_MN_N + \rho_0 b^{II}_{KLMN}E_{KL}W^I_MW^I_N + \rho_0 b^{ee}_{KLMN}E_{KL}W^e_MW^e_N + \rho_0 b^{Ie}_{KLMN}E_{KL}W^I_MW^e_N + \rho_0 b^{mI}_{KLMN}E_{KL}N_MW^I_N + \rho_0 b^{me}_{KLMN}E_{KL}N_MW^e_N + \rho_0^2 f^m_{KLM}G_{LM}N_K + \rho_0^2 f^I_{KLM}G_{LM}W^I_K + \rho_0^2 f^e_{KLM}G_{LM}W^e_K + \rho_0\gamma_{KLMN}E_{KL}G_{MN} + \rho_0^2\psi_{KLMN}G_{KL}Q_{MN} + \text{higher order terms}, \quad (7.26)$$

where the material coefficients C_{KLMN}, χ^m_{KL}, χ^I_{KL}, χ^e_{KL}, α_{KL}, β_{KLMN}, C, ϵ^m_{KLM}, ϵ^I_{KLM}, ϵ^e_{KLM}, ν_{KL}, λ^m_K, λ^I_K, λ^e_K, b^{mm}_{KLMN}, b^{II}_{KLMN}, b^{ee}_{KLMN}, and γ_{KLMN} are called the elastic, magnetic anisotropy, reciprocal ionic and electronic susceptibilities, exchange, ionic polarization gradient, thermal, piezomagnetic, ionic and electronic piezoelectric, thermoelastic, pyromagnetic, ionic and electronic pyroelectric, magnetostrictive, ionic and electronic electrostrictive and exchangestrictive constants, respectively, and the remainder of the coupling coefficients we do not bother to name. Actually, all coefficients in (7.26) may be regarded as functions of the absolute temperature θ (and perhaps even the fields through N_L, W^I_K, W^e_K in some instances). For *linear* heat conduction the form for L_K would be

$$L_K = -\mathcal{K}_{KN}\theta_{,N}, \quad (7.27)$$

where the linear thermal conductivity tensor $\mathcal{K}$ may be a function of θ, E_{LM}, N_L, W^I_L, and W^e_L.

As noted earlier, we can determine an energetic jump condition from the integral form (5.24), which condition is not needed in the determination of the solution of a boundary-value problem, but can be useful for obtaining information when the entire solution is not available. This jump condition is obtained by applying (5.24) to the aforementioned volumetric region surrounding the (not necessarily material) surface of discontinuity and assuming that $\rho, v_j, \epsilon, \mu_k'$, π^I_k, π^e_k, B^M_k, and E^M_k remain bounded, with the result

$$n_j\Big[\tau_{jk}v_k + \mathfrak{D}_{jk}\frac{d\pi^I_k}{dt} - A_{jk}\rho\frac{d\mu_k'}{dt} - q_j - \Big(\frac{C}{4\pi}\Big)e_{jkl}E^M_kH_l + v_jP_kE^M_k + v_jU^F\Big] + u_n[T + \rho\epsilon + U^F] - n_j[v_j(T + \rho\epsilon + U^F)] = 0, \quad (7.28)$$

where T and U^F are given in (5.3) and (5.23), respectively, and ϵ may be found from (5.14), (6.2), and (6.23). If the surface is material, we have (7.17), and (72.8) reduces to

$$n_j[\tau_{jk}v_k + \mathfrak{D}_{jk}\,d\pi^I_k/dt - A_{jk}\rho d\mu_k'/dt - q_j - (C/4\pi)e_{jkl}E^M_kH_l + v_jP_kE^M_k + v_jU^F] = 0. \quad (7.29)$$

8. AN ALTERNATE DESCRIPTION FOR THE USUAL CASE

Frequently the material resonances—ionic polarization, electronic polarization and magnetic spin—may be left out of account, as may the distinction between ionic and electronic polarizations, the exchange interaction and polarization gradient effects. In such simplified circumstances, the model used in Sec. 2 is needlessly complex and a simpler, but similar, model consisting of a single electronic charge and spin continuum coupled to a single lattice continuum is perfectly adequate. When this more common situation is considered, the saturation condition (3.24) is abandoned, and in place of (3.25) we have

$$\mathbf{B}^L = -\mathbf{B}^M. \quad (8.1)$$

In addition, in place of the equations of motion of the ionic and electronic polarizations (3.42) and (3.46), we have the equation for the single electronic charge continuum

$$\sigma^e\mathbf{E}^e = -\sigma^e\mathbf{E}^M - \frac{\sigma^e}{C}\mathbf{v}\times\mathbf{B}^M - \frac{\rho}{C}\frac{d\pi}{dt}\times\mathbf{B}^M - \mathbf{P}\cdot\nabla\mathbf{E}^M - \frac{\mathbf{v}}{C}\times(\mathbf{P}\cdot\nabla\mathbf{B}^M) - \mathbf{M}'\cdot\mathbf{B}^M\nabla, \quad (8.2)$$

in which σ^e is the charge density of the electronic charge continuum, $\sigma^e\mathbf{E}^e$ is the force exerted by the lattice continuum on the electronic charge continuum at the position of the charge continuum and

$$\pi = \mathbf{P}/\rho. \quad (8.3)$$

Since the electronic inertia is negligible in the case treated here, the stress equations of motion take the form

$$\nabla\cdot\tau + \mathbf{P}\cdot\nabla\mathbf{E}^M + \frac{\mathbf{v}}{C}\times(\mathbf{P}\cdot\nabla\mathbf{B}^M) + \frac{\rho}{C}\frac{d\pi}{dt}\times\mathbf{B}^M + \mathbf{M}'\cdot\mathbf{B}^M\nabla = \rho\frac{d\mathbf{v}}{dt}, \quad (8.4)$$

in place of the form shown in (3.44), and the conservation of mass is still given by (3.3). At this point it should be noted that Eqs. (8.1) and (8.2), although satisfied, are not actually needed in this formulation. The electromagnetic equations [Eqs. (4.1)–(4.17) and (4.19)] remain valid. In place of the integral form in (4.18), we have

$$\int_S \mathbf{n}\cdot(\tau + \mathbf{T}^{EM} + \mathbf{vg})\,dS = \frac{d}{dt}\int_V(\rho\mathbf{v} + \mathbf{g})\,dV. \quad (8.5)$$

In place of (5.8) the equation of the conservation of energy takes the somewhat reduced form

$$\frac{d}{dt}\int_V \rho\,(\tfrac{1}{2}\,\mathbf{v}\cdot\mathbf{v}+\epsilon)dV=\int_S(\mathbf{t}\cdot\mathbf{v}-\mathbf{n}\cdot\mathbf{q})dS+\int_V \Sigma dV, \tag{8.6}$$

where Σ is still given by (5.7). In place of (5.16), the first law of thermodynamics takes the simpler form

$$\rho\frac{d\epsilon}{dt}=\tau_{ij}v_{j,i}-\rho\mu_i'\frac{dB_i^M}{dt}+\rho E_i'^M\frac{d\pi_i}{dt}-q_{i,i}. \tag{8.7}$$

The second law of thermodynamics for the present case takes the form

$$\rho\frac{d\epsilon}{dt}-\tau_{ij}v_{j,i}+\rho\mu_i'\frac{dB_i^M}{dt}-\rho E_i'^M\frac{d\pi_i}{dt}=\rho\theta\frac{d\eta}{dt}, \tag{8.8}$$

in place of (5.17). The dissipation equation and the entropy inequality are still given by (5.18) and (5.19), respectively. In place of (5.24) the integral form of the energy equation without source terms in the present case takes the form

$$\frac{d}{dt}\int_V[\rho(\tfrac{1}{2}v_kv_k+\epsilon)+U^F]dV=\int_S n_j[\tau_{jk}v_k-q_j-(C/4\pi)e_{jkl}E_k^MH_l+v_jP_kE_k^M+v_jU^F]dS, \tag{8.9}$$

where U^F is given in (5.23).

In this rather simplified case, the most convenient form of the constitutive equations is considerably different than the forms used earlier in the treatment of the more complicated situation, which appear in Sec. 6. In view of (5.19) and (4.2),[61] it turns out to be particularly convenient in the present situation to define the thermodynamic function F by the Legendre transformation

$$F=\epsilon-E_i'^M\pi_i-\eta\theta. \tag{8.10}$$

The substitution of the material time derivative of (8.10) into (8.8) yields

$$\rho\frac{dF}{dt}=\tau_{ij}x_{M,i}\frac{d}{dt}(y_{j,M})-\rho\mu_i'\frac{dB_i^M}{dt}-\rho\pi_i\frac{dE_i'^M}{dt}-\rho\eta\frac{d\theta}{dt}, \tag{8.11}$$

where we have substituted for $v_{j,i}$ as in the beginning of Sec. 6. Motivated by (8.11), we assume

$$F=F(y_{j,M};B_i^M;E_i'^M;\theta). \tag{8.12}$$

From (8.11) and (8.12), in the usual way, we obtain the relations

$$\tau_{ij}=\rho y_{i,M}\frac{\partial F}{\partial(y_{j,M})},\quad \mu_i'=-\frac{\partial F}{\partial B_i^M},$$
$$\pi_i=-\frac{\partial F}{\partial E_i'^M},\quad \eta=-\frac{\partial F}{\partial\theta}. \tag{8.13}$$

As with ψ in Sec. 6, F must be invariant in a rigid rotation and, from Cauchy's theorem, we find that F can be an arbitrary function of 15 scalar products and 10 determinants as well as θ, for a total of 26 quantities. However, the 26 quantities are not all functionally independent, and it can be shown that the 26 variables are expressible in terms of 13 arguments consisting of

$$Z_K=y_{i,K}B_i^M,\quad \Omega_K=y_{i,K}E_i'^M,\quad \theta, \tag{8.14}$$

and the E_{KL} defined in (6.24). Thus we find that F may be reduced to the form

$$F=F(E_{KL},Z_K,\Omega_K,\theta) \tag{8.15}$$

in place of the form shown in (8.12). From (8.13)–(8.15) and (6.24), we obtain

$$\tau_{ij}=\rho y_{i,L}y_{j,M}\frac{\partial F}{\partial E_{LM}}+\rho y_{i,L}\frac{\partial F}{\partial Z_L}B_j^M+\rho y_{i,L}\frac{\partial F}{\partial\Omega_L}E_j'^M, \tag{8.16}$$

$$\mu_i'=-y_{i,L}\frac{\partial F}{\partial Z_L},\quad \pi_i=-y_{i,L}\frac{\partial F}{\partial\Omega_L},\quad \eta=-\frac{\partial F}{\partial\theta}, \tag{8.17}$$

where we have introduced the usual conventions concerning F and E_{KL}. At this point it is easy to show, with the aid of (8.16) and (8.17), that the conservation of angular momentum is satisfied. The heat conduction constitutive equation is still given by (6.41), but now

$$L_K=L_K(\theta,_M;E_{LM};Z_M;\Omega_M;\theta), \tag{8.18}$$

in place of the form given in (6.40). Once specific forms for F and L_K are chosen, the constitutive equations may be determined from (8.16), (8.17), and (6.41). Thus we now have a determinate theory, which, by appropriate substitution, can readily be reduced to ten equations in the ten dependent variables y_j, B_i^M, E_i^M, and θ. The ten equations are the three each of (8.4), (4.1), and (4.2) and (5.18).

The jump conditions at moving nonmaterial surfaces of discontinuity required to complete this description are (7.9), (7.11),

$$n_i[\tau_{ij}+T_{ij}^{EM}]+u_n[\rho v_j+g_j]-n_i[v_i\rho v_j]=0, \tag{8.19}$$

in place of (7.12), (7.13), (7.15), and (7.16). If the surface of discontinuity is material, (7.17) holds and the boundary conditions are (7.11), (7.15), (7.16), (7.19) and, if θ is continuous also, (7.22) and (7.23). If in addition material surfaces are attached, (7.24) holds as well. Clearly, these boundary conditions can be expressed in terms of the same ten field variables as the differential equations.

It should be noted that since no material resonances or other complex material behavior are included in this description and, fundamentally, the model consists of charge and circulating current densities, it should not be difficult to find the equivalent Lorentz invariant form of this description. It should also be noted that other convenient descriptions incorporating any one material resonance, while ignoring the others, may readily be obtained by making the appropriate Legendre transformation in the state function equation.

ACKNOWLEDGMENTS

This work was supported in part by the Office of Naval Research under Contract No. N00014-67-A-0117-0007 and the National Science Foundation under Grant No. GK-11195.

APPENDIX

Equation (3.6) may be written in the component form

$$\oint_{C'}[\mathbf{r}\times(d\mathbf{s}\times\mathbf{B}^M]_j=\int_{C'}g_{mn}\,dZ_m, \tag{A1}$$

in which the magnitude of the steady circulating current density i'^+ has been omitted and where

$$g_{mj} = e_{jkl}e_{lmn}Z_k B_n^M. \tag{A2}$$

From Stokes theorem for tensor point functions,[62] we obtain

$$\oint_{C'} g_{mj}dZ_m = \int_{S'} n_i e_{irm} g_{mj,r}\,dS, \tag{A3}$$

where S' is the area enclosed by the closed curve C'. Since

$$Z_{k,r} = \delta_{kr}, \tag{A4}$$

$$e_{ikm}e_{jkl} = \delta_{ml}\delta_{ij} - \delta_{mj}\delta_{il}, \tag{A5}$$

in the limit $|\mathbf{r}| \to 0$, Eq. (A3) yields

$$\oint_{C'} [\mathbf{r} \times (d\mathbf{s} \times \mathbf{B}^M)] = \int_{S'_0} \mathbf{n} \times \mathbf{B}^M dS. \tag{A6}$$

In Eq. (3.18) we employ the vector integral identity[63]

$$\oint_{C'} d\mathbf{s} \times \mathbf{B}^M = \int_{S'} [\mathbf{n}\cdot(\mathbf{B}^M\nabla) - \mathbf{n}\nabla\cdot\mathbf{B}^M]\,dS, \tag{A7}$$

which with (4.5), yields

$$\oint_{C'} d\mathbf{s} \times \mathbf{B}^M = \int_S \mathbf{n}\cdot(\mathbf{B}^M\nabla)\,dS, \tag{A8}$$

and where we have again omitted the constant i'^+.

1 R. A. Toupin. J. Ratl. Mech. Anal. **5**, 849 (1956).
2 A. C. Eringen, Intern. J. Eng. Sci. **1**, 127 (1963) and many subsequent papers in the same journal.
3 W. F. Brown, Jr., J. Appl. Phys. **36**, 994 (1965).
4 H. F. Tiersten, J. Math. Phys. **5**, 1298 (1964).
5 H. F. Tiersten, J. Math. Phys. **6**, 779 (1965).
6 D. E. Eastman, Phys. Rev. **148**, 530 (1966).
7 C. Kittel, Phys. Rev. **110**, 836 (1958).
8 E. Schlömann, J. Appl. Phys. **31**, 1647 (1960).
9 A. I. Akhiezer, V. G. Bar'iakhtar, and S. V. Peletminskii, Zh. Eksp. Teoret. Fiz. **35**, 228 (1958). [Sov. Phys. JETP **8**, 157 (1959)].
10 R. Becker and W. Döring, *Ferromagnetismus* (Springer-Verlag, Berlin, 1939).
11 H. F. Tiersten, Intern. J. Eng. Sci. **9**, 587 (1971).
12 H. F. Tiersten, *Linear Piezoelectric Plate Vibrations* (Plenum, New York, 1969), Chap. 4, Sec. 4.
13 R. A. Toupin, Intern. J. Eng. Sci. **1**, 101 (1963).
14 R. C. Dixon and A. C. Eringen, Intern. J. Eng. Sci. **3**, 359 (1965).
15 P. Penfield, Jr. and H. A. Haus, *Electrodynamics of Moving Media*, (MIT, Cambridge, Mass., 1967).
16 Reference 15, Sec. 7. 2.
17 D. F. Nelson and M. Lax, Phys. Rev. B **3**, 2778 (1971).
18 By redefining stress tensors and the stored energy function, equivalence may possibly be shown.
19 R. A. Grot and A. C. Eringen, Intern. J. Eng. Sci. **4**, 611 (1966).
20 L. Bragg, J. Math. Phys. **11**, 318 (1970).
21 R. A. Grot, "Relativistic and Nonrelativistic Continuum Theories for the Interaction of Electromagnetic Fields with Deformable Bodies," submitted to J. Math. Phys.
22 H. A. Lorentz, *The Theory of Electrons* (Dover, New York, 1952), 2nd ed.
23 B. A. Boley and J. H. Weiner, *Theory of Thermal Stresses* (Wiley, New York, 1960).
24 S. R. DeGroot and P. Mazur, *Non-Equilibrium Thermodynamics* (North Holland, Amsterdam, 1962).
25 A. C. Eringen, *Nonlinear Theory of Continuous Media* (McGraw-Hill, New York, 1962), Secs. 110 and 111.
26 C. Truesdell and R. A. Toupin, "The Classical Field Theories," in *Encyclopedia of Physics*, edited by S. Flügge (Springer-Verlag, Berlin, 1960), Vol. III, Secs. 293 and 296; Ref. 25, Secs. 27 and 44.
27 Reference 1, Sec. 11.
28 R. D. Mindlin, Intern. J. Solids Structures **4**, 637 (1968).
29 A. Askar, P. C. Y. Lee, and A. S. Cakmak, Phys. Rev. B **1**, 3525 (1970).
30 G. H. Livens, *The Theory of Electricity* (Cambridge U.P., London, 1962), second ed., Secs. 237-39.
31 Reference 26, Secs. 13, 15 and 16.
32 Reference 25, Sec. 17.
33 C. Moller, *The Theory of Relativity* (Oxford U.P., London, 1952), Sec. 29.
34 See Appendix for missing details.
35 Reference 4, Eq. (4. 2).
36 Reference 4, Eq. (4. 15) and the associated wording.
37 R. Becker and F. Sauter, *Electromagnetic Fields and Interactions* (Blaisdell, New York, 1964), Sec. 53.
38 Reference 37, Sec. 87.
39 J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941), Sec. 1. 4.
40 Reference 30, Sec. 239.
41 Reference 26, Sec. 81.
42 Reference 4, Sec. 4, Eqs. (4. 18)–(4. 21).
43 Reference 23, Secs. 1. 8–1. 12.
44 Reference 24, Chap. III, Secs. 1 and 2.
45 Reference 25, Secs. 36–40.
46 B. D. Coleman and W. Noll, Arch. Ratl. Mech. Anal. **13**, 167 (1963).
47 B. D. Coleman, Arch. Ratl. Mech. Anal. **17**, 1 (1964).
48 Reference 39, Sec. 2. 19.
49 Conduction can readily be included simply by adding the term $E_i^M J_i$ to the right-hand side of (5. 22), where J_i is the current.
50 In the principle of material objectivity, as ordinarily used in continuum mechanics, the required invariance is under *time-dependent* rigid rotations. For dissipative systems, in which there are rate dependent constitutive equations, this time-dependence is essential, as it is for the generalization mentioned in Sec. 5. However, for stored energy systems, invariance under static rigid rotations yields the *same* results as invariance under *dynamic* rigid rotations. On account of this, present Lorentz invariant theories[19] require invariance under the proper Lorentz group, for which the *static* rigid rotations are a subgroup. However, since this treatment is not Lorentz invariant, the point of view adopted is that the required invariance is under *time-dependent* three-dimensional rigid rotations, and the specific electric and magnetic variables appearing as scalar products in the first law of thermodynamics [(5. 16) and (8. 7)] are to be regarded as three-dimensional vectorial quantities in considerations of material objectivity. It is implicitly presumed, of course, that the appropriate *time-dependent* relativistically invariant objective transformations will reduce to the assumptions of this treatment at low material velocities.
51 A. L. Cauchy, Mem. Acad. Sci. **22**, 615 (1850) [Oevres (1) **2**, 351].
52 J. H. Van Vleck, Rev. Mod. Phys. **17**, 27 (1945).
53 This procedure was given the name "Principle of Equipresence" by Truesdell. See Ref. 26, Sec. 293.
54 A. E. Green and J. E. Adkins, *Large Elastic Deformations and Nonlinear Continuum Mechanics* (Oxford U.P, London, 1960), Chap. 8, Sec. 5.
55 A. C. Pipkin and R. S. Rivlin, "The Formulation of Constitutive Equations in Continuum Physics," Technical Report DS 4531/4 to the Department of the Army, Ordnance Corps (1958).
56 Actually the required invariance of L_K will be assured even if K_{LM} is used instead of G_{LM}. However, since on account of (3. 26) G_{LM}, rather than K_{LM}, appears in all the other constitutive equations, G_{LM} is included in (6. 40) rather than K_{LM}. This is tantamount to requiring the exchange dependence of the heat conduction constitutive equation to be invariant in a rigid rotation of the entire spin continuum with respect to the lattice continuum.
57 Reference 39, Sec. 1. 9.
58 Reference 26, Secs. 192 and 193.
59 Reference 39, Sec. 2. 5.
60 H. H. Woodson and J. R. Melcher, *Electromechanical Dynamics* (Wiley, New York, 1968), Sec. 6. 21.
61 Since Faraday's law contains only the free space electromagnetic field variables $\mathbf{E}^M$ and $\mathbf{B}^M$, which can be expressed in terms of the electromagnetic potentials, it is particularly advantageous to use $\mathbf{E}^M$ and $\mathbf{B}^M$ as the basic electromagnetic constitutive variables when material resonances and related effects are suppressed as in this section.
62 L. Brand, *Vector and Tensor Analysis* (Wiley, New York, 1947), Sec. 100, Eq. (3).
63 Reference 62, Sec. 100, Eq. (4).

Quantum Theory of Heat Transport in an Isotopically Substituted, One-Dimensional, Harmonic Crystal

William L. Greer*† and Robert J. Rubin
Institute for Materials Research, National Bureau of Standards, Washington, D.C. 20234
(Received 1 September 1971)

We present a quantum mechanical treatment of thermal transport in a one-dimensional isotopically substituted harmonic lattice. This work is an extension of a classical mechanical treatment. We find that the difference between the quantum and classical expressions for the thermal conductivity of a random chain vanishes in the limit $N \to \infty$, where N is the number of isotopes. Thus, as in the classical treatment, the thermal conductivity diverges as $N^{1/2}$. For a periodic diatomic lattice, we derive explicit formulas for the heat current as a function of temperature. At very low temperatures, this quantum mechanical current exhibits Kapitsa behavior.

1. INTRODUCTION

In a recent paper,[1] referred to as RGI, we examined heat transport in a model crystal. Using classical statistical mechanics, we derived exact expressions for the steady-state heat current passing through a finite isotopically disordered harmonic lattice, whose end atoms were harmonically bound to semi-infinite-perfect harmonic chains at different temperatures. From the definition of the coefficient of thermal conductivity in terms of the heat current and temperature gradient, we found the conductivity to be a function of the length of the disordered segment. The numerical calculations for $25 < N < 600$, where N is the number of defects, revealed that the conductivity is proportional to $\sqrt{N}$ within statistical uncertainty.

In this paper we present a quantum mechanical extension of our earlier treatment of thermal transport. Much of the discussion closely parallels that given in RGI to which the reader is referred for details. As far as the anomalous behavior of the thermal conductivity is concerned, there is no substantial difference between the classical and quantum mechanical results for randomly disordered lattices. We present the quantum version for completeness and to demonstrate the connection between our nonequilibrium treatment and the method of Kubo[2] and Green[3] used by Allen and Ford.[4]

We emphasize that the thermal conductivity which we calculate is for a finite isotopically disordered chain of atoms with a particular coupling to the thermal bath on either side of it. The baths consist of semi-infinite perfect lattices and the connection is via a harmonic force between the extreme atoms of the system and bath. Other boundary conditions or different baths give different results for the thermal conductivity.[5] While real experimental systems possess intrinsic thermal conductivities whose values are independent of the nature of heat baths or surfaces used, these one-dimensional harmonic crystals possess transport coefficients which depend on such details.

2. REVIEW AND SUMMARY OF CLASSICAL TREATMENT

Consider a finite chain of atoms bound to nearest neighbors by harmonic forces. The particles are labeled by an index r, $-\mathfrak{N} \le r \le \mathfrak{N}$. Except for N isotopic defects of mass M, at lattice sites $r = A_j$, $j = 1, 2, \ldots, N$, all particles have mass m. The ordering of the indices is such that $0 = A_1 < A_2 < \cdots < A_j < \cdots < A_N$ and the length of the disordered region $L \equiv A_N - A_1$ is much smaller than the length of the entire chain $\mathfrak{L} = 2\mathfrak{N} + 1$. Ultimately we allow $\mathfrak{N} \to \infty$ in order to provide a thermal bath of infinite extent on either side of the finite disordered region. The nearest-neighbor force constant f is the same everywhere in the crystal.

The formal solution to the quantum mechanical equations of motion for the displacement and momentum operators is equivalent to the classical solution.[6]

$$\begin{aligned}\mathbf{X}(\tau) &= \mathsf{M}^{-1/2}\mathsf{W}^{-1/2}\sin(\mathsf{W}^{1/2}\tau)\mathsf{M}^{-1/2}\mathbf{P}(0) \\ &\quad + \mathsf{M}^{-1/2}\cos(\mathsf{W}^{1/2}\tau)\mathsf{M}^{1/2}\mathbf{X}(0), \\ \mathbf{P}(\tau) &= \mathsf{M}^{1/2}\cos(\mathsf{W}^{1/2}\tau)\mathsf{M}^{-1/2}\mathbf{P}(0) \\ &\quad - \mathbf{M}^{1/2}\mathsf{W}^{1/2}\sin(\mathsf{W}^{1/2}\tau)\mathsf{M}^{1/2}\mathbf{X}(0),\end{aligned} \tag{1}$$

where all matrices are $(2\mathfrak{N} + 1) \times (2\mathfrak{N} + 1)$. M is a diagonal matrix whose rrth element is the mass of the rth atom, and $\mathsf{W} = \mathsf{M}^{-1/2}\mathsf{V}\mathsf{M}^{-1/2}$.

The matrix V is the symmetrical potential energy matrix:

$$\mathsf{V}_{r,s} = 2f[\delta_{r,s} - \tfrac{1}{2}\delta_{r,s+1} - \tfrac{1}{2}\delta_{r,s-1}], \quad r, s \ne \mathfrak{N}. \tag{2}$$

The end conditions of the $2\mathfrak{N} + 1$ atom chain are that the atoms at $r = -\mathfrak{N}$ and $r = +\mathfrak{N}$ are harmonically bound to fixed positions (i.e., the hypothetical atom at $r = -\mathfrak{N} - 1$ or $+\mathfrak{N} + 1$ is fixed in its equilibrium position). Thus

$$\begin{aligned}\mathsf{V}_{r,-\mathfrak{N}} &= f[2\delta_{r,\mathfrak{N}} - \delta_{r-1,-}\ ], \\ \mathsf{V}_{r\,\mathfrak{N}} &= f[2\delta_{r,\mathfrak{N}} - \delta_{r+1\,\mathfrak{N}}].\end{aligned} \tag{3}$$

The time variable τ is dimensionless; with respect to the real time t, $\tau \equiv 2(f/m)^{1/2}t$, where f is the nearest neighbor force constant. In this model we considered an ensemble of initial conditions in which all atoms of index $-R < r \le +\mathfrak{N}$ are held fixed with zero velocity, and in which the initial velocities and displacements of the atoms of index $-\mathfrak{N} \le r \le -R$ are specified by a canonical distribution at temperature T.[1] These initial conditions are specified by the phase-space distribution function

$$\begin{aligned}\mathcal{W}\left(\begin{bmatrix}\dot{\mathbf{X}}_h(0)\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{X}_h(0)\\ \mathbf{0}\end{bmatrix}\right) &= Z\exp\left(-\frac{1}{2kT}\dot{\mathbf{X}}_h^T(0)\mathsf{M}_h\dot{\mathbf{X}}_h(0)\right. \\ &\quad \left. -\frac{1}{2kT}\mathbf{X}_h^T(0)\mathsf{V}_h\mathbf{X}_h(0)\right)\prod_{r=-R+1}^{\mathfrak{N}}\delta[\dot{x}(r,0)]\delta[x(r,0)].\end{aligned} \tag{4}$$

In this expression Z is the normalizing constant.

Here we have partitioned $\mathbf{X}(0)$ and $\dot{\mathbf{X}}(0)$ into

$$\mathbf{X}(0) = \begin{bmatrix}\mathbf{X}_h(0)\\ \mathbf{0}\end{bmatrix}, \quad \dot{\mathbf{X}}(0) = \begin{bmatrix}\dot{\mathbf{X}}_h(0)\\ \mathbf{0}\end{bmatrix}, \tag{5}$$

where the components of $\mathbf{X}_h(0)$ and $\dot{\mathbf{X}}_h(0)$ are the atomic displacements and velocities, respectively, of

particles of index $-\mathfrak{N} \le r \le -R$. The $(\mathfrak{N}+1-R) \times (\mathfrak{N}+1-R)$ matrix M_h is diagonal with elements equal to the mass m. The matrix V_h is defined by the following set of equations:

$$\begin{aligned}
[\mathsf{V}_h]_{r,s} &= 2f[\delta_{r,s} - \tfrac{1}{2}\delta_{r,s+1} - \tfrac{1}{2}\delta_{r,s-1}], \\
&\quad r,s \neq -R, -\mathfrak{N}, \\
[\mathsf{V}_h]_{r,-\mathfrak{N}} &= 2f[\delta_{r,-\mathfrak{N}} - \tfrac{1}{2}\delta_{r,-\mathfrak{N}+1}], \\
[\mathsf{V}_h]_{r,R} &= 2f[\delta_{r,-R} - \tfrac{1}{2}\delta_{r,-R-1}].
\end{aligned} \tag{6}$$

Similar definitions cover $\mathsf{V}_d, \mathsf{V}_c, \mathsf{M}_d, \mathsf{M}_c$.

This nonequilibrium temperature distribution gives rise to a net flow of heat from the "hot" region to the colder one. The average value of the heat current past atom r at any time $\tau > 0$ is

$$\langle J_N(r,\tau)\rangle = \tfrac{1}{4}m\langle \dot{x}(r,\tau)[x(r-1,\tau) - x(r,\tau)]\rangle, \tag{7}$$

where the classical phase space average implies, as usual,

$$\begin{aligned}
\langle A(\mathbf{X},\dot{\mathbf{X}};t)\rangle &\equiv \int dx_{-\mathfrak{N}}\, dx_{-\mathfrak{N}+1}\cdots \\
&\quad \times dx_{\mathfrak{N}}\, d\dot{x}_{-\mathfrak{N}}\, d\dot{x}_{-\mathfrak{N}+1}\cdots d\dot{x}_{\mathfrak{N}}\, \mathcal{W}[\dot{\mathbf{X}},\mathbf{X}]A(\mathbf{X},\dot{\mathbf{X}};t).
\end{aligned} \tag{8}$$

In RGI we showed that as $\tau \to \infty$, $\mathfrak{N} \to \infty$, and $\mathfrak{N}/\tau \to \infty$, the heat current $\langle J_N(r,\tau)\rangle$ becomes independent of r and τ:

$$\lim_{\tau,\mathfrak{N}\to\infty} \langle J_N(r,\tau)\rangle \equiv J_N = \frac{kT}{2\pi}\int_0^1 d\omega \mathcal{T}_N^2(\omega), \tag{9}$$

where $\mathcal{T}_N^2(\omega)$ is the square of the transmitted amplitude for a wave of frequency ω. The explicit expression for $\mathcal{T}_N^2(\omega)$ is given in RGI and in two earlier papers.[7,8]

In RGI we found that, because $\mathcal{T}_N^2(\omega)$ is negligible for frequencies much larger than $\omega \sim N^{-1/2}$, the heat current is

$$J_N \sim N^{-1/2}. \tag{10}$$

Hence, the thermal conductivity K_N, defined via

$$J_N = -K_N(\Delta T/L), \tag{11}$$

is

$$K_N \sim N^{1/2}. \tag{12}$$

In fact, an adequate approximation for K_N was developed in RGI by a Gaussian approximation for $\mathcal{T}_N^2(\omega)$. The result is

$$\tilde{K}_N = \frac{k}{4QC}\left(\frac{1+QC}{\pi(1-C)}\right)^{1/2} N^{1/2} \tag{13}$$

where $Q = (M-m)/m$ and C is the fractional concentration of defects in the segment of length $L = N/C$.

3. QUANTUM THEORY OF HEAT FLOW

In this quantum mechanical modification to the results of the previous section, we wish to retain the basic model: an initial thermal nonequilibrium condition which approaches a steady state as $\tau \to \infty$. This is accomplished simply by replacing $\mathcal{W}[\dot{\mathbf{X}}(0),\mathbf{X}(0)]$ of Eq. (4) by an initial density matrix operator $\rho(0)$ and by replacing the averaging prescription in Eq. (8) by a trace over a complete set of states.

Consider the three, uncoupled chains of atoms in Fig. 1. The left-most and right-most circular chains have periodic boundary conditions. The linear chain of length $2R-1 > A_N - A_1$ has fixed boundary conditions, although as we shall see, the exact nature of the boundaries for this chain are not important.

It is easy to see that these three chains may be assembled into the chain defined in the previous section by breaking and forming certain bonds.

The Hamiltonians for each chain are

$$\begin{aligned}
\mathcal{H}_h &= \sum_{n=-\mathfrak{N}}^{-R} \frac{-\hbar^2}{2m}\frac{\partial^2}{\partial x_n^2} + \tfrac{1}{2}f\sum_{n=-\mathfrak{N}}^{-R}(x_n - x_{n-1})^2, \\
\mathcal{H}_c &= \sum_{n=R}^{\mathfrak{N}} \frac{-\hbar^2}{2m}\frac{\partial^2}{\partial x_n^2} + \tfrac{1}{2}f\sum_{n=R}^{\mathfrak{N}}(x_n - x_{n+1})^2, \\
\mathcal{H}_d &= \sum_{n=-R+1}^{R-1} \frac{-\hbar^2}{2M_n}\frac{\partial^2}{\partial x_n^2} + \tfrac{1}{2}f\sum_{n=-R+2}^{R-1}(x_n - x_{n-1})^2 \\
&\quad + \tfrac{1}{2}fx_{R-1}^2 + \tfrac{1}{2}fx_{-R+1}^2.
\end{aligned} \tag{14}$$

Because of cyclic boundary conditions in the "h" and "c" chains, the atom of index $-\mathfrak{N}-1$ and $\mathfrak{N}+1$ are, respectively, $-R$ and R in the summations above. We use cyclic boundary conditions in order to facilitate the normal coordinate analysis. The chain whose thermal transport properties we are investigating has a Hamiltonian

$$\begin{aligned}
\mathcal{H} &= \sum_{n=-\mathfrak{N}}^{\mathfrak{N}} \frac{-\hbar^2}{2M_n}\frac{\partial^2}{\partial x_n^2} + \tfrac{1}{2}f\sum_{n=-\mathfrak{N}+1}^{\mathfrak{N}}(x_n - x_{n-1})^2 \\
&\quad + \tfrac{1}{2}fx_{\mathfrak{N}}^2 + \tfrac{1}{2}fx^2.
\end{aligned} \tag{15}$$

We now define the initial density matrix operator to be

$$\rho(0) \equiv \frac{\exp(-\beta_h\mathcal{H}_h)\exp(-\beta_d\mathcal{H}_d)\exp(-\beta_c\mathcal{H}_c)}{\mathrm{Tr}[\exp(-\beta_h\mathcal{H}_h)\exp(-\beta_d\mathcal{H}_d)\exp(-\beta_c\mathcal{H}_c)]}, \tag{16}$$

where $\beta_j \equiv (kT_j)^{-1}$ and T_j is the initial temperature of the jth chain for $j = h, d, c$.

Because the eigenstates of $\rho(0)$ are not the eigenstates of $\mathcal{H}$, the average value of any dynamic variable A changes with time:

$$\langle A(t)\rangle \equiv \mathrm{Tr}[\rho(0)e^{i\mathcal{H}t/\hbar}A(0)e^{-i\mathcal{H}t/\hbar}]. \tag{17}$$

The variable in which we are interested is the heat current operator

$$\begin{aligned}
J_N(r,\tau) &= \tfrac{1}{8}\{p(r,\tau)[x(r-1,\tau) - x(r,\tau)] \\
&\quad + [x(r-1,\tau) - x(r,\tau)]p(r,\tau)\}.
\end{aligned} \tag{18}$$

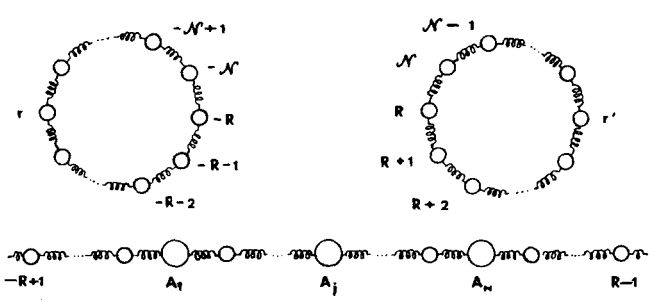


FIG. 1. Uncoupled harmonic chains used to define the initial conditions.

The trace is taken most conveniently in the basis set defined by the normal modes of the three independent chains h, d, and c.

In matrix notation,

$$\begin{aligned}\mathcal{H}_h &= \tfrac{1}{2}\mathbf{P}_h^T\mathsf{M}_h\mathbf{P}_h + \tfrac{1}{2}\mathbf{X}_h^T\mathsf{V}_h\mathbf{X}_h,\\ \mathcal{H}_c &= \tfrac{1}{2}\mathbf{P}_c^T\mathsf{M}_c\mathbf{P}_c + \tfrac{1}{2}\mathbf{X}_c^T\mathsf{V}_c\mathbf{X}_c,\\ \mathcal{H}_d &= \tfrac{1}{2}\mathbf{P}_d^T\mathsf{M}_d\mathbf{P}_d + \tfrac{1}{2}\mathbf{X}_d^T\mathsf{V}_d\mathbf{X}_d,\end{aligned} \tag{19}$$

where the elements of $\mathbf{X}_c$ are the displacements of all atoms of index $+R \le r \le \mathfrak{N}$ and $\mathbf{P}_c$ contains the conjugate momentum operators to those displacements. Similar definitions obtain for $\mathbf{X}_d$ and $\mathbf{P}_d$ and $\mathbf{X}_h$ and $\mathbf{P}_h$. Obviously,

$$\mathbf{X}(0) = \begin{bmatrix}\mathbf{X}_h(0)\\ \mathbf{X}_d(0)\\ \mathbf{X}_c(0)\end{bmatrix}, \qquad \mathbf{P}(0) = \begin{bmatrix}\mathbf{P}_h(0)\\ \mathbf{P}_d(0)\\ \mathbf{P}_c(0)\end{bmatrix}. \tag{20}$$

Normal coordinates for each of the three initially unconnected chains are

$$\begin{aligned}\mathcal{Q}_j &= \mathsf{S}_j^T\mathsf{M}_j^{1/2}\mathbf{X}_j,\\ \mathcal{P}_j &= \mathsf{S}_j^T\mathsf{M}_j^{-1/2}\mathbf{P}_j, \qquad j = h, c, \text{or } d,\end{aligned} \tag{21}$$

whence

$$\mathcal{H}_j = \tfrac{1}{2}\mathcal{P}_j^T\mathcal{P}_j + \tfrac{1}{2}\mathcal{Q}_j^T\Omega_j^2\mathcal{Q}_j. \tag{22}$$

In Eqs. (21) and (22),

$$\Omega_j^2 \equiv \mathsf{S}_j^T\mathsf{M}_j^{-1/2}\mathsf{V}_j\mathsf{M}_j^{-1/2}\mathsf{S}_j \equiv \mathsf{S}_j^T\mathsf{W}_j\mathsf{S}_j, \qquad j = h, c, \text{or } d. \tag{23}$$

The S_j matrix is found from the requirement that the Ω_j^2 matrices be diagonal with elements ω_k^2 which are solutions to the secular equation

$$\det[\mathsf{M}_j^{-1/2}\mathsf{V}_j\mathsf{M}_j^{-1/2} - \omega^2\,{}_j] = 0, \qquad j = h, c, \text{or } d.$$

Quantum mechanical averages now are[6]

$$\begin{aligned}\langle\mathcal{P}_j\mathcal{P}_{j'}^T\rangle &= \delta_{j,j'}[\tfrac{1}{2}\hbar\Omega_j\coth(\tfrac{1}{2}\hbar\beta_j\Omega_j)],\\ \langle\mathcal{Q}_j\mathcal{Q}_{j'}^T\rangle &= \delta_{j,j'}[\tfrac{1}{2}\hbar\Omega_j^{-1}\coth(\tfrac{1}{2}\hbar\beta_j\Omega_j)],\\ \langle\mathcal{Q}_j\mathcal{P}_{j'}^T\rangle &= \delta_{j,j'}[\tfrac{1}{2}i\hbar\,\mathbb{1}_j],\\ \langle\mathcal{P}_j\mathcal{Q}_{j'}^T\rangle &= \delta_{j,j'}[-\tfrac{1}{2}i\hbar\,\mathbb{1}_j], \qquad j = h, c, \text{or } d.\end{aligned} \tag{24}$$

The calculation of the average of $J_N(\tau, \tau)$ is outlined in the Appendix. The final $\tau \to \infty$, $\mathfrak{N} \to \infty$ steady state heat current is given by Eq. (A32) of the Appendix:

$$J_N = \frac{1}{2\pi}\int_0^1 d\omega\mathcal{T}_N^2(\omega)\left(\frac{\hbar\omega}{(e^{\hbar\omega/kT_h}-1)} - \frac{\hbar\omega}{(e^{\hbar\omega/kT_c}-1)}\right). \tag{25}$$

The result is independent of T_d, the initial temperature of the finite defect chain.

We also see that, in general, the heat flux is not proportional to the temperature difference $T_h - T_c = \delta T$. However, if δT is small, we may expand Eq. (25) in a power series in δT:

$$J_N = \frac{\delta T}{2\pi kT^2}\int_0^1 d\omega\frac{\hbar^2\omega^2 e^{\hbar\omega/kT}}{(e^{\hbar\omega/kT}-1)^2}\mathcal{T}_N^2(\omega) + \mathcal{O}[(\delta T)^2] + \cdots, \tag{26}$$

where $T = T_h$.

Thus, by Eq. (11), we identify the thermal conductivity from the linear term

$$K_N = \frac{L}{2\pi kT^2}\int_0^1 d\omega\hbar^2\omega^2\frac{e^{\hbar\omega/kT}}{(e^{\hbar\omega/kT}-1)^2}\mathcal{T}_N^2(\omega). \tag{27}$$

This is in exact agreement with the results of Allen and Ford,[4] who use the Kubo linear response formalism. As $\hbar \to 0$, we retrieve the classical result obtained earlier in RGI:

$$\lim_{\hbar\to 0} K_N = \frac{Lk}{2\pi}\int_0^1 d\omega\mathcal{T}_N^2(\omega). \tag{28}$$

It is noteworthy that our model, which places the entire chain in an extreme nonequilibrium situation initially and which follows the average heat flow as a function of time until a steady state obtains, gives exactly the same result as the Kubo formalism in which an infinitesimal temperature gradient is imposed initially.

Is there any quantum effect on K_N for an isotopically disordered lattice of N defects at random atomic sites within a length L? Unfortunately, no. In RGI we demonstrated that $\mathcal{T}_N^2(\omega)$ is essentially zero for $\omega > N^{-1/2}$. Hence quantum effects which are distinct from classical ones will be seen only if $\hbar\omega/kT \gg 1$, for these values of ω which contribute to the integral, or, $kT \ll \hbar\omega < \hbar N^{-1/2}$. Thus the temperature must be small indeed for a quantum contribution to K_N to be sizeable if N is macroscopically large.

4. PERIODIC ISOTOPIC SUBSTITUTION

In the cases when $\mathcal{T}_N^2(\omega)$ is nonvanishing over an appreciable frequency range the fore-going conclusions are not applicable. One particular situation which is easy to treat is the case in which the spacing between successive defects is a constant, i.e., a periodically substituted lattice. If we denote the inter-defect separation by a, the transmission coefficient is given by Rubin,[7] Appendix B;

$$\mathcal{T}_N^2(\omega) = |D_N|^{-2}, \tag{29}$$

where

$$D_N = [U_N(\chi) - (1 - i\Delta)\exp(-i\kappa a)U_{N-1}(\chi)]\exp(-iN\kappa a), \tag{30}$$

and where $U_N(\chi) = \sin(N+1)\eta/\sin\eta$ is a Tchebycheff polynomial, $\eta = \cos^{-1}\chi$ and

$$\begin{aligned}\Delta &= Q\omega(1-\omega^2)^{-1/2},\\ \chi &= \tfrac{1}{2}(1 + i\Delta)\exp(i\kappa a) + \tfrac{1}{2}(1 - i\Delta)\exp(-i\kappa a).\end{aligned} \tag{31}$$

One finds that

$$\mathcal{T}_N^2(\omega) = \left(1 + \frac{\Delta^2\sin^2(N\eta)}{\sin^2\eta}\right)^{-1}. \tag{32}$$

As long as $|\chi| < 1$, η is real and $\mathcal{T}_N^2(\omega)$ is an oscillating function of ω. When $|\chi| > 1$, η becomes imaginary, and $\mathcal{T}_N^2(\omega)$ is an exponentially decreasing function of N.

A rearrangement of Eq. (31) yields

$$\sin^2\eta = 1 - (1 + \Delta^2)\cos^2(\kappa a + \phi), \tag{33}$$

where $\phi = \tan^{-1}\Delta$. The transmission cut-off frequencies $\{\omega_j\}$ at which $|\chi| = 1$ are determined from the condition

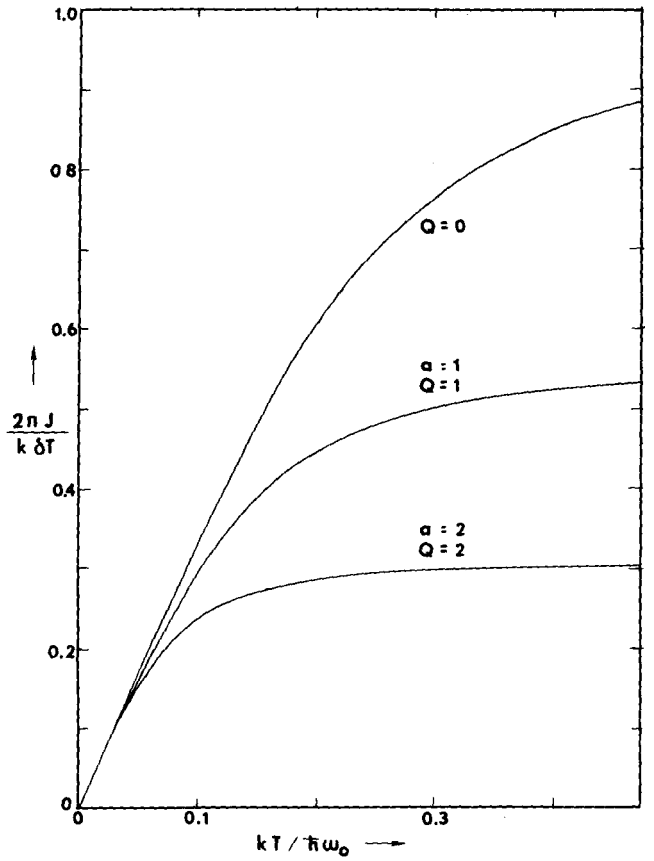


FIG. 2. Heat current J (multiplied by $2\pi/k\delta T$) as a function of temperature. Here $\omega_0 = 2(f/m)^{1/2}$ is unity in the units employed in this paper.

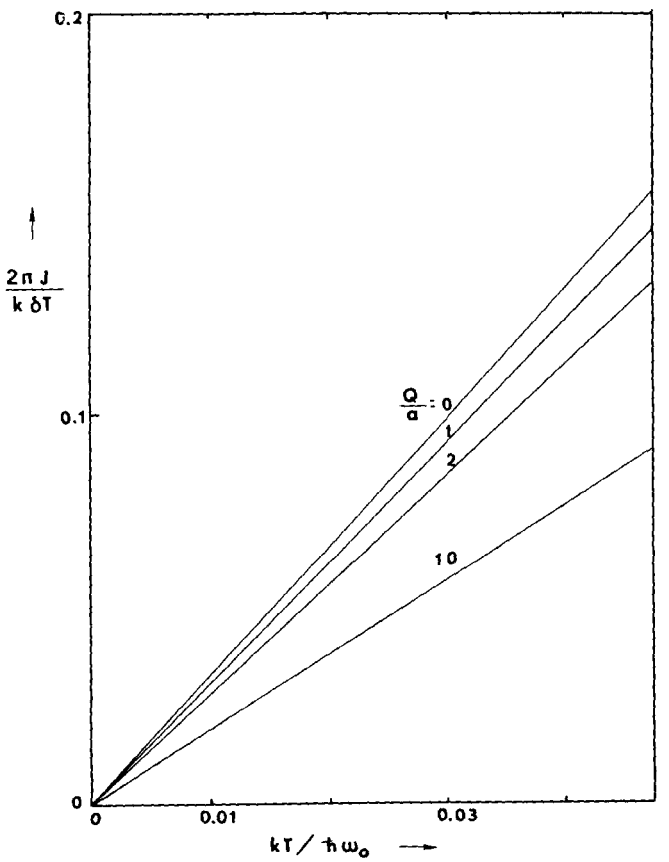


FIG. 3. Low temperature behavior of $2\pi J/k\delta T$ for choices of Q/a.

$$\sin^2\eta = 0,$$

or

$$(1 + \Delta^2)\cos^2(\kappa a + \phi) = 1,$$

or

$$\cos(\kappa a) - \Delta \sin(\kappa a) = \pm 1. \tag{34}$$

Equation (34) provides the frequencies at which transmission regions are separated by stopping bands, for N sufficiently large.

In the limit $N \to \infty$, we may replace $\mathcal{T}_N^2(\omega)$ by a smoothed function $\mathcal{T}_\infty^2(\omega)$:

$$\mathcal{T}_\infty^2(\omega) \equiv \begin{cases} 0 & , \quad \eta \text{ imaginary} \\ (1 + \tfrac{1}{2}\Delta^2/\sin^2\eta)^{-1}, & \eta \text{ real}, \end{cases} \tag{35}$$

for all frequencies except $\omega = 0$ at which $\mathcal{T}_\infty^2(\omega) = 1$.

This simplification to Eq. (32) occurs because, for real η, $\sin^2(N\eta)$ is rapidly oscillatory about $\frac{1}{2}$ over any small frequency range $[\omega, \omega + \delta\omega]$. As $N \to \infty$, the interval $\delta\omega$ becomes infinitesimal. The smoothed transmission coefficient of Eq. (35) is useful in numerical integrations.

Because the smoothed coefficient for a periodically substituted lattice becomes independent of N as $N \to \infty$, the thermal conductivity defined by Eq. (11) diverges linearly with N. This feature has been noted for perfect harmonic lattices also.[1,9]

The heat current is

$$J = \frac{\delta T}{2\pi k T^2} \int_0^1 d\omega \mathcal{T}_\infty^2(\omega) \hbar^2\omega^2 \frac{e^{\hbar\omega/kT}}{(e^{\hbar\omega/kT} - 1)^2}. \tag{36}$$

In the high temperature limit for which $kT \gg \hbar\omega$, the heat current is proportional to δT and independent of T itself:

$$J = \frac{k\delta T}{2\pi} \int_0^1 d\omega \mathcal{T}_\infty^2(\omega). \tag{37}$$

However, in the low temperature region for which $kT \lesssim \hbar\omega$, a distinct quantum effect emerges, and J is a function of T and of δT. At very low temperatures, the temperature dependent terms in the integrand of Eq. (36) decay rapidly as a function of ω. As a consequence, the transmission coefficient is essentially constant over that frequency range for which the integrand contributes to the integral.

Hence, at sufficiently small T,

$$J = \frac{\delta T}{2\pi k T^2} \mathcal{T}^2 \int_0^\infty d\omega \frac{\hbar^2\omega^2 e^{\hbar\omega/kT}}{(e^{\hbar\omega/kT} - 1)^2}, \tag{38}$$

where

$$\mathcal{T}^2 = \lim_{\omega \to 0^+} \mathcal{T}_\infty^2(\omega) = \left(1 + \frac{(Q/a)^2}{8(1 + Q/a)}\right)^{-1}. \tag{39}$$

We note that $\mathcal{T}^2$ is a function of Q/a which implies that the low temperature limiting behavior of the heat current for periodic crystals of the same ratio Q/a is identical.

For sufficiently low temperatures ($kT \ll \hbar$), the integral in Eq. (38) differs negligibly from the heat capacity per atom of the heat reservior harmonic chains:

$$C(T) \equiv \frac{2}{\pi} \int_0^1 d\omega \frac{\hbar^2\omega^2 e^{\hbar\omega/kT}}{(e^{\hbar\omega/kT} - 1)^2} (1 - \omega^2)^{-1/2}$$

or

$$C(T) \simeq \frac{2}{\pi} \int_0^1 d\omega \frac{\hbar^2\omega^2 e^{\hbar\omega/kT}}{(e^{\hbar\omega/kT} - 1)^2}, \quad \text{if } kT \ll \hbar\omega. \tag{40}$$

Thus we write

$$J = \frac{\delta T}{4kT^2}\mathcal{T}^2 C(T). \tag{41}$$

The exact T-dependence of the current as $T \to 0$ is

$$J = \left[\frac{\mathcal{T}^2\zeta(2)}{\pi}k^2\right] T\delta T, \tag{42}$$

where $\zeta(n)$ is the Riemann Zeta Function of argument n. This proportionality of J to $T\delta T$ is typical of Kapitsa resistance in one dimension. In three dimensions, one observes a $T^3\delta T^{10}$ behavior which is related to the difference in the density of states near $\omega = 0$ in one and three dimensions.

In Figs. (2) and (3), the heat current vs. temperature for choices of Q and a is displayed. Both the classical $kT \gg \hbar\omega$ and Kapitsa $kT \ll \hbar\omega$ are clearly seen.

5. DISCUSSION

We have extended our earlier[1] classical treatment of heat transport across an isotopically substituted linear crystal to the quantum domain. In the limit of an infinitesimal temperature gradient δT, our result reduces to the Kubo–Green result worked out by Allen and Ford.[4] From a knowledge of the transmission coefficient $\mathcal{T}_N^2(\omega)$ as a function of frequency, we are able to deduce the following results.

(a) In an isotopically *disordered* one-dimensional lattice connected to harmonic heat baths, the thermal conductivity is divergent as $N^{1/2}$, where N is the number of isotopic impurities. This general result is true in the quantum and classical domains.

(b) In a *periodic* isotopically substituted lattice, the low temperature heat current is given by Eqs. (38) and (39). All crystals with the same ratio of Q/a have exactly the same low temperature current. For large values of N, the thermal conductivity is proportional to N.

APPENDIX

According to Eq. (18), the average heat current past atom r at time τ is

$$\langle J_N(r,\tau)\rangle = \tfrac{1}{8}\langle p(r,\tau)[x(r-1,\tau) - x(r,\tau)] + [x(r-1,\tau) - x(r,\tau)]p(r,\tau)\rangle. \tag{A1}$$

For the finite $2\mathfrak{N}+1$ particle chain, we use the matrix notation introduced previously;

$$\begin{aligned} X(r,\tau) &= \Delta_r^T[\mathsf{M}^{-1/2}\mathsf{W}^{-1/2}\sin(\mathsf{W}^{1/2}\tau)\mathsf{M}^{-1/2}\mathbf{P}(0) \\ &\quad + \mathsf{M}^{-1/2}\cos(\mathsf{W}^{1/2}\tau)\mathsf{M}^{1/2}\mathbf{X}(0)], \\ P(r,\tau) &= \Delta_r^T[\mathsf{M}^{1/2}\cos(\mathsf{W}^{1/2}\tau)\mathsf{M}^{-1/2}\mathbf{P}(0) \\ &\quad - \mathsf{M}^{1/2}\mathsf{W}^{1/2}\sin(\mathsf{W}^{1/2}\tau)\mathsf{M}^{1/2}\mathbf{X}(0)]. \end{aligned} \tag{A2}$$

The $2\mathfrak{N}+1$ component column vector Δ_r has only one nonzero element, the rth one which is unity. Using the normal coordinates of the initially uncoupled lattices, Eqs. (20) and (21), we define

$$\begin{aligned} \mathbf{X}(0) &= \mathsf{M}^{-1/2}\mathsf{S}^{(0)}\mathcal{Q}^{(0)}, \\ \mathbf{P}(0) &= \mathsf{M}^{1/2}\mathsf{S}^{(0)}\mathcal{P}^{(0)}, \end{aligned} \tag{A3}$$

where

$$\mathcal{Q}^{(0)} \equiv \begin{bmatrix} \mathcal{Q}_h \\ \mathcal{Q}_d \\ \mathcal{Q}_c \end{bmatrix}, \quad \mathcal{P}^{(0)} \equiv \begin{bmatrix} \mathcal{P}_h \\ \mathcal{P}_d \\ \mathcal{P}_c \end{bmatrix}, \quad \mathsf{S}^{(0)} \equiv \begin{bmatrix} \mathsf{S}_h & 0 & 0 \\ 0 & \mathsf{S}_d & 0 \\ 0 & 0 & \mathsf{S}_c \end{bmatrix}. \tag{A4}$$

With these definitions, the general average value of interest in the heat current expressions, Eq. (A1), may be written as

$$\begin{aligned} \tfrac{1}{8}\langle P(r,\tau)x(r',\tau) + x(r',\tau)P(r,\tau)\rangle &\equiv \mathcal{J}_{rr'}(\tau) \\ &= \tfrac{1}{4}\Delta_r^T[\mathsf{M}^{1/2}\cos(\mathsf{W}^{1/2}\tau)\mathsf{S}^{(0)}\langle\mathcal{P}^{(0)}\mathcal{P}^{(0)T}\rangle\mathsf{S}^{(0)T} \\ &\quad \times \sin(\mathsf{W}^{1/2}\tau)\mathsf{W}^{-1/2}\mathsf{M}^{-1/2} - \mathsf{M}^{1/2}\mathsf{W}^{1/2}\sin(\mathsf{W}^{1/2}\tau) \\ &\quad \times \mathsf{S}^{(0)}\langle\mathcal{Q}^{(0)}\mathcal{Q}^{(0)T}\rangle\mathsf{S}^{(0)T}\cos(\mathsf{W}^{1/2}\tau)\mathsf{M}^{-1/2}]\Delta_{r'}. \end{aligned} \tag{A5}$$

In Eq. (A5) the quantum mechanical averages of products of normal coordinates and conjugate momenta are obtained by using Eqs. (A4) and (24). The cross terms containing $\langle\mathcal{Q}^{(0)}\mathcal{P}^{(0)T}\rangle$ and $\langle\mathcal{P}^{(0)}\mathcal{Q}^{(0)T}\rangle$ cancel exactly.

The matrices, M, W, and $\mathsf{S}^{(0)}$ are the same for both classical and quantum mechanics. Only the averages $\langle\mathcal{Q}^{(0)}\mathcal{Q}^{(0)T}\rangle$ and $\langle\mathcal{P}^{(0)}\mathcal{P}^{(0)T}\rangle$ appearing in Eq. (A5) are different in quantum and classical calculations.

We write out explicitly the result of matrix multiplication,

$$\mathcal{J}_{rr'}(\tau) = \tfrac{1}{4}\sum_q [P_{rq}(\tau)\langle\mathcal{P}^{(0)}\mathcal{P}^{(0)T}\rangle_{qq}X_{r'q}(\tau) + \dot{P}_{rq}(\tau)\langle\mathcal{Q}^{(0)}\mathcal{Q}^{(0)T}\rangle_{qq}\dot{X}_{r'q}(\tau)], \tag{A6}$$

where the index q denotes normal modes of the initially uncoupled lattices and

$$X_{r'q}(\tau) \equiv [\mathsf{M}^{-1/2}\mathsf{W}^{-1/2}\sin(\mathsf{W}^{1/2}\tau)\mathsf{S}^{(0)}]_{r'q}, \tag{A7a}$$

$$P_{rq}(\tau) = M_r\dot{X}_{rq}(\tau) = [\mathsf{M}^{1/2}\cos(\mathsf{W}^{1/2}\tau)\mathsf{S}^{(0)}]_{rq}, \tag{A7b}$$

and the dot over $X_{rq}(\tau)$ or $P_{rq}(\tau)$ denotes a time derivative. From Eqs. (A2), we observe that $X_{r'q}(\tau)$ as defined in Eq. (A7a) is the displacement of the r'th atom at time τ when the initial lattice conditions are defined by

$$\mathcal{Q}^{(0)} = 0, \qquad \mathcal{P}^{(0)} = \Delta_q. \tag{A8}$$

Thus $X_{r'q}(\tau)$ is the *classical* displacement of the r'th atom, while $P_{rq}(\tau)$ is the *classical* momentum of the rth atom at time τ when the initial lattice conditions are given in Eq. (A8). The identification of these functions at classical displacements and momenta allows contact with our earlier[1] classical treatment of the disordered harmonic chain.

The quantum mechanics enters Eq. (A5) from

$$\begin{aligned} \langle\mathcal{P}^{(0)}\mathcal{P}^{(0)T}\rangle_{qq'} &= \tfrac{1}{2}\hbar\omega_q\coth(\tfrac{1}{2}\hbar\omega_q/kT_j)\delta_{q,q'}, \\ \langle\mathcal{Q}^{(0)}\mathcal{Q}^{(0)T}\rangle_{qq'} &= \tfrac{1}{2}\hbar\omega_q^{-1}\coth(\tfrac{1}{2}\hbar\omega_q/kT_j)\delta_{q,q'}, \end{aligned} \tag{A9}$$

where the normal mode frequency of index q is ω_q, k is Boltzmann's constant, and the temperature T_j is T_h, T_d, or T_c, depending on the chain of Fig. 1 for which q is a normal mode index.

We are interested in the long time behavior of the heat flow when transients introduced by the initial nonequilibrium temperature profile have decayed. Because the defect bearing region of length L if finite in extent, for τ sufficiently large that

$$L \ll \tau \ll \mathfrak{N}, \tag{A10}$$

the contributions to the heat current J_N arising from the initial temperature T_d of the defect region are much smaller than the contributions to J_N from the two large heat baths at temperatures T_h and T_c. In the limit as $\mathfrak{N} \to \infty, \tau \to \infty$ and $\mathfrak{N}/\tau \to \infty$, the contributions from the defect region's initial temperature are vanishingly small. If we anticipate the limits $\mathfrak{N} \to \infty$ and $\tau \to \infty$ and neglect contributions to J_N or $\mathcal{J}_{r,r'}(\tau)$ from the defect modes, we find $\mathcal{J}_{r,r'}(\tau)$ to be a sum of two contributions, one from the lattice section initially at temperature T_h, the other from the lattice section initially at T_c:

$$\begin{aligned}
&\mathcal{J}_{r,r'}(\tau) \\
&= \sum_{q\in h} [P_{rq}(\tau)X_{r'q}(\tau)\langle \mathcal{P}_h \mathcal{P}_h^T\rangle_{qq} + \dot{P}_{rq}(\tau)\dot{X}_{r'q}(\tau)\langle \mathcal{Q}_h \mathcal{Q}_h^T\rangle_{qq}] \\
&\quad + \sum_{q\in c} [P_{rq}(\tau)X_{r'q}(\tau)\langle \mathcal{P}_c \mathcal{P}_c^T\rangle_{qq} + \dot{P}_{rq}(\tau)\dot{X}_{r'q}(\tau)\langle \mathcal{Q}_c \mathcal{Q}_c^T\rangle_{qq}] \\
&\equiv \mathcal{J}^{(h)}_{r,r'}(\tau) + \mathcal{J}^{(c)}_{r,r'}(\tau).
\end{aligned} \tag{A11}$$

Next we evaluate $\mathcal{J}^{(h)}_{r,r'}(\tau)$ explicitly and appeal to simple physical symmetry to deduce $\mathcal{J}^{(c)}_{r,r'}(\tau)$. In this calculation we require an explicit form for S_h. For mathematical convenience, we have chosen the left-most chain to be initially characterized by cyclic boundary conditions. In general we are free to specify any arbitrary boundary conditions, but we know that as $\mathfrak{N} \to \infty$ and as $\tau \to \infty$, the contributions of the boundaries to $\mathcal{J}_{r,r'}(\tau)$ vanish, provided r and r' are finite.

For the cyclic boundary conditions cited previously and depicted in Fig. 1, the initial velocity (classical) of the nth particle, $-\mathfrak{N} \le n \le -R$, is given in terms of the classical momenta $[\mathcal{P}_h]_q$ by[11]

$$\begin{aligned}
\dot{x}(n,0) = \left(\frac{2}{m\mathfrak{N}'}\right)^{1/2} \sum_{s=1}^{\mathfrak{N}'/2} &\left[\mathcal{P}_{s,1} \cos\left(\frac{2\pi sn}{\mathfrak{N}'}\right)\right. \\
&\left. + \mathcal{P}_{s,2} \sin\left(\frac{2\pi sn}{\mathfrak{N}'}\right)\right], \quad \mathfrak{N}' \equiv \mathfrak{N} - R + 1.
\end{aligned} \tag{A12}$$

Here the normal mode index $q = (s,j)$, where $j = 1, 2$. Therefore, if $q = (s,1)$ and $\mathcal{P}_h = \Delta_q$, then

$$\dot{x}(n,0) = \left(\frac{2}{m\mathfrak{N}'}\right)^{1/2} \cos\left(\frac{2\pi sn}{\mathfrak{N}'}\right), \tag{A13}$$

and if $q = (s,2)$ when $\mathcal{P}_h = \Delta_q$,

then

$$\dot{x}(n,0) = \left(\frac{2}{m\mathfrak{N}'}\right)^{1/2} \sin\left(\frac{2\pi sn}{\mathfrak{N}'}\right). \tag{A14}$$

The normal mode frequencies are

$$\omega_q = \sin(\pi s/\mathfrak{N}'), \quad j = 1, 2. \tag{A15}$$

From Eqs. (A3), (A7), (A13), and (A14) we find that the displacement $X_{r',q}(\tau)$ is

$$X_{r',(sj)}(\tau) = \left(\frac{2}{M_r\mathfrak{N}'}\right)^{1/2} \sum_{n=-\mathfrak{N}}^{-R} [W^{-1/2} \sin(W^{1/2}\tau)]_{rn} \times \begin{Bmatrix} \cos\left(\frac{2\pi sn}{\mathfrak{N}'}\right) \\ \sin\left(\frac{2\pi sn}{\mathfrak{N}'}\right) \end{Bmatrix}, \tag{A16}$$

where the upper of the two trigonometric functions in curly brackets $\{\cdots\}$ is appropriate to $j = 1$, the lower to $j = 2$.

From the definition of $\mathcal{J}^{(h)}_{rr'}(\tau)$ and Eq. (A16),

$$\begin{aligned}
\mathcal{J}_{rr'}(\tau) = \tfrac{1}{8} \sum_{s=1}^{\mathfrak{N}'/2} &\hbar\omega_s \coth\left(\frac{\frac{1}{2}\hbar\omega_s}{kT_h}\right)\left(\frac{2}{\mathfrak{N}'}\right) \\
&\times \sum_{n=-\mathfrak{N}}^{-R} \sum_{n'=-\mathfrak{N}}^{-R} \left[\cos\left(\frac{2\pi sn}{\mathfrak{N}'}\right) \cos\left(\frac{2\pi sn'}{\mathfrak{N}'}\right)\right. \\
&\left. + \sin\left(\frac{2\pi sn}{\mathfrak{N}'}\right) \sin\left(\frac{2\pi sn'}{\mathfrak{N}'}\right)\right] \{[W^{-1/2} \sin(W^{1/2}\tau)]_{r'n} \\
&\times [\cos(W^{1/2}\tau)]_{rn'} - \omega_s^{-2} [\cos(W^{1/2}\tau)]_{r'n} \\
&\times [W^{1/2} \sin(W^{1/2}\tau]_{rn'}\}.
\end{aligned} \tag{A17}$$

At this point we immediately proceed to the limit as $\mathfrak{N} \to \infty$. In this limit the discrete argument $q = 2\pi s/\mathfrak{N}'$ becomes continuous in the interval $[0, \pi]$ as s ranges over $[1, \frac{1}{2}\mathfrak{N}']$. That is to say, the sum over s becomes an integral over q:

$$\lim_{\mathfrak{N}\to\infty} \frac{2}{\mathfrak{N}'} \sum_{s=1}^{\mathfrak{N}'/2} f\left(\frac{2\pi s}{\mathfrak{N}'}\right) = \frac{1}{\pi} \int_0^\pi dq f(q). \tag{A18}$$

Hence, for $\mathfrak{N} \to \infty$, and τ finite,

$$\begin{aligned}
\mathcal{J}^{(h)}_{rr'}(\tau) = \frac{1}{8\pi} \int_0^\pi dq \hbar\omega_q \coth(\tfrac{1}{2}\hbar\omega_q/kT_h) &\sum_{n=-\infty}^{-R} \sum_{n'=-\infty}^{-R} \\
&\times \{[W^{-1/2} \sin(W^{1/2}\tau)]_{r'n} [\cos(W^{1/2}\tau)]_{rn'} \\
&- \omega_q^{-2} [\cos(W^{1/2}\tau)]_{r'n} [W^{1/2} \sin(W^{1/2}\tau)]_{rn'}\} \\
&\times [\cos(qn)\cos(qn') + \sin(qn)\sin(qn')].
\end{aligned} \tag{A19}$$

Again we make a physical interpretation of the matrix elements of Eq. (A19) and use results derived previously.[1]

From Eq. (A2) we note that

$$X_{\text{trig}}(r',\tau) = \sum_{n=-\infty}^{-R} [W^{-1/2} \sin(W^{1/2}\tau)]_{r'n} \text{ trig}(qn), \tag{A20}$$

where trig(qn) is either $\sin(qn)$ or $\cos(qn)$, is the displacement of the rth atom at time τ when the initial, classical conditions of the lattice are

$$\begin{aligned}
&X(n,0) = 0, && \text{all } n, \\
&\dot{X}(n,0) = \text{trig}(qn), && n \le -R, \\
&\dot{X}(n,0) = 0, && n > -R.
\end{aligned} \tag{A21}$$

when $r' > A_N$ this displacement, $X_{\text{trig}}(r'\tau)$ is[1]

$$\begin{aligned}
X_{\text{trig}}(r'\tau) = \frac{1}{2\pi i} \int_L dp \frac{e^{p\tau}}{p(1+p^2)^{1/2} D_N(p)} &\\
\times \sum_{n=-\infty}^{-R} E^{2(r'-n)}(p) \text{ trig}(qn), &
\end{aligned} \tag{A22}$$

where L is a contour in the complex p-plane to the right of all singularities of the p-dependent integrand,

$$\begin{aligned}
&E(p) \equiv [p + (1+p^2)^{1/2}]^{-1}, \\
&D_N(p) = \det\left(\delta_{r,s} + \frac{Qp}{(1+p^2)^{1/2}} E^{2|A_r - A_s|}\right),
\end{aligned}$$

$$Q = \left(\frac{M}{m} - 1\right). \tag{A23}$$

When trig = cos,

$$X_{\cos}(r'\tau) = \frac{1}{2\pi i}\int_L dp\, \frac{e^{p\tau}E^{2(r'+R)}}{p(1+p^2)^{1/2}D_N(p)} \times \left[\frac{\cos(qR)}{2}\left(\frac{1}{1-E^2e^{iq}} + \frac{1}{1-E^2e^{-iq}}\right) - \frac{\sin(qR)}{2i}\left(\frac{1}{1-E^2e^{iq}} - \frac{1}{1-E^2e^{-iq}}\right)\right], \tag{A24}$$

and when trig = sin,

$$X_{\sin}(r',\tau) = \frac{-1}{2\pi i}\int_L dp\, \frac{e^{p\tau}E^{2(r'+R)}}{p(1+p^2)^{1/2}D_N(p)} \times \left[\frac{\sin(qR)}{2}\left(\frac{1}{1-E^2e^{iq}} + \frac{1}{1-E^2e^{-iq}}\right) + \frac{\cos(qR)}{2i}\left(\frac{1}{1-E^2e^{iq}} - \frac{1}{1-E^2e^{-iq}}\right)\right]. \tag{A25}$$

The integrand in either case has branch points at $p = \pm i$, and simple poles at

$$p = 0, \quad p = \pm i \sin^{-1}(\tfrac{1}{2}q) = \pm i\omega_q. \tag{A26}$$

If the two branch points are connected by a cut which lies to the left of the imaginary p-axis, we may make the following arguments for large τ: The contour L is deformed to encircle the branch cut and isolated poles on the imaginary p-axis. As τ increases without bound, the contribution to the integral from the branch cut contour vanishes because $e^{p\tau}$ falls exponentially to zero if $\mathrm{Re}\{p\} < 0$. Thus only the simple isolated poles on the imaginary p-axis determine $X_{\mathrm{trig}}(r',\tau)$ as $\tau \to \infty$.

After the evaluation of the residues at these poles, the displacements are

$$X_{\cos}(r'\tau) = \tfrac{1}{2}\cos(qR) + \tfrac{1}{2}\frac{\mathcal{T}_N(\omega_q)}{\omega_q}\sin[\omega_q\tau - qr' - \psi_N(q)], \tag{A27}$$

$$X_{\sin}(r',\tau) = -\tfrac{1}{2}\sin(qR) + \tfrac{1}{2}\frac{\mathcal{T}_N(\omega_q)}{\omega_q}\cos[\omega_q\tau - qr' - \psi_N(q)],$$

where we have used[1]

$$D_N^{-1}(i\omega_q) = \mathcal{T}_N(\omega_q)e^{-i\psi_N(q)}. \tag{A28}$$

Here $\mathcal{T}_N(\omega_q)$ is the transmitted amplitude of a wave of unit amplitude and frequency ω_q incident on the array of N defects. The phase shift is $\psi_N(q)$.

Finally, from the use of Eq. (A27), (A19), and (A20), we derive

$$\lim_{\tau\to\infty} \mathcal{J}^{(h)}_{r,r'}(\tau) = \frac{1}{16\pi}\int_0^\pi dq\hbar\omega_q \frac{\mathcal{T}_N^2(\omega_q)}{\omega_q}\sin[q(r-r') \times \coth\frac{\tfrac{1}{2}\omega_q}{kT_h}, \tag{A29}$$

which is independent of τ and R and only depends on r and r' through their difference.

Since

$$\lim_{\tau\to\infty}\langle J_N(r,\tau)\rangle = \lim_{\tau\to\infty}[\mathcal{J}^{(h)}_{r\,r-1}(\tau) - \mathcal{J}^{(h)}_{r,r}(\tau)], \tag{A30}$$

The asymptotic contribution to the heat current past atom $r > A_N$ is

$$\langle J_N(r)\rangle = \frac{1}{16\pi}\int_0^\pi dq\hbar\mathcal{T}_N^2(\omega_q)\coth\left(\frac{\tfrac{1}{2}\hbar\omega_q}{kT_h}\right)\sin q = \frac{1}{2\pi}\int_0^1 d\omega\mathcal{T}_N^2(\omega)\hbar\omega\left(\frac{1}{e^{\hbar\omega/kT_h}-1} + \tfrac{1}{2}\right). \tag{A31}$$

Since $\langle J_N(r)\rangle$ is independent of r, the heat current is constant everywhere in vicinity of defect region; i.e., a steady state has developed. Obviously the current arising from the region initially at temperature T_c is the same as Eq. (A31), except for a change of sign and replacement of T_h by T_c. Thus the steady state heat current across an N-defect lattice connected to two semi-infinite harmonic lattice heat baths at temperatures T_h and T_c is

$$J_N = \frac{1}{2\pi}\int_0^1 d\omega\hbar\omega\mathcal{T}_N^2(\omega)\left(\frac{1}{e^{\hbar\omega/kT_h}-1} - \frac{1}{e^{\hbar\omega/kT_c}-1}\right). \tag{A32}$$

With the interpretation of

$$\langle n(\omega,\tau)\rangle \equiv (e^{\hbar\omega/kT} - 1)^{-1}$$

as the mean number of phonons of frequency ω present at temperature T, the physical meaning of the integrand in Eq. (A32) is clear.

---

* NRC–NBS Postdoctoral Research Associate, 1969–71.
† Present Address: Institute for Molecular Physics, University of Maryland, College Park, Md. 20742.

1 R. J. Rubin and W. L. Greer, J. Math. Phys. 12, 1686 (1971).
2 R. Kubo, J. Phys. Soc. Japan 12, 570 (1957).
3 M. S. Green, J. Chem. Phys. 20, 1281 (1952); 22, 398 (1954).
4 K. R. Allen and J. Ford, Phys. Rev. 176, 1046 (1968).
5 Unpublished calculation by the authors and private communication from Dr. W. M. Visscher.
6 R. J Rubin, Phys. Rev. 131, 964 (1963).
7 R. J. Rubin, J. Math. Phys. 9, 2252 (1968).
8 R. J. Rubin, J. Math. Phys. 11, 1857 (1970).
9 Z. Rieder, J. L Lebowitz, and E. Lieb, J. Math. Phys. 8, 1073 (1967).
10 I. M. Khalatnikov, *Introduction to the Theory of Superfluidity* (Benjamin, New York, 1965), Chap. 23.
11 C. Kittel, *Quantum Theory of Solids* (Wiley, New York, 1963), p. 17.

Velocity-Dominated Singularities in Irrotational Hydrodynamic Cosmological Models

E. P. T. Liang*
Department of Physics, University of California at Berkeley, Berkeley, California†
(Received 14 May 1971)

We consider irrotational perfect fluid solutions of the Einstein equations with an equation of state $p = \gamma\rho$. We define "velocity-dominated" singularities of these solutions, a notion previously introduced for dust models. We demonstrate explicitly that uniquely and invariantly defined inner metric tensor, extrinsic curvature tensor, and scalar bang-time function can be assigned to these singularities, as in the dust case. We study the effects of a time-varying equation of state and viscosities on these singularities, and show by order-of-magnitude estimates that they do not change the structure of the singularity provided $\gamma > 0$. Some known exact perfect fluid solutions, both homogereous and inhomogeneous, are listed as examples.

1. INTRODUCTION

In a previous paper,[1] we introduced "velocity-dominated" singularities of irrotational dust solutions of the Einstein equations. However, it is possible that primordial matter at the very beginning of the big bang was more complicated than dust, although recent investigations of the microscopic properties of hadronic matter[2,3] seem to indicate, that dust might be a good approximation to the state of matter up to a microsecond after the explosion. Moreover, if inhomogeneous singularities are to play any role in the construction of realistic models for quasars or galactic centers, it would be very naive and artificial to use dust as the matter source. The study of the effects of realistic matter on the structure and properties of the cosmological singularity (or singularities) is thus of interest.

The purpose of this paper is to generalize the notion of velocity-dominated singularity to solutions of Einstein equations with irrotational perfect fluid sources having a barotropic equation of state $p = \gamma\rho$. It turns out that in this case the Einstein equations can also be explicitly integrated to give the form of the metric near the singularity. Similarly, one can identify some functions of integration as the metric, extrinsic curvature tensor, and bang-time functions of the three-dimensional singularity manifold. The results show that the inclusion of pressure does not affect the behavior of Kasner-like singularities, in agreement with the findings of Lifshitz and Khalatnikov.[4] However, pressure does play a role in the structure of the Friedman-like singularities, a phenomenon well known in the homogeneous Friedman models. Moreover, it is interesting to note in the Kasner-like case that our analysis, which is done in comoving nongeodesic frames, shows the same kind of singularity structures as those found by Lifshitz and Khalatnikov, who use geodesic normal coordinates.[4]

What about more complicated states of matter? In Sec. III of this paper, we try to estimate the effects of viscosity, and a time-varying equation of state. We tentatively conclude that as far as the velocity-dominated approximation goes, they do not affect the structure of the singularity in the sense that if we ignore them in the first order, then the correction terms will only show up in higher orders. Of course, this kind of argument does not prove anything, and the question will only be settled when examples of exact analytic or numerical solutions are found.

In the final section, we present some example of exact solutions with perfect fluid sources. Inhomogeneous exact models are very hard to find, and except for the case of $p = \rho$, we are only able to show the existence of velocity-dominated singularities in the plane and spherical symmetric models through some ad hoc approximation scheme.

2. THE STRUCTURE OF THE SINGULARITY

The Einstein equations with perfect fluid sources are

$$G^\mu_\nu = -T^\mu_\nu, \quad \mu\nu = 0 \cdots 3, \quad a, b = 1 \cdots 3, \quad \mathrm{sgn}(-+++), \tag{1}$$

where

$$T^\mu_\nu = \rho u^\mu u_\nu + p h^\mu_\nu, \quad h^\mu_\nu \equiv \delta^\mu_\nu + u^\mu u_\nu,$$
$$h^\mu_\nu u_\mu = 0, \quad u^\mu u_\mu = -1.$$

Throughout this section we assume a barotropic equation of state for the matter $p = \gamma\rho$, where γ is a constant and $\rho > 0$. $\gamma = 0$ reduces to the dust case. $\gamma = \frac{1}{3}$ gives radiation or ultrarelativistic matter whereas $\gamma = 1$ corresponds to stiff matter (e.g., the Brans-Dicke bosons). Although the solutions of the velocity-dominated equations do not depend on specific values of γ from other considerations (e.g., the existence of the singularity,[5] causality, etc.), it is reasonable to restrict γ to the range $-\frac{1}{3} \leq \gamma \leq 1$. We also assume that the fluid is irrotational: $\omega_{\mu\nu} \equiv h^\alpha_\mu h^\beta_\nu u_{[\alpha;\beta]} = 0$. Then in "comoving normal" coordinates, the metric can be written as[6]

$$ds^2 = -e^{2\sigma(x^c,t)}dt^2 + g_{ab}(x^c, t)dx^a dx^b \tag{2}$$

and $u^\mu = e^{-\sigma}\delta^\mu_0$ is the unit vector field tangent to the $x^a = \text{const}$ lines. The remaining allowed coordinate transformations are

$$x'^a = x'^a(x^b), \quad t' = t'(t). \tag{3}$$

We shall denote ordinary spatial derivative by a comma (,), covariant derivative with respect to $g_{\mu\nu}$ by a semi colon (;), covariant derivative with respect to the spatial metric $g_{ab}(x^c, t = \text{const})$ by a stroke (|), and geometrical quantities of the $t = \text{const}$ hypersurfaces by a superscript 3.

The "field velocity"

$$K^a_b \equiv \tfrac{1}{2}g^{ac}\partial_t g_{cb} \tag{4}$$

is related to the second fundamental form Ω^a_b of the $t = \text{const}$ surfaces by $\Omega^a_b = e^{-\sigma}K^a_b$. K^a_b can be split into traceless and isotropic parts:

$$K^a_b = L^a_b + \tfrac{1}{3}\delta^a_b K, \quad K \equiv K^a_a = \partial_t(\ln\alpha),$$
$$\alpha \equiv (\mathrm{Det}\,|g_{ab}|)^{1/2} \Rightarrow L^a_a = 0; \tag{5}$$

and let $L^2 \equiv L^a_b L^b_a \geq 0$. Using metric (2) and the contracted Bianchi Identities $T^{\mu\nu}_{;\nu} = 0$ one obtains

(i) the conservation equation $u_\mu T^{\mu\nu}_{;\nu} = 0$

$$\Rightarrow \partial_t p = -(\rho + p)K, \tag{6}$$

(ii) the equations of motion $h_{a\mu} T^{\mu\nu}_{;\nu} = 0$

$$\Rightarrow \sigma_{,a} = -p_{,a}/(\rho + p). \tag{7}$$

Applying the equation of state $p = \gamma\rho$, we obtain the first integral of (7)

$$\sigma = -[\gamma/(\gamma + 1)] \ln\rho + f(t).$$

For convenience we set $f(t)$ equal to zero by rescaling t and write from now on

$$\sigma = -[\gamma/(\gamma + 1)] \ln\rho. \tag{8}$$

The freedom of arbitrary coordination transformations is restricted to

$$x'^a = x'^a(x^b), \qquad t' = t + \text{const.} \tag{3'}$$

Equation (6) also has a first integral

$$\rho = {}_0\rho(x^c)\alpha^{-(\gamma+1)}. \tag{9}$$

Let us now look at the Einstein equations. Using standard techniques (e.g., Eisenhart[7]) we compute

$$\begin{aligned} G^a_b = {}^3G^a_b &+ \sigma^{,a}\sigma_{,b} + \sigma^{;a}_{|b} - e^{-2\sigma}(\partial_t + K - \partial_t\sigma)L^a_b \\ &+ e^{-2\sigma}\delta^a_b[\tfrac{2}{3}(\partial_t + \tfrac{1}{2}K - \partial_t\sigma)K + \tfrac{1}{2}L^2 \\ &- e^{2\sigma}(\sigma^{,c}\sigma_{,c} + \sigma^{;c}_{|c})] = -p\delta^a_b, \end{aligned} \tag{10}$$

$$G^0_0 = \tfrac{1}{2}[e^{-2\sigma}(\tfrac{2}{3}K^2 - L^2) - {}^3R] = \rho, \tag{11}$$

$$G^0_a = e^{-\sigma}[e^{-\sigma}(K^b_a - \delta^b_a K)]_{|b} = 0. \tag{12}$$

Adding (10) to (11) multiplied by $\gamma\delta^a_b$ we obtain what we shall call the "evolution equations" free of matter terms:

$$\begin{aligned} {}^3G^a_b - \tfrac{1}{2}{}^3R\gamma\delta^a_b &+ \sigma^{,a}\sigma_{,b} + \sigma^{;a}_{|b} - e^{-2\sigma}(\partial_t + K - \partial_t\sigma)L^a_b \\ &+ e^{-2\sigma}\delta^a_b\{\tfrac{2}{3}[\partial_t + \tfrac{1}{2}(1+\gamma)K - \partial_t\sigma]K + \tfrac{1}{2}(1-\gamma)L^2 \\ &- e^{2\sigma}(\sigma^{,c}\sigma_{,c} + \sigma^{;c}_{|c})\} = 0, \end{aligned} \tag{13a}$$

in which

$$\partial_t\sigma = \gamma K \tag{13b}$$

is obtained by substituting (8) into (6). Equation (13b) has the integral

$$\sigma = {}_0\sigma(x^c) + \gamma \ln\alpha. \tag{13c}$$

Equations (13) define a closed system for the unknown functions g_{ab} and σ when combined with Eq. (4). As is already shown in Ref. 1, they can in principle be converted into a very complicated, covariant, but not very useful functional equation for g_{ab}. ρ can then be computed from (11).

Now from the following generalization of Raychaudhuri's equation[8] to include pressure gradients

$$\begin{aligned} \dot\theta = &-\sigma^{\mu\nu}\sigma_{\mu\nu} - \tfrac{1}{3}\theta^2 - \tfrac{1}{2}(\rho + 3p) - [1/(\rho + p)] \\ &\times (\Box^2 p + \ddot p + 2\dot p\theta) + [1/(\rho + p)^2][\dot p^2 + p^{,\mu}(\rho + p)_{,\mu}], \end{aligned}$$

where

$$\theta \equiv u^\mu_{;\mu}, \qquad \cdot \equiv u^\mu\partial_\mu, \qquad \Box^2 \equiv \nabla_\mu\partial^\mu,$$
$$\sigma_{\mu\nu} \equiv h^\alpha_\mu h^\beta_\nu u_{(\alpha;\beta)} - \tfrac{1}{3}\theta h_{\mu\nu},$$

one checks that if along a particular matter line the pressure gradients (spatial) are small and $\rho + 3p \geq 0$, so that the right-hand side remains negative, then $\dot\theta < 0$ always implies $\theta \to \infty$ for some $t = {}_0t$, so that a physical singularity (at which curvature invariants blow up) exists at ${}_0t$. Whether this actually happens to all matter lines in exact solutions depends on the specific initial data. But suppose it does, so that $t = {}_0t(x^c)$ defines a three-dimensional singularity manifold, then we are interested in characterizing the structure and properties of the metric near such a singularity. We call ${}_0t(x^c)$ the bang time and assume that it is C^∞, and we want to construct a differentiable manifold out of $t = {}_0t(x^c)$, which we will call the singularity. Following the velocity-dominated approach for the dust models, the essential idea here is to drop the spatial curvature and pressure gradient terms in Eqs. (13), and use the solution as a first approximation to g_{ab} near the singularity. We first generalize the definition of velocity-dominated singularity introduced in the previous paper.

Let $\hat g_{ab}$ with corresponding $\hat K^a_b$, $\hat\sigma$ and bang-time function ${}_0\hat t \equiv {}_0t$ be an exact perfect fluid solution. Suppose there is another g_{ab} which in some neighborhood of ${}_0t$ satisfies (a) $g_{ab} = g_{ba}$, $\operatorname{sgn}(g_{ab}) = +3$; (b) g_{ab} obeys (4) and

$$\begin{aligned} [\partial_t + (1-\gamma)K]L^a_b = \delta^a_b\{&\tfrac{2}{3}[\partial_t + \tfrac{1}{2}(1-\gamma)K]K \\ &+ \tfrac{1}{2}(1-\gamma)L^2\}; \end{aligned} \tag{14}$$

(c) there exists some component, say W, of g_{ab} such that for fixed x^c and $t \to {}_0t$,

$$\begin{aligned} (\hat K^a_b - K^a_b)(K^c_d K^d_c)^{-\frac{1}{2}} \to 0,\ \hat\sigma - \sigma \to 0, \\ (\hat g_{ab} - g_{ab})W^{-1} \to 0; \end{aligned} \tag{15}$$

Then we call $\hat g_{ab}$ "velocity-dominated" and call g_{ab} its first approximation. In order that a solution of (4) and (14) be a first approximation, the consistency conditions

$$e^{2\sigma}\,{}^3R^a_b(K^c_d K^d_c)^{-1} \underset{t\to{}_0t}{\longrightarrow} 0, \tag{16a}$$

$$e^{2\sigma}(\sigma^{,a}\sigma_{,b} + \sigma^{;a}_{|b})(K^c_d K^d_c)^{-1} \underset{t\to{}_0t}{\longrightarrow} 0, \tag{16b}$$

together with $\rho > 0$ in (11), and Eq. (12) must be satisfied. These will place restrictions on g_{ab}. The above definition is invariant under the transformations (3).

The task now is to integrate Eqs. (4) and (14) for g_{ab}, K^a_b, and σ, and check consistency conditions (12) and (16). If everything is all right, then putting the solutions back into (13) will generate a second-order approximation, and so on. ρ and p can be computed order by order from Eq. (11).

Setting the traceless and isotropic parts of Eq. (14) to zero separately gives

$$[\partial_t + (1-\gamma)K]L^a_b = 0, \tag{17a}$$

$$[\partial_t + \tfrac{1}{2}(1-\gamma)K]K = -\tfrac{3}{4}(1-\gamma)L^2. \tag{17b}$$

We consider two different cases.

Case A: $\gamma \neq 1$. The integral of (17a) is

$$L^a_b = {}_0M^a_b(x^c)\alpha^{\gamma-1}, \tag{18}$$

where ${}_0M^a_b$ is some arbitrary symmetric traceless 3-tensor density. Substituting solution (18) into (17b) we obtain a second-order equation for α:

$$\alpha\frac{\partial^2\alpha}{\partial t^2} - \frac{1+\gamma}{2}\left(\frac{\partial\alpha}{\partial t}\right)^2 + \tfrac{3}{4}(1-\gamma){}_0M^2\alpha^{2\gamma} = 0, \qquad {}_0M^2 \equiv {}_0M^a_b\,{}_0M^b_a \geq 0, \tag{19}$$

which has the general solution

$$\alpha = {}_0\alpha(x^c)\{[t - {}_0t(x^c)][t - {}_0t'(x^c)]\}^{1/(1-\gamma)}, \tag{20}$$

where ${}_0\alpha$, ${}_0t$, ${}_0t'$ are arbitrary functions of integration related to ${}_0M$. Putting (20) back into (18), (13c), (5), (4) and integrating give

$$K^a_b = {}_0K^a_b(x^c)\left(\frac{1}{(t-{}_0t)} - \frac{1}{(t-{}_0t')}\right) + \frac{2}{3(1-\gamma)} \times \delta^a_b\,\frac{1}{t-{}_0t'},$$

$${}_0K^a_b\,{}_0K^b_a = \frac{1}{(1-\gamma)^2}, \qquad {}_0K^a_a = \frac{1}{1-\gamma},$$

$$g_{ab} = {}_0g_{ac}(x^c)\exp 2\left({}_0K^c_b[\ln(t-{}_0t) - \ln(t-{}_0t')] + \frac{2}{3(1-\gamma)}\delta^c_b\ln(t-{}_0t')\right), \tag{21}$$

$$\sigma = {}_0\sigma(x^c) + [\gamma/(1-\gamma)]\ln[(t-{}_0t)(t-{}_0t')].$$

Following the terminology of our last paper, we call this the "Heckmann–Schucking-like"[1] solution. When ${}_0t \to {}_0t'$, we obtain the Friedman-like[1] solution

$$g_{ab} = {}_0g_{ab}(t-{}_0t)^{4/[3(1-\gamma)]}, \quad K^a_b = \frac{2}{3(1-\gamma)}\delta^a_b\frac{1}{(t-{}_0t)},$$
$$\sigma = {}_0\sigma + \frac{2\gamma}{(1-\gamma)}\ln(t-{}_0t), \quad \alpha = {}_0\alpha(t-{}_0t)^{2/(1-\gamma)}, \tag{22}$$

Also we get back the Kasner-like[1] solution by letting

$${}_0g_{ab} = {}_0g'_{ac}\exp(2\{{}_0K^c_b - [1/3(1-\gamma)]\delta^c_b\}\ln(-{}_0t'))$$

and then taking the limit ${}_0t' \to -\infty$ (after dropping the prime):

$$g_{ab} = {}_0g_{ac}\exp[2\,{}_0K^c_b\ln(t-{}_0t)], \quad K^a_b = {}_0K^a_b(t-{}_0t)^{-1},$$
$$\sigma = {}_0\sigma + [\gamma/(1-\gamma)]\ln(t-{}_0t), \quad \alpha = {}_0\alpha(t-{}_0t)^{1/(1-\gamma)}. \tag{23}$$

Note that ${}_0\sigma$ is not arbitrary but is determined by (8) and (11) when one computes the first nonvanishing coefficient of ρ, whose order should be consistent with the exact relation (9). If we now express the proper time τ of a particular matter line, say $x^c =$ const, in terms of the coordinate time t, then we get, for the Kasner-like case (everything being evaluated at $x^c =$ const)

$$\tau - {}_0\tau \sim (t-{}_0t)^{1/(1-\gamma)} \tag{24}$$

so that Eqs. (23) become

$$\alpha \sim (\tau - {}_0\tau), \quad K^a_b \sim {}_0K^a_b(\tau-{}_0\tau)^{-(1-\gamma)} \Rightarrow \Omega^a_b = {}_0\Omega^a_b(\tau-{}_0\tau)^{-1} \quad \text{and} \quad \times g_{ab} \sim {}_0g_{ac}\exp[2\,{}_0\Omega^c_b\ln(\tau-{}_0\tau)], \tag{23'}$$

where

$${}_0\Omega^a_b \equiv {}_0K^a_b(1-\gamma)$$

is such that

$${}_0\Omega^a_b\,{}_0\Omega^b_a = {}_0\Omega^a_a = 1.$$

Thus, the local behavior is the same as in the dust model, independent of the value of γ. We can therefore say, that pressure does not affect the structure of Kasner-like singularities. On the other hand, for the Friedman-like case, we have

$$\tau - {}_0\tau \sim (t-{}_0t)^{(1+\gamma)/(1-\gamma)}, \tag{25}$$

and Eqs. (22) become

$$\alpha \sim (\tau-{}_0\tau)^{2/(1+\gamma)}, \quad K^a_b \sim \frac{2}{3(1-\gamma)}\delta^a_b(\tau-{}_0\tau)^{-(1-\gamma)/(1+\gamma)} \Rightarrow \Omega^a_b = \frac{2}{3(1+\gamma)}\delta^a_b(\tau-{}_0\tau)^{-1} \quad \text{and} \quad g_{ab} \sim (\tau-{}_0\tau)^{4/[3(1+\gamma)]}. \tag{22'}$$

Thus, we see that in terms of the local proper time of the comoving observers, positive pressure ($\gamma > 0$) has the effect of lowering the power of the time dependence of the metric of the Friedman-like solutions.

Both ${}_0g_{ab}$ and ${}_0K^a_b$ transform as 3-tensors under (3a) and are uniquely defined for t fixed as in (8). From their definitions, we see that they can be simultaneously diagonalized at any one point of the singularity by a local real coordinate transformation.

Case B: $\gamma = 1$. In this case we simply combine (17a) and (17b) and write

$$\partial_t K^a_b = 0 \tag{26}$$

The solution is

$$K^a_b = {}_0K^a_b(x^c), \quad g_{ab} = {}_0g_{ac}(x^c)\exp 2\{{}_0K^c_b t\}, \quad \alpha = {}_0\alpha\,\exp{}_0Kt, \quad \sigma = {}_0\sigma + {}_0Kt, \quad {}_0K \equiv {}_0K^a_a, \quad {}_0\alpha \equiv (\det|{}_0g_{ab}|)^{1/2}. \tag{27}$$

For definiteness we assume ${}_0K(x^c) > 0$ in the following, so that a singularity exists for all world lines at $t = -\infty$. To see the structure of the singularity more

clearly, we redefine the time coordinate as $t = e^{\bar{t}}$ or $\bar{t} = \ln t$. Then in the new coordinate system we have

$$ds^2 = -e^{2\bar{\sigma}(x,\bar{t})}d\bar{t}^2 + \bar{g}_{ab}(x,\bar{t})dx^a\,dx^b,$$
$$\bar{\sigma} = {}_0\sigma + ({}_0K - 1)\ln\bar{t}, \qquad \alpha = {}_0\alpha\bar{t}^{\,{}_0K}, \tag{28}$$
$$\bar{g}_{ab} = {}_0g_{ac}\exp(2\,{}_0K^c_b\ln\bar{t}).$$

The singularity is now shifted to $\bar{t} = 0$ and the allowed coordinate transformations remain the same as in (3). On the other hand, if we again express everything in terms of the proper time τ of a particular comoving observer, then at, say $x^c = \text{const}$, we have

$$\tau - {}_0\tau \sim \bar{t}^{\,{}_0K} \tag{29}$$

and solutions (28) become

$$\alpha \sim (\tau - {}_0\tau), \qquad g_{ab} \sim {}_0g_{ac}\exp[2\,{}_0\Omega^c_b\ln(\tau - {}_0\tau)], \tag{28'}$$

where

$${}_0\Omega^a_b \equiv {}_0K^a_b/{}_0K$$

is such that

$${}_0\Omega^a_a = 1.$$

Equations (28′) differ from (23′), in that the condition ${}_0\Omega^a_b\,{}_0\Omega^b_a = 1$ is relaxed. We call this type of singularity "semi-Kasner-like". With $\bar{t}$ fixed as in (28), ${}_0g_{ab}$ and ${}_0K^a_b$ are again uniquely defined and have similar properties (except for the relaxation of the constraints on ${}_0K^a_b$) as in the previous cases. We now identify them as the metric and extrinsic curvature tensor of the three-dimensional singularity manifold for all cases. In the semi-Kasner-like case, however, the bang-time function is amazingly constrained to be a constant, so that the price of relaxing the constraints on ${}_0K^a_b$ here is to lose the two arbitrary functions ${}_0t$ and ${}_0t'$, at least in the first order.

The task of checking the consistency conditions and estimating correction terms is very similar to the dust case. We omit the details in the following and discuss briefly the main results. We consider only the Friedman-like case and the Kasner-like case. (The Heckmann-Schucking-like case is completely equivalent to the Kasner-like case near any one of its singularities.) The semi-Kasner-like case, being pathological, remains to be studied in more detail. For simplicity, we analyze everything in an open region on which one of the solutions (22) and (23) holds.

From the Bianchi identities $G^{\mu}_{\nu;\mu} = 0$, we know that if a metric satisfies Eqs. (4), (10), and (11) together with $p = \gamma\rho$, then

$$G^0_a = {}_0G^0_a(x^c)e^{-\sigma}\alpha^{-1}, \tag{30}$$

where ${}_0G^0_a$ are functions of integration. Moreover, if (12) is also satisfied, then

$$G^0_0 = {}_0G^0_0(x^c)\alpha^{-(1+\gamma)}. \tag{9'}$$

In the Friedman-like case, using solutions (22), Eq. (12) becomes

$$G^0_a = \left[(t - {}_0t)^{-2}\,{}_0t_{,a}\left(-\frac{4}{3(1-\gamma)} - \frac{8\gamma}{3(1-\gamma)^2}\right) + (t - {}_0t)^{-1}\right.$$
$$\left.\times\left(\frac{4}{3(1-\gamma)}\,{}_0\sigma_{,a}\right)\right]e^{-2\,{}_0\sigma}(t - {}_0t)^{-4\gamma/(1-\gamma)}, \tag{31}$$

while (11) gives ρ to the lowest order:

$$\rho = G^0_0 = [\tfrac{4}{3}(1-\gamma)^2]^{(1+\gamma)/(1-\gamma)}(t - {}_0t)^{-2(1+\gamma)/(1-\gamma)}. \tag{32}$$

Thus (32) and (8) give ${}_0\sigma$ a constant, so that the second term in Eq. (31) vanishes identically. Then from (12) and (30), we see that we have to set the first term in (31) equal to zero. Hence we need

$${}_0t_{,a}(x^c) = 0. \tag{33}$$

Assuming (33) one obtains for the 3-Ricci-tensor

$${}^3R^a_b = {}^3_0R^a_b(x^c)(t - {}_0t)^{-4/3(1-\gamma)}, \tag{34a}$$

where ${}^3_0R^a_b$ is the Ricci tensor formed with ${}_0g_{ab}$. Similarly, the pressure gradient terms give

$$\sigma^{,a}\sigma_{,b} + \sigma_{|b}{}^{|a} = {}_0g^{ac}\,{}_0\sigma_{,c\|b}(t - {}_0t)^{-4/3(1-\gamma)}, \tag{34b}$$

where from now on we denote covariant derivative with respect to ${}_0g_{ab}$ by a double stroke $\|$. Equations (34) satisfy conditions (16) iff $\gamma > -\frac{1}{3}$, which is satisfied by all known or conceivable states of matter.[2] Assuming this and putting (34) back into (13) gives second-order approximations which can be checked to be consistent with (15). Since now both ${}_0\sigma$ and ${}_0t$ are constants, the matter lines are geodesics to the lowest order and therefore our Friedman-like case is completely equivalent to the isotropic case of Lifshitz and Khalatnikov.[4]

The Kasner-like case is more involved and one has to resort to the use of Cartan frames and stretched metrics developed in our last paper.[1] If $\tau^a_A(x^c)$ are the unit eigenvectors of ${}_0K^a_b$ orthonormal with respect to ${}_0g_{ab}$ ($A, B = 1, 2, 3$ not subject to summation convention), we can write

$${}_0K^a_b = \sum_A \frac{P_A}{1-\gamma}\,\tau^a_A\tau_{Ab}, \tag{35}$$

where

$$\tau_{Aa} \equiv {}_0g_{ab}\tau^b_A,$$

etc. Conditions (21) now require

$$\sum_A P_A = \sum_A P_A^2 = 1. \tag{36}$$

For convenience, we arrange P_A according to the convention $1 \geq P_1 \geq P_2 \geq 0 \geq P_3 \geq -\frac{1}{3}$. Substituting solutions (23) into the left-hand side of (12) gives a leading term identically equal to zero, and the first nonvanishing term is

$$\tau^a_B G^0_a = (t - {}_0t)^{-2\gamma/(1-\gamma)}\{{}_0K^a_{b\|a}\tau^b_B$$
$$+ [(1-P_B)/(1-\gamma)]\,{}_0\sigma_{,b}\tau^b_B\}(t - {}_0t)^{-1}e^{-2\,{}_0\sigma}, \tag{37}$$

having the same order as given by (30). However, in this case we cannot simply set the coefficients equal to zero because higher-order corrections [e.g., $\sim \ln(t - {}_0t)$] to the solutions (23) will in general con-

tribute to this term. But in any case ${}_0G^0_a = 0$ are three constraint conditions on the metric of the singularity. On the other hand, ρ to the first order is zero from Eq. (11), so the contribution to ρ must come from higher-order terms, the coefficients of which are in general functions of spatial variables. Thus, ${}_0\sigma$ of the solutions (23) will in general be spatially dependent and the matter lines are *not* geodesics even in the lowest order. In this sense, our analysis differs from that of Lifshitz and Khalatnikov[4] who use normal geodesic congruence as the time lines, although the results agree essentially.

The conditions (16) can be checked using (35), etc. and the formulas derived in the Appendix of our last paper.[1] For ${}_0t$ constant and $\mathbf{P} \equiv (P_1, P_2, P_3) \neq (1, 0, 0)$, one again obtains the hypersurface orthogonal condition[4] for τ_{3a}:

$$\tau_{31,2} - \tau_{32,1} = 0. \tag{38}$$

For ${}_0t \neq$ const, many other conditions arise as in the dust models, e.g.,

$${}_0t_{,a}\tau^a_A = 0, \quad A = 1, 2 \tag{39}$$

plus new conditions from (16b), so that ${}_0t \neq$ const does not give any extra arbitrary function of three variables. For $\mathbf{P} = (1, 0, 0)$ even for ${}_0t =$ const a variety of conditions results. For example, (16b) requires

$$({}_0\sigma_{,a}\ {}_0\sigma_{,b} + {}_0\sigma_{,a\|b})\tau^a_1\tau^b_1 = 0. \tag{40}$$

The meaning and interpretations of these numerous conditions for all cases are being investigated. In any case, once all consistency conditions are satisfied, then ${}^3R^a_b$ and $\sigma'^a\sigma_{,b} + \sigma^{\prime a}_{|b}$ put back into (13) will generate the second-order approximation.

What is the degree of generality of these solutions?

To look into this, let us first recall the number counting of the essential arbitrary functions for the generic case. To get any general irrotational perfect fluid solution with a specified equation of state $p = p(\rho)$, one has to give g_{ab} and K^a_b on one spacelike hypersurface $t =$ const subject to the three constraints (12). The initial data for σ are not arbitrary but are related through the integrated Bianchi identity (8) to the initial value of ρ, which in turn is constrained by Eq. (11). Let us also assume that the requirement $\rho > 0$ does not affect the counting. Then one has 12 functions of three variables subject to three constraints. Taking into account the three coordinate transformations (3), one ends up with six essential functions in general. On the other hand, the Heckmann–Schücking-like solutions with ${}_0t =$ const are generated by ${}_0g_{ab}$, ${}_0K^a_b$, and ${}_0t'$ subject to ${}_0G^0_b = 0$ in (12). This leaves four essential functions of three variables. The other cases are not more general. We thus conclude that the general irrotational perfect fluid solution is *not* velocity-dominated. In particular, the very interesting mixmaster–like solutions[9] do not seem to fall into our category.

3. OTHER STATES OF MATTER

Because of the limited knowledge one has at present about the properties of matter near a big bang, it would be more realistic to consider other plausible states of matter different from that of an irrotational perfect fluid with the simple equation of state $p = \gamma\rho$. In the following, we try to consider the effects of (i) a time-varying equation of state $p = \gamma(t)\rho$, and (ii) viscosity (shear and bulk), again by the method of estimating the order of magnitudes of the correction terms. Models with rotations, being more involved, merit an independent investigation.

A. Time-Varying Equation of State

If the recent theories of hadronic matter[2,3] are correct, then the ratio of pressure to matter energy density is essentially zero near the big bang, but gradually increases as the universe expands until it reaches the ultrarelativistic value of $p = \frac{1}{3}\rho$ in the lepton era.[2] One way to take into account this property of matter is to consider it as being described by a time-varying equation of state. (We ignore the spatial variation of γ for simplicity.) A reasonable theoretical model to incorporate this is to consider γ as dependent on the spatial volume element α through some power law

$$\gamma = {}_0\gamma\alpha^n, \quad {}_0\gamma \text{ const}, \quad n \text{ any real no.} > 0, \tag{41}$$

so that $\gamma \to 0$ as $\alpha \to 0$ at the singularity. Repeating the analysis of Sec. I, but dropping γ compared to 1, whenever $\gamma + 1$ appears, since we are only interested in the region $\alpha \to 0$, we obtain from (6)–(9),

$$\partial_t\sigma = \partial_t(\gamma \ln\alpha) \tag{42}$$

instead of (13b), while (13a) remains valid.

Putting (42) into (14), the traceless part can be integrated. The solution is again Eq. (18) with $\gamma = 0$. The trace of (14) now becomes, after plugging in solution (18) for L^a_b,

$$\alpha^{-1}\frac{\partial^2\alpha}{\partial t^2} - \tfrac{1}{2}\alpha^{-2}\left(\frac{\partial\alpha}{\partial t}\right)^2 - \alpha^{-1}\left(\frac{\partial\alpha}{\partial t}\right)\left(\frac{\partial\gamma}{\partial t}\right)\ln\alpha + \tfrac{3}{4}\,{}_0M^2\alpha^{-2} = 0, \tag{43}$$

which differs from the corresponding equation for dust by the additional third term. Suppose α has the first-order solution (22) or (23) (with $\gamma = 0$) near the singularity; then all other terms in (43) are $\sim (t - {}_0t)^{-2}$, whereas the third term is $\sim (t - {}_0t)^{-2+n} \ln(t - {}_0t)$ or $(t - {}_0t)^{-2+2n} \ln(t - {}_0t)$, which is of a little higher order as long as $n > 0$. Replacing (41) by other elementary functions (e.g., $(\ln\alpha)^{-1}$, $e^{-\alpha^{-1}}$, etc.) would not make the $\partial_t\gamma$ term any more dominant. Thus, it is *at least consistent* to ignore it in the first order. *In this sense*, we say that a time-varying γ with $\gamma \to 0$ as $\alpha \to 0$ has no essential effect on the structure of velocity-dominated singularities.

B. Viscosity

Let us first consider shear viscosity. T^μ_ν in Eq. (1) is changed to

$$T^\mu_\nu = \rho u^\mu u_\nu + p h^\mu_\nu - \lambda\sigma^\mu_\nu, \tag{44}$$

where the viscosity coefficient λ is assumed to be constant and $\sigma^\mu_\nu \equiv h^{\mu\alpha}h^\beta_\nu u_{(\alpha;\beta)} - \frac{1}{3}\theta h^\mu_\nu$ is the shear tensor.

Again going to comoving normal coordinates (2) we obtain

$$\sigma^a_b = e^{-\sigma}L^a_b, \qquad \sigma^0_a = \sigma^a_0 = \sigma^0_0 = 0. \tag{45}$$

The conservation equation is now

$$\partial_t\rho + (\rho + p)K - \lambda L^2 e^{-\sigma} = 0, \tag{46}$$

and the equations of motion $h_{a\mu}T^{\mu\nu}_{;\nu} = 0$ become

$$(\rho + p)\sigma_{,a} + p_{,a} - \lambda e^{-\sigma}L^c_{a\|c} = 0, \tag{47}$$

while the evolution equations (14) give

$$(\partial_t + K - \partial_t\sigma + \lambda e^{\sigma})L^a_b = 0, \tag{48a}$$

$$(\partial_t + \tfrac{1}{2}(1+\gamma)K - \partial_t\sigma)K + \tfrac{3}{4}(1-\gamma)L^2 = 0. \tag{48b}$$

Equations (46) and (47) cannot be integrated because of the extra viscosity terms. We now assume that they are of a higher order and drop them in the first approximation and then check consistency. Then Eqs. (6)–(9), (13b) and (13c) remain valid. Equations (13b) and (13c) put back into (48a) give the first integral for L^a_b:

$$L^a_b = {}_0M^a_b(x^c)\alpha^{\gamma-1}\exp\left(-\lambda e^{0\sigma}\int_{0t}^{t}\alpha^\gamma(t')dt'\right). \tag{49}$$

We see that $\gamma = 0$ gives the well-known exponential decay of the shear[10,11] due to viscosity. Since we are primarily interested in the region near $\alpha = 0$, ${}_0t$ is to be chosen as the bang time for convenience (this can always be done due to the remaining freedom in redefining ${}_0M^a_b$). From (49) we see that as $t \to {}_0t$, exp( ) $\to 1$ and L^a_b goes to solution (18), so that the first-order solutions are again (22) or (23). (We ignore the Heckmann–Schucking-like case because here we are only interested in the region near one of the singularities.) With solutions (22), $L^2e^{-\sigma}$ in Eq. (46) is $\sim (t - {}_0t)^{-2/(1-\gamma)}$, obviously of higher order than the other terms $[\sim (t - {}_0t)^{-(3+\gamma)/(1-\gamma)}]$.

With solutions (23), $L^2e^{-\sigma}$ is $\sim (t - {}_0t)^{(-2+\gamma)/(1-\gamma)}$, which is of order higher than the other terms $[\sim (t - {}_0t)^{-2/(1-\gamma)}]$ provided $\gamma > 0$. In Eq. (47) with solutions (23) the leading viscosity term is $\sim (t - {}_0t)^{(\gamma-2)/(1-\gamma)}$, while the other terms are $\sim (t - {}_0t)^{-2/(1-\gamma)}$; with solutions (22), Eq. (47) is identically satisfied to lowest order. Thus, dropping the viscosity terms in the lowest order is at least self-consistent as long as $\gamma > 0$.

Next let us consider bulk viscosity. T^μ_ν can be written as

$$T^\mu_\nu = \rho u^\mu u_\nu + ph^\mu_\nu - \eta\theta h^\mu_\nu, \qquad \theta \equiv u^\mu_{;\mu}, \tag{50}$$

where η the bulk viscosity coefficient is assumed constant. The conservation equation and equations of motions are (in comoving normal coordinates):

$$\partial_t\rho + (\rho + p)K = \eta e^{-\sigma}K^2, \tag{51}$$

$$(\rho + p)\sigma_{,a} + p_{,a} = \eta e^{-\sigma}K_{,a}. \tag{52}$$

Since these equations cannot be integrated, we again resort to estimates by first dropping the viscosity terms. Then Eqs. (6)–(9), (13b), and (13c) remain valid. The evolution equations (14) give in this case

$$(\partial_t + K - \partial_t\sigma)L^a_b = 0, \tag{53a}$$

$$[\partial_t + \tfrac{1}{2}(1+\gamma)K - \partial_t\sigma - \eta e^{\sigma}]K + \tfrac{3}{4}(1-\gamma)L^2 = 0. \tag{53b}$$

Solving (53a) as before and putting the solution into (53b) one obtains

$$\alpha^{-1}\frac{\partial^2\alpha}{\partial t^2} - \frac{(1+\gamma)}{2}\alpha^{-2}\left(\frac{\partial\alpha}{\partial t}\right)^2 + \tfrac{3}{4}(1-\gamma){}_0M^2\alpha^{2\gamma-2} - \tfrac{3}{2}\eta e^{\sigma_0}\alpha^{\gamma-1}\left(\frac{\partial\alpha}{\partial t}\right) = 0. \tag{54}$$

Here we again see that the viscosity term will be of higher order, if we assume a first-order solution for α using only the other three terms. In fact, the last term will be of higher order as long as $\alpha \to 0$ according to some power law, and $\gamma \geq 0$, because then the first three terms are $\sim (t - {}_0t)^{-2}$, whereas this term is $\sim (t - {}_0t)^{-1+n}$, where $n \geq 0$. Dropping this last term in (54) we again obtain the solutions (22) and (23). With solutions (22), $e^{-\sigma}K^2$ in (51) is $\sim (t - {}_0t)^{-2/(1-\gamma)}$, while the other terms are $\sim (t - {}_0t)^{-(3+\gamma)/(1-\gamma)}$. Equation (52) is identically satisfied to lowest order since ${}_0t_{,a} = 0$ for this case. With solutions (23), $e^{-\sigma}K^2$ is $\sim (t - {}_0t)^{(-2+\gamma)/(1-\gamma)}$, while the other terms in (51) are $\sim (t - {}_0t)^{-2/(1-\gamma)}$. Also $e^{-\sigma}K_{,a}$ is $\sim (t - {}_0t)^{(-2+\gamma)/(1-\gamma)}$ compared to the $(t - {}_0t)^{-2/(1-\gamma)}$ dependence of the other terms in (52). Therefore we again obtain the condition $\gamma > 0$ for the viscosity terms to be negligible in all equations in the limit $\alpha \to 0$.

To summarize, we have shown that viscosity effects are negligible near the singularity provided $p > 0$, in the sense that if we drop them in the first order, then the solutions are at least self-consistent.

Finally, one observes trivially that the inclusion of the cosmological term (Λ = const) in the field equations does not change any of our previous analysis.

4. EXAMPLES OF PERFECT FLUID MODELS

In the following, we list some of the exact solutions of the Einstein equations we were able to find, simply to demonstrate the existence of velocity-dominated singularities. No exhaustive classification is intended.

For homogeneous models, the well-known Tolman[12] and Bianchi I[13] radiation models ($\gamma = \frac{1}{3}$) and their generalizations to other values of γ are trivially velocity-dominated, their singularities being the prototypes of Friedman-like and Heckmann–Schucking-like singularities, respectively. The radiation ($\gamma = \frac{1}{3}$) models of Kastowski[14] have both a Friedman-like ($\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$) and a Kasner-like ($\frac{2}{3}, \frac{2}{3}, -\frac{1}{3}$) singularity. All Ellis–MacCallum models with 3-spaces of constant curvature[15] (i.e., ${}^3R_{ab}$'s are isotropic) are explicitly velocity-dominated. On the other hand, the Collin's special type II axis-symmetric models[16] are *not* velocity-dominated. A structure can be assigned to them, but then $\mathbf{P} = [(1-\gamma)/2(1+\gamma), (3+\gamma)/4(1+\gamma), (3+\gamma)/4(1+\gamma)]$. In general, the asymptotic behaviors of the Bianchi models are more

easily seen using the potential method of Misner[17] and Jacobs–Hughston,[18] since the "potentials" $V(\beta)$ are essentially the 3-curvatures. A solution will be velocity-dominated near the singularity $\Omega \to \infty$ if the initial conditions are such that $V(\beta)$ becomes negligible compared to the "kinetic energy" $(d\beta/d\Omega)$ terms in the limit $\Omega \to \infty$. In this sense, types VIII and IX seem not to be velocity-dominated because of the recurrent dominance of the 3-curvature terms as the system point gets scattered off by the rising potential walls, whereas all other types seem at least to admit subclasses of solutions which are velocity-dominated. At this point, we like to clarify a little the terminology used by other groups[15,18]. The "cigar" singularity corresponds to our case $\mathbf{P} = (P_1, P_2 > 0 > P_3)$. The "pancake" singularity corresponds to our case $\mathbf{P} = (1, 0, 0)$. The "point" singularity corresponds to our Friedman-like case if $\gamma \neq 1$. Finally, a velocity-dominated "barrel" singularity can only occur when $\gamma = 1$.

As far as we know, no exact analytic solutions of inhomogeneous models with perfect fluid ($\gamma \neq 0$) sources have been found except for the special case $p = \rho$.[19] However, in both the plane and spherical symmetric[20] models, one can use ad hoc approximation methods to study the behavior of the metric near the singularity. The metric for these models can be written in the form

$$ds^2 = -e^{2\sigma(x^3,t)}dt^2 + \varphi^2(x^3,t)\,dl^2 + \psi^2(x^3,t)(dx^3)^2,$$
$$u^\mu = e^{-\sigma}\delta^\mu_0, \qquad p, \rho \text{ functions of } x^3, t \text{ only}, \tag{55}$$

where (a) in the spherical case,

$$dl^2 = d\Omega^2, \qquad (x^1, x^2, x^3) = (\theta, \phi, r),$$

(b) in the plane case,

$$dl^2 = dx^2 + dy^2, \qquad (x^1, x^2, x^3) = (x, y, z),$$

and

$$\sigma = \gamma(\ln\psi + 2\ln\phi), \tag{56}$$

follows from the Bianchi identities after fixing the x^3 and t coordinates. Suppose we assume a first-order solution of the form

$$\varphi \simeq {}_0\varphi(x^3)[t - {}_0t(x^3)]^{2/[3(1-\gamma)]}, \tag{57a}$$

$$\psi \simeq {}_0\psi(x^3)[t - {}_0t(x^3)]^{-1/[3(1-\gamma)]}. \tag{57b}$$

Then putting these into the field equations, one checks that all lowest-order terms cancel out identically and the correction terms are of higher order $[{}_1\varphi \sim (t - {}_0t)^{(6\gamma+4)/3(1-\gamma)}; \quad {}_1\psi \sim (t - {}_0t)^{(6\gamma+1)/3(1-\gamma)}]$ as long as $\gamma > -\frac{1}{3}$. Thus one sees that the exact solution will admit singularity of the type $(\frac{2}{3}, \frac{2}{3}, -\frac{1}{3})$. Similarly one checks that a first-order Friedman-like solution

$$\varphi \simeq {}_0\varphi(x^3)(t - {}_0t)^{2/[3(1-\gamma)]}, \tag{58a}$$

$$\psi \simeq {}_0\psi(x^3)(t - {}_0t)^{2/[3(1-\gamma)]} \tag{58b}$$

is also consistent with the field equations provided the conditions ${}_0t_{,3} = 0$ and $\gamma > -\frac{1}{3}$ are satisfied. Moreover, it turns out that the general solution contains the same number of essential arbitrary functions (two of them) as the first approximation, so that the data on the singularity uniquely generate the full metric.

For the special case $p = \rho(\gamma = 1)$, exact solutions with plane symmetry can be obtained and are being studied in detail by Tabensky.[19] At least one class of exact solutions turns out to be velocity-dominated. The form of the metric near the singularity is

$$e^\sigma \simeq K(z)t^{3K(z)/4+1/2K(z)-1},$$
$$\varphi \simeq t^{K(z)/2}; \qquad \psi \simeq K(z)t^{1/2K(z)-K(z)/4}, \tag{59}$$

where $K(z)$ is some arbitrary function of z. In terms of the proper time of a particular comoving observer it can be considered as having the structure

$$\mathbf{P} = \left(\frac{2K^2}{2+3K^2}; \frac{2K^2}{2+3K^2}; \frac{2-K^2}{2+3K^2}\right),$$

where now $\mathbf{P} \equiv (P_1, P_2, P_3)$ are the power dependences of φ and ψ on the *proper time* of the matter lines.

The interesting feature about this class of solution is that one of the arbitrary functions gets wiped out as one approaches the singularity, so that the data on the singularity $(K(z))$ is *not* enough to generate the full metric, in contrast to the $p < \rho$ case above. Another class of special $p = \rho$ solutions has been found, however, which has the singularity structure $\mathbf{P} = (1/c, 1/c, 1)$ (c any constant $\neq -2$), and which does not seem to be velocity-dominated.

5. CONCLUSIONS

We have shown in this paper by explicit demonstration that a unique and elegant structure can be assigned to velocity-dominated singularities of perfect fluid models with a barotropic equation of state $p = \gamma\rho$, whenever such singularities appear in exact solutions of the Einstein equations. Furthermore, we show that it is at least self-consistent to consider the effects due to a time-varying equation of state in which $\gamma \to 0$ near the singularity, and viscosity (for $\gamma > 0$), to be of higher order so that they do not change the structure of the singularity.

A great deal of work remains to be done. The immensely complicated and numerous consistency conditions remain to be disentangled and hopefully interpreted, especially for the (1, 0, 0) and $\gamma = 1$ cases. The Lifshitz and Khalatnikov conjecture,[4] that the (1, 0, 0) singularities are merely caustics of matter flow lines and will be removed by pressure gradients, is perhaps somehow buried in the mess of the consistency conditions. This is of particular interest because of its relation to the problem of horizons.[21] The effect of rotations is another problem that might be of interest and importance. The construction of the b boundary[1] seems also to be possible. Finally, the study of the mixmaster-type singularities, which do not seem to fall into the velocity-dominated category as it is defined now, is of crucial importance.

To a large extent, our approach is very similar to the approach of Lifshitz and Khalatnikov, except maybe a little more systematic and geometrically intepretable. DeWitt has also integrated the velocity-dom-

inated equations for the vacuum case, though in a completely different context. [22] In any case, the number counting tells us that it is still far short of being generic and generalization in some other direction must be searched if one hopes to obtain the most general singularity structure of cosmological solutions of the Einstein equations.

ACKNOWLEDGMENTS

The author is greatly indebted to Professor R. K. Sachs for his guidance, advice, and assistance, among many other things, throughout this work. He is also grateful to D. Eardley, R. Tabensky, and Professor J. Silk for many helpful discussions.

---

* Partially supported by NSF GP-21495, University of California at Berkeley, and NSF Grants GU-1598 and GP-20033, The University of Texas at Austin.

† Present Address: Center for Relativity Theory, Department of Physics, University of Texas at Austin, Austin, Texas 78712.

[1] D. Eardley, E. Liang, and R. Sachs, Velocity-Dominated Singularities in Irrotational Dust Cosmologies. J. Math. Phys. **23**, 99 (1972).

[2] See W. Kundt, "Survey on Cosmology" (unpublished), and references therein.

[3] K. Huang and S. Weinberg, Phys. Rev. Letters **25**, 895 (1970).

[4] E. M. Lifshitz and I. M. Khalatnikov, Advan. Phys. **12**, 185 (1963).

[5] S. W. Hawking, Phys. Rev. Letters **17**, 444 (1966).

[6] See, for example, J. L. Synge, *Relativity, The General Theory* (North-Holland, Amsterdam, 1964).

[7] L. P. Eisenhart, *Riemannian Geometry* (Princeton U.P. Princeton, N.J. 1966).

[8] J. Ehlers, Akad. Wiss Lit. (Mainz), Abhandl. Math-Nat. Kl. (No. 11), (1961).

[9] V. A. Benlinskii and I. M. Khalatnikov, Zh. Eksp. Teor. Fiz. **57**, 2163 (1969) [Sov. Phys. JETP **30**, 1174 (1970)]

[10] C. W. Misner, Astrophys. J. **151**, 431 (1968).

[11] G. F. R. Ellis, "Relativistic Cosmology", in International School of Physics, "Enrico Fermi," XLVII General Relativity and Cosmology, 1969.

[12] See for example L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, Rev. 2nd. ed. (Pergamon, London, 1962), Chap. 12.

[13] K. C. Jacobs, Astrophys. J. **153**, 661 (1968); Bianchi I Cosmological Models, Ph.D. Thesis California Inst. of Tech., 1969 (University Microfilms, 1969.)

[14] R. Kantowski, Ph.D. Thesis, U. of Texas at Austin, 1966.

[15] G. F. R. Ellis and M. A. H. MacCallum, Commun Math. Phys. **12**, 2, 108 (1969); M. A. H. MacCallum, Commun. Math. Phys. **20**, 57 (1971).

[16] C. B. Collins and J. M. Stewart, "Qualitative Cosmology" (unpublished).

[17] C. W. Misner, *Quantum Cosmology I* Phys. Rev. **186**, 5, 1319 (1969); "Classical and Quantum Dynamics of a Closed Universe," *Relativity*, edited by M. Carmeli, S. I. Ficklev, and L. Witten (Plenum, N.Y., 1970).

[18] K. C. Jacobs and L. P. Hughston, "More Quantum Cosmology," (unpublished).

[19] R. Tabensky (private communication) paper to appear on exact $p = \rho$ plane and spherical symmetric solutions.

[20] See, for example, G. C. McVittie, *General Relativity and Cosmology* (U. of Illinois Press, Urbana, 1965).

[21] See, for example, C. W. Misner, Phys. Rev. Letters. **22**, 1071 (1969) or S. W. Hawking, Adams Prize Essay, Cambridge Univ., 1966 (unpublished).

[22] B. S. DeWitt, Phys. Rev. **160**, 1113 (1967). We thank Dr. A. Fischer for informing us of the previous works on velocity-dominated solutions done in this paper.

A Method for Generating New Solutions of Einstein's Equation. II

Robert Geroch*
Department of Physics, University of Texas, Austin, Texas
(Received 20 August 1971)

A scheme is introduced which yields, beginning with any source-free solution of Einstein's equation with two commuting Killing fields for which a certain pair of constants vanish (e.g., the exterior field of a rotating star), a family of new exact solutions. To obtain a new solution, one must specify an arbitrary curve (up to parametrization) in a certain three-dimensional vector space. Thus, a single solution of Einstein's equation generates a family of new solutions involving two arbitrary functions of one variable. These transformations on exact solutions form a non-Abelian group. The extensive algebraic structure thereby induced on such solutions is studied.

1. INTRODUCTION

In Ref. 1, a general scheme was described which yields, beginning with any source-free solution of Einstein's equation with a Killing vector ξ^a, a one-parameter family of (in general, distinct) exact solutions. The method is as follows. Solve

$$\nabla_a \omega = \epsilon_{abcd}\xi^b \nabla^c \xi^d, \tag{1}$$

$$\nabla_{[a}\alpha_{b]} = \tfrac{1}{2}\epsilon_{abcd}\nabla^c \xi^d, \qquad \xi^a \alpha_a = \omega, \tag{2}$$

$$\nabla_{[a}\beta_{b]} = 2\lambda\nabla_a \xi_b + \omega\epsilon_{abcd}\nabla^c \xi^d, \qquad \xi^a\beta_a = \lambda^2 + \omega^2 - 1, \tag{3}$$

for ω, α_a, and β_a, where $\lambda = \xi^a\xi_a$. (Existence of solutions, locally, is guaranteed by Einstein's and Killing's equations.) Then, for each real number θ, the metric

$$\tilde{g}_{ab} = \lambda\tilde{\lambda}^{-1}(g_{ab} - \lambda^{-1}\xi_a\xi_b) + \tilde{\lambda}\eta_a\eta_b \tag{4}$$

is again an exact source-free solution of Einstein's equation,[2] where we have set

$$\tilde{\lambda} = \lambda[(\cos\theta - \omega\sin\theta)^2 + \lambda^2\sin^2\theta]^{-1}, \tag{5}$$

$$\eta_a = \tilde{\lambda}^{-1}\xi_a + 2\alpha_a\cos\theta\sin\theta - \beta_a\sin^2\theta. \tag{6}$$

Now suppose we begin with a solution g_{ab} which admits two Killing fields $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$. Then any linear combination of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ (with constant coefficients) is again a Killing field. We can perform the transformation above with respect to that linear combination. It turns out, furthermore, that the resulting metric $\tilde{g}_{ab}$ again admits $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ as Killing fields, provided (i) $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ commute, and (ii) a certain pair of constants vanish. Finally if (i) and (ii) are satisfied initially, then they continue to be satisfied after the transformation (4).

It is clear from the remarks above that the possibility is now open[3] to generate still larger classes of exact solutions by iterating the transformations (4). To our original metric g_{ab}, apply the transformation with respect to some linear combination of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$. To the resulting metric, apply the transformation with respect to some other linear combination of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$, etc. The crucial question is: Does the result of applying two successive transformations (4) (with different Killing vectors) depend on the order in which they are applied? If not, iterations of the transformation would be unnecessary: the entire class of metrics would be obtained already after the first application of (4). It turns out, however, that the metric which results from two successive applications of (4) *does* depend on the order in which they are applied. Thus, after the first application of (4), we obtain a two-parameter family of solutions. (One parameter represents the θ value of the transformation, the other the linear combination of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ used.[4]) Applying the transformations (4) to each metric in this family, we obtain a four-parameter family. The third iteration yields a six-parameter family, etc. Thus, beginning with just one solution of Einstein's equation (with a pair of commuting Killing fields), we expect to obtain, by successively iterating the transformations (4), a class of exact solutions too large to be characterized by any finite number of parameters. That is, the resulting class of solutions should involve arbitrary functions.

The purpose of this paper is to carry out the program outlined above.

The iteration process can be described more conveniently by using infinitesimal transformations (4) (i.e., $\theta \ll 1$) rather than finite ones. Let $\xi^a(t)$ $(0 \le t \le 1)$ be a one-parameter family of Killing fields on the space-time M, g_{ab}. That is to say, for each value of t, $\xi^a(t)$ is some linear combination of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$. Fix a small interval Δt. Now apply to our space-time successive infinitesimal ($\theta = \Delta t$) transformations (4), first with respect to the Killing field $(\Delta t)^{-1}[\xi^a(\Delta t) - \xi^a(0)]$, then with respect to $(\Delta t)^{-1}[\xi^a(2\Delta t) - \xi^a(\Delta t)]$, etc., ending with the transformation with respect to the Killing field $(\Delta t)^{-1}[\xi^a(1) - \xi^a(1 - \Delta t)]$. The result is some exact solution of Einstein's equation. Taking the limit as $\Delta t \to 0$, we obtain, for each curve $\xi^a(t)$ in the space of Killing vectors, a solution of Einstein's equation. (In this language, iterating finite transformations amounts to considering only those curves which consist of broken straight-line segments.) The noncommutativity of successive transformations implies that the final metric defined by a curve $\xi^a(t)$ depends on the detailed behavior of the curve, not just on its endpoints. We thus expect to obtain, from one solution of Einstein's equation, a class of solutions which involves one arbitrary funtion of one variable. [$\xi^a(t)$ is a curve in the two-dimensional space of Killing fields. But parametrization is irrrelevant.]

It turns out, in fact, that the presence of two commuting Killing fields gives rise to transformations more general than (4). As a consequence, one obtains, beginning with a single solution of Einstein's equation, a class of new exact solutions which depend on two arbitrary functions of one variable.

Does one expect these new solutions to have any physical significance? Suppose we begin with a Weyl solution. It has a pair of commuting Killing fields

(static and axially symmetric), and the two constants referred to earlier vanish. Applying all possible transformations above to this solution, we expect to obtain a class of exact solutions involving two arbitrary functions of one variable. Each of these solutions will be stationary and axially symmetric. But the most general stationary, axially symmetric, source-free solution of Einstein's equation involves just two arbitrary functions of one variable (physically, the "mass distribution and angular momentum distribution along the axis"). Thus, at least by this crude argument, we expect a single Weyl solution to yield the general exterior metric of a rotating star! In other words, it is reasonable to suppose that any two exact solutions with a pair of commuting Killing fields (and whose two constants vanish) will, at least locally, be related by one of the transformations described here. (I know of no proof of this conjecture, however.)

Before one can determine a new solution using (4), it is first necessary to integrate certain differential equations, namely, (1), (2), and (3). Now suppose we have performed one transformation, and wish to carry out a second. One must first carry out integrals analogous to (1), (2), and (3), but now referring to the metric $\tilde{g}_{ab}$, which resulted from the first transformation. Similarly, the third iteration requires the values of integrals referring to the metric obtained after the second iteration, etc. Ultimately, we wish to express the final metric in terms of quantities which refer only to the initial metric g_{ab}, and not to the intermediate ones. Only in this way can we retain any hope of describing the final class of new metrics in any simple way. Therefore, one must reexpress each of the intermediate integrals in terms of the original metric. We wish to write down, at the outset, the results of all the integrations which will ever have to be performed in iterating the transformations (4).

It is instructive to attempt to proceed directly. To the metric g_{ab}, apply the transformation (4) with respect to the Killing vector $\overset{0}{\xi}{}^a$. Now write down the expressions (1), (2), and (3) in preparation for the application of a second transformation with respect to say, $\overset{1}{\xi}{}^a$. Express the (curl-free) right sides in terms of the original metric, and integrate. After some effort, one obtains new fields, which are defined by g_{ab}, but which one would have been unlikely to discover *a priori*. Now attempt a third transformation, with respect to $\overset{0}{\xi}{}^a$, and again express the required integrals in terms of g_{ab}. The effort is considerable this time, but one eventually discovers still other new fields. The various fields do not appear to fall into any simple pattern, and so it is difficult, proceeding in this way, to write down all the fields (or even to tell if their number is finite).

The solution of this problem—exhausting the list of fields defined by a single exact solution of Einstein's equation with a pair of commuting Killing fields—requires a digression into the properties of such solutions. This is done in Sec. 2. It is convenient to adopt a two-dimensional formalism, in which the action of the isometries is "divided out." Einstein's equation is then expressed as a set of differential equations involving certain scalar and tensor fields on a 2-manifold. (These equations are remarkably simple.) We then define the collection of new fields, which, in fact, form an infinite sequence. It turns out to be convenient to divide the fields into pairs to obtain complex fields. The equations can then also be combined into complex equations. These fields are potentially important in any discussion of stationary axially symmetric solutions.

The transformations themselves are discussed in Sec. 3. Once all the fields have been defined, in Sec. 2, things become relatively simple. We obtain a set of differential equations giving the change in the fields (as a function of the curve parameter t) in terms of their values at t. The effect of the transformations is thus merely to shuffle the fields among themselves. The fact that no finite number of fields will suffice is reflected in the fact that the change in the nth field is expressed in terms of the values of the first $n + 1$ fields. This feature also makes the equations difficult to solve.

Unfortunately, it has not been possible to solve the equations in Sec. 3 in closed form, and thus to write down explicitly the complete class of new solutions of Einstein's equation. This deficiency is perhaps not so serious as it may appear, however. In a large number of special cases, one can obtain new solutions in closed form, and, even when no explicit form exists, one can still establish certain properties of the resulting solutions. What is it in the structure of Einstein's equation that causes its solutions to be subject to the transformations defined here? These topics are discussed in Sec. 4.

2. SUPPLEMENTARY FIELDS

Let M be a four-dimensional manifold, and g_{ab} a metric on M of signature $(-,+,+,+)$ which is a solution of Einstein's (source-free) equation $R_{ab} = 0$. Let $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ be a pair of Killing fields on M which commute:

$$\overset{0}{\xi}{}^b\nabla_b\overset{1}{\xi}{}^a - \overset{1}{\xi}{}^b\nabla_b\overset{0}{\xi}{}^a = 0. \tag{7}$$

As described in Sec. 1, the result of successive application of the transformations (4) to such a solution involves an infinite collection of additional fields, each of which can be expressed in terms of g_{ab}, $\overset{0}{\xi}{}^a$, and $\overset{1}{\xi}{}^a$. The purpose of this section is to define these fields. The discussion is simplified considerably by passing to a two-dimensional formalism, in which the action of the isometries is divided out. Our starting point, then, is the two-dimensional formulas giving Einstein's equation for a space-time with a pair of commuting Killing fields. (The formalism is described in detail, and the equations derived, in Appendix A.) Using an inductive argument, we then define the sequence of new fields. These fields, whose mere existence requires Einstein's equation, have a number of interesting properties. A few of these are discussed here.

Let S denote the 2-manifold whose points are the orbits, in M, under the isometry group generated by $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$. We denote by h_{ab} the induced metric on S (A5), and by D_a the (covariant) derivative on S with respect to this metric (A7). We next introduce five

scalar fields on S. Three are given by the inner products of the Killing fields:

$$\lambda_{00} = \overset{0}{\xi}{}^a \overset{0}{\xi}_a, \quad \lambda_{01} = \lambda_{10} = \overset{0}{\xi}{}^a \overset{1}{\xi}_a, \quad \lambda_{11} = \overset{1}{\xi}{}^a \overset{1}{\xi}_a, \tag{8}$$

and the other two by the formulas

$$c_0 = \epsilon_{abcd} \overset{0}{\xi}{}^a \overset{1}{\xi}{}^b \nabla^c \overset{0}{\xi}{}^d, \quad c_1 = \epsilon_{abcd} \overset{0}{\xi}{}^a \overset{1}{\xi}{}^b \nabla^c \overset{1}{\xi}{}^d. \tag{9}$$

It follows from Einstein's and Killing's equations and (7) that c_0 and c_1 are constants. Expressed in terms of $h_{ab}, \lambda_{00}, \lambda_{01}, \lambda_{11}, c_0$, and c_1, Einstein's equation takes the form (see Appendix A)

$$\begin{aligned} D^a[\tau^{-1} D_a \lambda_{00}] &= 2\tau^{-3}\lambda_{00}[(D^a\lambda_{00})(D_a\lambda_{11}) \\ &\quad - (D^a\lambda_{01})(D_a\lambda_{01})] + 2\tau^{-3}(c_0)^2, \\ D^a[\tau^{-1} D_a \lambda_{01}] &= 2\tau^{-3}\lambda_{01}[(D^a\lambda_{00})(D_a\lambda_{11}) \\ &\quad - (D^a\lambda_{01})(D_a\lambda_{01})] + 2\tau^{-3}c_0 c_1, \\ D^a[\tau^{-1} D_a \lambda_{11}] &= 2\tau^{-3}\lambda_{11}[(D^a\lambda_{00})(D_a\lambda_{11}) \\ &\quad - (D^a\lambda_{01})(D_a\lambda_{01})] + 2\tau^{-3}(c_1)^2, \end{aligned} \tag{10}$$

$$\begin{aligned} \mathcal{R} &= \tau^{-2}[(D^a\lambda_{00})(D_a\lambda_{11}) - (D^a\lambda_{01})(D_a\lambda_{01})] \\ &\quad + 6\tau^{-4}[2c_0c_1\lambda_{01} - (c_0)^2\lambda_{11} - (c_1)^2\lambda_{00}], \end{aligned} \tag{11}$$

where $\mathcal{R}$ is the scalar curvature[5] of S, and where we have defined

$$\tau^2 = 2[(\lambda_{01})^2 - \lambda_{00}\lambda_{11}]. \tag{12}$$

These equations can be written more concisely using an index notation. We introduce upper case Latin indices with range 0, 1. The three fields $\lambda_{00}, \lambda_{01}$, and λ_{11} then define a symmetric tensor λ_{AB}, while the constants c_0 and c_1 define a vector c_A. Mathematically, a quantity with upper case Latin indices represents a multilinear mapping (linear if one index, bilinear if two, etc.) from the two-dimensional vector space of Killing vectors to the vector space of scalar fields on S. These Latin indices will be raised and lowered with an antisymmetric ϵ_{AB} (i.e., $\epsilon_{00} = \epsilon_{11} = 0, \epsilon_{01} = -\epsilon_{10} = 1$), using the same rule as for ordinary spinors[6]:

$$\epsilon^{AC}\epsilon_{BC} = \delta^A_B, \quad p^A = \epsilon^{AB}p_B, \quad p_B = p^C\epsilon_{CB} \tag{13}$$

In this notation, Eqs. (10) and (11) take the form

$$\begin{aligned} D^a[\tau^{-1} D_a \lambda_{AB}] &= \tau^{-3}\lambda_{AB}[(D_a\lambda_{MN})(D^a\lambda^{MN})] \\ &\quad + 2\tau^{-3}c_A c_B, \end{aligned} \tag{14}$$

$$\mathcal{R} = \tfrac{1}{2}\tau^{-2}[(D_a\lambda_{MN})(D^a\lambda^{MN})] - 6\tau^{-4}\lambda_{AB}c^A c^B. \tag{15}$$

We must now make an additional assumption, namely, that the constant c_A vanishes. This assumption is required because of the following result: a transformation (4) on a solution of Einstein's equation destroys the presence of two Killing vectors unless $c_A = 0$. (See Appendix B.) Thus, it is only under this condition that we retain the possibility of iterating the transformations. (This assumption does not appear to be very restrictive. In fact, I am aware of no solution of Einstein's equation for which $c_A \neq 0$.) We impose $c_A = 0$ at this point because only then do the fields we wish to define exist.

Note that now the upper case Latin indices appear in (14) and (15) only in symmetric pairs. It is convenient, therefore, to replace such pairs by Greek indices. Thus, for example, the field λ_{AB} will be written λ_α. The "α" merely stands for the pair of indices "AB" that it replaces. More precisely, objects with Greek indices represent multilinear mappings from a three-dimensional vector space (symmetric, second-rank tensors over the vector space of Killing vectors) to the vector space of scalar fields on S. The ϵ_{AB} defines a symmetric metric:

$$G_{\alpha\beta} = -\epsilon_{A(C}\epsilon_{D)B}, \quad \alpha \to AB, \quad \beta \to CD, \tag{16}$$

with signature $(+,-,-)$. We shall raise and lower Greek indices with this metric, an operation equivalent, by (16), to raising and lowering the corresponding Latin indices with ϵ_{AB}.

We have now reduced our basic fields to h_{ab} and λ_α, and Einstein's equation to

$$D^a[\tau^{-1} D_a \lambda_\alpha] = \tau^{-3}\lambda_\alpha\,[(D_a\lambda_\mu)(D^a\lambda^\mu)], \tag{17}$$

$$\mathcal{R} = \tfrac{1}{2}\tau^{-2}\,[(D_a\lambda_\mu)(D^a\lambda^\mu)], \tag{18}$$

where, from (12),

$$\tau^2 = -\lambda^\alpha\lambda_\alpha. \tag{19}$$

These are the equations on which the remainder of this section will be based.

We introduce two further bits of notation. Define the totally antisymmetric field

$$\epsilon_{\alpha\beta\gamma} = -2^{-1/2}[\epsilon_{A(C}\epsilon_{D)(E}\epsilon_{F)B} + \epsilon_{B(C}\epsilon_{D)(E}\epsilon_{F)A}], \quad \alpha \to AB, \ \beta \to CD, \gamma \to EF \tag{20}$$

so $\epsilon^{\alpha\beta\gamma}\epsilon_{\alpha\beta\gamma} = 6$. We denote by ϵ_{ab} the alternating tensor field on S[(A6)]. Finally, we shall use a star, applied to a vector field on S, to denote the dual, e.g.,

$$^*V_a = \epsilon_a{}^b V_b. \tag{21}$$

Thus, $^{**} = -1$.

A solution of (17) and (18) leads to an infinite sequence of further fields, each of which has a single Greek index. [In fact, one only uses (17). Note that this equation is conformally invariant.] We now define these. Contracting (17) with λ^α, and using (19), we have

$$D^a D_a \tau = 0. \tag{22}$$

That is to say, τ is an analytic function on S. Its conjugate function, defined by

$$D_a\sigma = {}^*D_a\tau, \tag{23}$$

is therefore also analytic. Next note that (17) implies that the right side of

$$D_a\omega_\alpha = \tau^{-1}\epsilon_{\alpha\mu\nu}\lambda^\mu\,{}^*D_a\lambda^\nu \tag{24}$$

is curl-free. Hence, (24) defines a new field, ω_α. [This ω_α is, in fact, precisely the twists defined by (A8).] We introduce the complex linear combination

$$\overset{0}{A}_\alpha = \omega_\alpha + i\lambda_\alpha. \tag{25}$$

It follows immediately from (17) and (24) that $\overset{0}{A}_\alpha$ satisfies

$$D^a D_a \overset{0}{A}_\alpha = -\tfrac{1}{2}\tau^{-1}\epsilon_{\alpha\mu\nu}(D^a \overset{0}{A}^\mu)(*D_a \overset{0}{A}^\nu). \tag{26}$$

This $\overset{0}{A}_\alpha$ is the first of our sequence of fields. It follows from (26) and (23) that the right side of

$$D_a \overset{1}{A}_\alpha = -\epsilon_{\alpha\mu\nu}\overset{0}{A}^\mu D_a \overset{0}{A}^\nu + 2(\sigma + \tau^*) D_a \overset{0}{A}_\alpha \tag{27}$$

is curl-free, and hence that (27) defines a field $\overset{1}{A}_\alpha$. But now the right side of

$$D_a \overset{2}{A}_\alpha = -\tfrac{1}{2}\epsilon_{\alpha\mu\nu}[\overset{0}{A}^\mu D_a \overset{1}{A}^\nu + \overset{1}{A}^\mu D_a \overset{0}{A}^\nu] + (\sigma + \tau^*) D_a \overset{1}{A}_\alpha + (\sigma + \tau^*)^2 D_a \overset{0}{A}_\alpha \tag{28}$$

is curl-free, defining $\overset{2}{A}_\alpha$. [Polynomials in * are defined by expanding and setting $** = -1$. For example, $(\sigma + \tau^*)^2 = \sigma^2 - \tau^2 + 2\sigma\tau^*$.] Proceeding inductively, we define $\overset{n}{A}_\alpha$ in terms of $\overset{n-1}{A}_\alpha, \overset{n-2}{A}_\alpha, \ldots, \overset{0}{A}_\alpha$ by

$$D_a \overset{n}{A}_\alpha = -(1/n)\epsilon_{\alpha\mu\nu}(A^\mu D_a \overset{n-1}{A}^\nu + \cdots + \overset{n-1}{A}^\mu D_a \overset{0}{A}^\nu) + (2/n)[(\sigma + \tau^*) D_\alpha \overset{n-1}{A}_\alpha + \cdots + (\sigma + \tau^*)^n D_a \overset{0}{A}_\alpha]. \tag{29}$$

The existence of a solution (i.e., the vanishing of the curl of the right side) follows from this same equation for lower values of n.

To summarize, a solution of Einstein's equation with two commuting Killing vectors (and with $c_A = 0$) is characterized by two fields, h_{ab} and λ_α, on a 2-manifold S, subject to (17) and (18). Such a solution then defines, via (25) and (29), an infinite sequence of complex fields, $\overset{0}{A}_\alpha, \overset{1}{A}_\alpha, \ldots$.

Are the A's all algebraically independent? The answer is no. In fact, the imaginary part of $\overset{n}{A}_\alpha$ can be expressed in terms of $\overset{n-1}{A}_\alpha, \overset{n-2}{A}_\alpha, \ldots, \overset{0}{A}_\alpha$ by the equation

$$\operatorname{Im}\overset{n}{A}_a = -(1/n)\epsilon_{\alpha\mu\nu}(\operatorname{Re}\overset{0}{A}^\mu \operatorname{Im}\overset{n-1}{A}^\nu + \cdots + \operatorname{Re}\overset{n-1}{A}^\mu \operatorname{Im}\overset{0}{A}^\nu) + (2/n)[\operatorname{Re}(\sigma + i\tau)\operatorname{Im}\overset{n-1}{A}_\alpha + \cdots + \operatorname{Re}(\sigma + i\tau)^n \operatorname{Im}\overset{0}{A}_\alpha] \tag{30}$$

To prove this equation, one observes, from (24) and (27), that it is valid for $n = 1$, and, from (29), that its validity for n implies its validity for $n + 1$. Although one could certainly introduce only the independent variables (λ_α and the real parts of the A's), and write all expressions in terms of these, the result is to complicate rather than simplify equations. Strangely enough, Eq. (30) seems never to play a significant role. It is used at only one point in Sec. 3: to get an induction argument started.

Finally, note that (23) defines σ, and (29) each $\overset{n}{A}_\alpha$, only up to an additive constant. This feature gives rise to an (infinite-dimensional) group of gauge transformations on σ and the A's. Gauge invariance provides a powerful check on formulas. In fact, an expression involving τ, σ, and the A's is essentially uniquely determined by its behavior under all gauge transformations. (It was through this fact that the inductive formulas in this and Sec. 3 were obtained.) The gauge transformations, along with some parity and dimension transformations, are discussed in Appendix C.

3. GENERATING NEW SOLUTIONS

Roughly speaking, the action of the transformations (4) is to shuffle the $\overset{n}{A}_\alpha$ among themselves. This "shuffling" is such that if one iterates only a finite number of transformations, then only a finite number of A's are involved in the final solution. However, as more and more transformations are applied in succession, the number of A's required increases without bound. In this section we shall derive the equations governing the behavior of these fields on S under the transformations (4), first in the infinitesimal and then in the finite case.

Let the metric g_{ab} of Sec. 1 admit commuting Killing fields $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$, and let $\overset{0}{\xi}{}^a$ be the Killing vector denoted ξ^a in Eqs. (1)–(6). Then the transformed metric $\tilde{g}_{ab}$ of (4) also admits $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ as Killing fields provided

$$\pounds_{\overset{1}{\xi}}\omega = 0, \quad \pounds_{\overset{1}{\xi}}\alpha_a = 0, \quad \pounds_{\overset{1}{\xi}}\beta_a = 0, \tag{31}$$

where $\pounds$ denotes the Lie derivative. These equations can be satisfied if (and, in general, only if) $c_A = 0$. (See Appendix B.) Thus, in this case our new solution—and all solutions obtained by further iterations of the transformations—have $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ as commuting Killing fields. Note that, while the transformations alter the underlying metric on M, they leave invariant the vector fields $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ on M.[7]

The general transformation can be obtained by iterating infinitesimal ones, and so we are led to consider the infinitesimal version (i.e., $\theta \ll 1$) of (4). To first order in θ, Eq. (4) takes the form

$$\tilde{g}_{ab} = g_{ab} - 2\theta\omega_{00} g_{ab} + 4\theta \overset{0}{\xi}_{(a}\alpha_{b)}. \tag{32}$$

Equation (32) defines an infinitesimal change in our solution of Einstein's equation, and hence some infinitesimal change in the variables on S—$h_{ab}, \lambda_{00}, \lambda_{01}$ and λ_{11}—which define that solution. Using the identity

$$2\overset{1}{\xi}{}^b\nabla_{[a}\alpha_{b]} + \pounds_{\overset{1}{\xi}}\alpha_a - \nabla_a(\overset{1}{\xi}{}^b\alpha_b) = 0 \tag{33}$$

and Eqs. (2) and (31), we have

$$\overset{1}{\xi}{}^b\alpha_b = -2^{-1/2}\sigma + \omega_{01}. \tag{34}$$

Hence, when the four-dimensional metric undergoes (32), the variables on S undergo

$$\begin{aligned}
&\lambda_{00} \to \overset{0}{\xi}{}^a\overset{0}{\xi}{}^b\tilde{g}_{ab} = \lambda_{00} + 2\theta\omega_{00}\lambda_{00},\\
&\lambda_{01} \to \overset{0}{\xi}{}^a\overset{1}{\xi}{}^b\tilde{g}_{ab} = \lambda_{01} + 2\theta\omega_{01}\lambda_{00} - 2^{1/2}\theta\sigma\lambda_{00},\\
&\lambda_{11} \to \overset{1}{\xi}{}^a\overset{1}{\xi}{}^b\tilde{g}_{ab} = \lambda_{11} - 2\theta\omega_{00}\lambda_{11} + 4\theta\omega_{01}\lambda_{01} - 2^{3/2}\theta\sigma\lambda_{00},\\
&h_{ab} \to h_{ab} - 2\theta\omega_{00}h_{ab}.
\end{aligned} \tag{35}$$

This equation expresses, in two-dimensional language, the action of an infinitesimal transformation (4).

Eq. (35) can be written more concisely using the index notation. Define an operator $\mathcal{T}_\beta$ on λ_α and h_{ab} by

$$\mathcal{T}_\beta \lambda_\alpha = \lambda_\beta \omega_\alpha - G_{\alpha\beta}(\omega^\mu \lambda_\mu) + \sigma \epsilon_{\beta\alpha\mu}\lambda^\mu, \tag{36}$$

$$\mathcal{T}_\beta h_{ab} = -\omega_\beta h_{ab}. \tag{37}$$

Then (35) can be written

$$\begin{aligned} \lambda_\alpha &\to \lambda_\alpha + 2\theta K^\beta \mathcal{T}_\beta \lambda_\alpha, \\ h_{ab} &\to h_{ab} + 2\theta K^\beta \mathcal{T}_\beta h_{ab}, \end{aligned} \tag{38}$$

where K^β has components $K^{00} = 1, K^{01} = K^{10} = K^{11} = 0$, so K^β is real, constant, and null. Infinitesimal transformations (4) defined by other Killing vectors (other linear combinations of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$) are obtained by letting K^β in (38) be an arbitrary real, constant, null vector.

These results can be checked directly. Applying the operator $\mathcal{T}_\beta$ to both sides of (17) and (18), using (36) and (37), we obtain an identity in each case. [Note that (17) is conformally invariant, while (37) is an infinitesimal conformal transformation on S.] This remark applies, of course, whether $\mathcal{T}_\beta$ is contracted with a vector K^β or not. Therefore, we may drop the condition that K^β be null. Equation (38) defines an infinitesimal transformation on exact solutions for an arbitrary real, constant K^β. This is an important point: The presence of two commuting Killing vectors gives rise to new infinitesimal transformations (i.e., those for which K^β is not null) which are not of the form (4) for *any* linear combination of the Killing fields. Whereas one might have expected at first that the class of infinitesimal transformations with two Killing vectors would be two dimensional, in fact it is three dimensional. This feature comes about essentially because the operator $\mathcal{T}_\beta$ has a single Greek index, i.e., because an infinitesimal transformation (32) is "quadratic" in the Killing vector.

Note that, from (19) and (36),

$$\mathcal{T}_\beta \tau^2 = -2\lambda^\alpha \mathcal{T}_\beta \lambda_\alpha = 0. \tag{39}$$

That is, τ [and hence, by (23), also σ] is invariant under all the transformations.

We can now describe the iteration process in more detail. Let V denote the three-dimensional vector space of the α's, and let $\gamma(t)$ be a curve in V, parametrized by the real variable $t (0 \le t \le 1)$. Let $\gamma(0)$ be the origin of V, and let $K^\beta(t)$ be the tangent vector to this curve. By iterating the sequence of infinitesimal transformations (38) defined by this curve, we obtain a one-parameter family, $\lambda_\alpha(t), h_{ab}(t)$, of solutions of (17) and (18), and hence a one-parameter family of exact solutions of Einstein's equation. That is to say, this family of solutions must satisfy

$$\frac{d}{dt}\lambda_\alpha(t) = K^\beta(t)\mathcal{T}_\beta \lambda_\alpha(t), \tag{40}$$

$$\frac{d}{dt} h_{ab}(t) = K^\beta(t)\mathcal{T}_\beta h_{ab}(t), \tag{41}$$

where the right sides are given by (36) and (37). The problem is to solve these differential equations given initial conditions: the values of $\lambda_\alpha(t)$ and $h_{ab}(t)$ at $t = 0$.

That (40) and (41) are not easy to solve arises from the fact that the right sides of these equations contain not only $\lambda_\alpha(t)$ and $h_{ab}(t)$, but also $\omega_\alpha(t)$, and this quantity is defined only implicity in terms of $\lambda_\alpha(t)$ by (24). Thus, it is necessary to determine how $\omega_\alpha(t)$ varies along the curve $\gamma(t)$. Applying $\mathcal{T}_\beta$ to (24), and using (36), we obtain

$$\begin{aligned} \mathcal{T}_\beta \omega_\alpha = {} & \tfrac{1}{2}(\omega_\alpha \omega_\beta - \lambda_\alpha \lambda_\beta) \\ & - \tfrac{1}{2} G_{\alpha\beta}(\omega^\mu \omega_\mu - \lambda^\mu \lambda_\mu) - \tfrac{1}{2}\epsilon_{\beta\mu\alpha}\ \mathrm{Re}\overset{1}{A}{}^\mu. \end{aligned} \tag{42}$$

That is, the behavior of $\omega_\alpha(t)$ along the curve depends on that of $\overset{1}{A}_\alpha t)$. Equations (36) and (42) can be written neatly in the complex notation [see (25) and (30)]:

$$\mathcal{T}_\beta \overset{0}{A}_\alpha = \tfrac{1}{2}\overset{0}{A}_\alpha \overset{0}{A}_\beta - \tfrac{1}{2} G_{\alpha\beta}(\overset{0}{A}{}^\mu \overset{0}{A}_\mu) - \tfrac{1}{2}\epsilon_{\beta\mu\alpha}\overset{1}{A}{}^\mu. \tag{43}$$

To determine the behavior of $\overset{1}{A}_\alpha(t)$ along the curve, we apply $\mathcal{T}_\beta$ to (27), using (43), to obtain

$$\begin{aligned} \mathcal{T}_\beta \overset{1}{A}_\alpha = {} & \tfrac{1}{6}\overset{0}{A}_\beta \overset{1}{A}_\alpha + \tfrac{1}{3}\overset{1}{A}_\beta \overset{0}{A}_\alpha - \tfrac{1}{2} G_{\alpha\beta}(\overset{0}{A}{}^\mu \overset{1}{A}_\mu) \\ & - \tfrac{2}{3}\epsilon_{\beta\mu\alpha}\overset{2}{A}{}^\mu - \tfrac{1}{6}\epsilon_{\beta\mu\alpha}\overset{0}{A}{}^\mu(\overset{0}{A}{}^\nu \overset{0}{A}_\nu). \end{aligned} \tag{44}$$

The behavior of $\overset{1}{A}_\alpha(t)$ depends on that $\overset{2}{A}_\alpha(t)$. We proceed by induction. Consider the equation

$$\begin{aligned} \mathcal{T}_\beta \overset{n}{A}{}^\alpha = {} & \sum (a_1)^{-1}(a_1 + a_2)^{-1}\cdots(n+2)^{-1} \\ & \times (\epsilon_{\beta\mu}{}^\nu \overset{a_1-1}{A}{}^\mu)(\epsilon_{\nu\rho}{}^k \overset{a_2-1}{A}{}^\rho)\cdots(\epsilon_{\tau\lambda}{}^\alpha \overset{a_m-1}{A}{}^\lambda) \\ & - \epsilon_{\beta\mu}{}^\alpha \overset{n+1}{A}{}^\mu \end{aligned} \tag{45}$$

where the sum extends over all ordered collections, $(a_1, a_2, \cdots, a_m)$ of positive integers satisfying $a_1 + a_2 + \ldots + a_m = n + 2$. (Hence, $1 \le m \le n + 2$.) Equation (45) reduces to (43) when $n = 0$, and to (44) when $n = 1$. Furthermore, applying $\mathcal{T}_\beta$ to (29), we see that the validity of (45) for $n < s$ implies its validity for $n = s$. Hence, (45) determines the behavior of all the $\overset{n}{A}_\alpha$ along the curve $\gamma(t)$.

To summarize, a solution of Einstein's equation, g_{ab}, with a pair of commuting Killing vectors and $c_A = 0$, defines a sequence, $\overset{0}{A}_\alpha, \overset{1}{A}_\alpha, \ldots$, of fields on S. To obtain a new solution, we must first specify any curve, $\gamma(t) (0 \le t \le 1)$, in the three-dimensional vector space V, with $\gamma(0)$ at the origin. Let $K^\beta(t)$ be its tangent vector. We must solve[8] the equations

$$\frac{d}{dt}\overset{n}{A}_\alpha(t) = K^\beta(t)\mathcal{T}_\beta \overset{n}{A}_\alpha(t), \tag{46}$$

$$\frac{d}{dt} h_{ab}(t) = K^\beta(t)\mathcal{T}_\beta h_{ab}(t) \tag{47}$$

for $\overset{n}{A}_\alpha(t)$ and $h_{ab}(t)$, subject to initial conditions: At $t = 0$, $\overset{n}{A}_\alpha(t)$ and $h_{ab}(t)$ reduce to the quantities determined by our original metric g_{ab}. The right sides of (46) and (47) are evaluated using (45) and (37). The values of these fields at $t = 1$ then define a new exact solution of Einstein's equation, again with a pair of commuting Killing vectors. Thus, beginning with a curve in V and an exact solution, we obtain a new

exact solution. (The resulting solution is, of course, independent of the parametrization of $\gamma(t)$.) Equations (46) and (47) are not easy to solve because, by (45), the right side of the nth equation (46) involves the A's up to $\overset{n+1}{A}_\alpha$. Thus, it is not possible to solve the equations one at a time: they must all be solved simultaneously.[9] (This feature, of course, reflects the fact that the entire sequence of A fields is required to discuss the transformations.)

Finally, we verify an assertion made in Sec. 1, namely, that the result of integrating (46) and (47) in general depends on the details of the curve $\gamma(t)$, not just on the endpoint of that curve. Clearly, nontrivial dependence on the curve itself reduces to the assertion that the operator $\mathcal{T}_{[\alpha}\mathcal{T}_{\beta]}$ is nonzero. But, for example, from (45), we have

$$\begin{aligned}\mathcal{T}_{[\alpha}\mathcal{T}_{\beta]}\overset{0}{A}_\gamma &= \tfrac{1}{6}\overset{0}{A}_{[\alpha}G_{\beta]\gamma}(\overset{0}{A}^\mu\overset{0}{A}_\mu) + \tfrac{1}{4}\overset{0}{A}_\gamma\epsilon_{\alpha\beta\mu}\overset{1}{A}^\mu \\ &\quad - \tfrac{1}{6}\overset{0}{A}_{[\alpha}\epsilon_{\beta]\gamma\mu}\overset{1}{A}^\mu + \tfrac{1}{6}\overset{1}{A}_{[\alpha}\epsilon_{\beta]\gamma\mu}\overset{0}{A}^\mu - \tfrac{1}{4}\epsilon_{\alpha\beta\gamma}(\overset{0}{A}^\mu\overset{1}{A}_\mu) \\ &\quad + \tfrac{1}{2}G_{\gamma[\alpha}\epsilon_{\beta]\mu\nu}\overset{0}{A}^\mu\overset{1}{A}^\nu - \tfrac{1}{3}\overset{2}{A}_{[\alpha}G_{\beta]\gamma},\end{aligned} \tag{48}$$

and the right side is nether zero nor a gauge transformation (Appendix C). Thus, we expect the general solution of Einstein's equation, obtained from a single solution, to depend on two arbitrary functions of one variable. (A solution is defined by a curve, up to parametrization, in the three-dimensional vector space V.)

4. CONCLUSION

The general solution of Einstein's equation in the static, axially symmetric case [i.e., of Eqs. (17) and (18)] is not known. But neither is the general solution of (46) and (47) known. Do the transformations described here contribute, then, to the goal of obtaining new exact solutions of Einstein's equation? In fact, it is considerably easier to obtain specific solutions of (46) and (47) than of (17) and (18). The reason is that, to fix a solution of the elliptic partial differential equations (17) and (18), one must specify the boundary conditions. It is difficult in practice to select special cases (i.e., simple boundary conditions) for which these equations can be solved. Equations (46) and (47), on the other hand, involve an arbitrary curve in V. If one chooses an explicit, and not too complicated, form for the curve $\gamma(t)$ (e.g., any broken, null straight line), the equations can be integrated explicitly. One obtains in this way many new, exact solutions of Einstein's equation in closed form—solutions one would never have guessed from (17) and (18) (much less from $R_{ab} = 0$).

Furthermore, many of the questions one might wish to ask of an exact solution can be answered without an explicit form. (After all, most solutions, just as most functions, don't have any "explicit form.") Is a solution asymptotically flat? Static? What are the multipole moments?[10] What is the nature of the singularities? It turns out that the transformations defined here are well-suited for discussing questions of this type. The reason, again, is that the solutions defined by (46) and (47) from a given solution of Einstein's equation are labeled in a particularly simple way—by curves in V—while solutions of (17) and (18) are labeled by boundary conditions. One could ask, for example, for the necessary and sufficient conditions that a certain property (e.g., staticity) be preserved by $\gamma(t)$—a question having no analog in (17) and (18).

Of course, it is not only formulas for and descriptions of exact solutions which are of interest in general relativity. In particular, one would like to understand more deeply the structure of Einstein's equation, or, what is perhaps the same thing, of the set of solutions of that equation. What structure arises from the transformations obtained here? Let V be a real, three-dimensional vector space, and consider the collection of all piecewise smooth[11] curves $\gamma(t)$ $(0 \le t \le 1)$ in V for which $\gamma(0)$ is the origin of V. We regard two such curves as equivalent if they differ by a reparametrization or a retracing.[12] (Equivalent curves clearly define the same transformation on solutions.) On the collection $\mathcal{G}$ of equivalence classes we define a composition law. If $\gamma'(t)$ and $\gamma''(t)$ represent elements of $\mathcal{G}$, their product is the element of $\mathcal{G}$ represented by

$$\gamma'\gamma''(t) = \begin{cases}\gamma'(2t), & 0 \le t \le \tfrac{1}{2} \\ \gamma'(1) + \gamma''(2t-1), & \tfrac{1}{2} \le t \le 1.\end{cases} \tag{49}$$

Thus, $\mathcal{G}$ becomes a (non-Abelian, infinite-dimensional) group. [The identity is the constant curve which remains at the origin. The inverse of $\gamma(t)$ is $-\gamma(t)$.] This $\mathcal{G}$, the group of "effective transformations," acts on the collection of all exact solutions of Einstein's equation with a pair of commuting Killing vectors for which the constants (9) vanish. The crude argument (counting functions) of Sec. 1 suggests that this action is simply transitive[13] (i.e., that any two solutions are related by exactly one element of $\mathcal{G}$). Thus, these solutions of Einstein's equation are acted upon, perhaps simply transitively, by some group $\mathcal{G}$. What is it about Einstein's equation that should cause its solutions to carry so rich a structure?

APPENDIX A: EINSTEIN'S EQUATION WITH TWO COMMUTING KILLING VECTORS

Let M be a four-dimensional manifold, and g_{ab} a metric on M of signature $(-,+,+,+)$ which satisfies Einstein's equation, $R_{ab} = 0$. Let $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ be a pair of Killing vectors on M which commute with each other (7). Because of the presence of a two-parameter group of motions, the metric essentially depends on only two independent variables. One is led, therefore, to introduce a 2-manifold S such that position in S is the "independent variable." The metric and Killing vectors can then be expressed in terms of certain fields on S. The purposes of this Appendix are to introduce a formalism based on this idea, and to rewrite Einstein's equation in the formalism. (For a completely analogous procedure with one Killing vector, see the Appendix of Ref. 1.)

The first step is to divide the space–time M into orbits under the isometries. Two points p and q of M are defined as lying in the same orbit if there is a curve from p to q whose tangent vector is everywhere a linear combination of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$. It will be convenient to impose three additional conditions on the orbits. Firstly, we assume that, at each point of M, $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ are linearly independent. (This assumption

eliminates regions such as the axis in the static, axially symmetric solutions.) Fortunately, the set of points at which the Killing vectors are linearly dependent is necessarily closed and of dimensionality less than four,[14] and so such points, if initially present in M, can be excised. The local linear independence of the Killing fields, together with the vanishing of their commutator, implies[15] that the orbits are two-dimensional surfaces in M. Secondly, we require that the subspace of the tangent space at each point spanned by $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ be timelike. Thus, the orbits are timelike 2-surfaces. The choice of "timelike" rather than "spacelike" is merely for definiteness: The spacelike case differs from the one treated here simply by a few reversals of sign. On the other hand, it is important that the orbits do not become null, for we shall soon introduce a metric on the set S of orbits, and nonsingularity of the metric will be essential. Finally, we assume that S is a smooth (two-dimensional, Hausdorff) manifold, and that the natural mapping $\Psi: M \to S$ is smooth. [Ψ is the mapping which takes each point of M to the orbit (point of S) which passes through that point.] This condition always hold locally. Intuitively, it imposes the global requirement that no orbit "comes back arbitrarily near to itself."

It should be emphasized that the sole purpose of these three assumptions is to permit the introduction of the two-dimensional formalism below. This formalism in turn serves only to (considerably) simplify the calculations. The final transformations (Sec. 3) on exact solutions of Einstein's equation, when expressed in four-dimensional language, do not require any of the assumptions above.

We now have a 2-manifold S. The next step is to discuss tensor fields on S. The treatment of such fields is based on the following result: *there is a natural, one-to-one correspondence (induced by Ψ) between tensor fields $\tilde{T}^{a\cdots c}{}_{b\ldots d}$ on S and tensor fields $T^{a\cdots c}{}_{b\ldots d}$ on M which satisfy*

$$\overset{0}{\xi}{}_a T^{a\cdots c}{}_{b\ldots d} = 0, \cdots, \overset{0}{\xi}{}^d T^{a\cdots c}{}_{b\ldots d} = 0,$$
$$\overset{1}{\xi}{}_a T^{a\cdots c}{}_{b\ldots d} = 0, \cdots, \overset{1}{\xi}{}^d T^{a\cdots c}{}_{b\ldots d} = 0, \quad \text{(A1)}$$

$$\mathcal{L}_{\overset{0}{\xi}} T^{a\cdots c}{}_{b\ldots d} = 0, \quad \mathcal{L}_{\overset{1}{\xi}} T^{a\cdots c}{}_{b\ldots d} = 0. \quad \text{(A2)}$$

Furthermore, this mapping from fields on S to fields on M commutes with the algebraic tensor operations (addition, outer product, and contraction). The proof is essentially the same as that of the analogous result for one Killing vector.[1] As a consequence of this theorem, tensor calculations on the 2-manifold S can be performed without ever leaving M. We simply drop the distinction between tensor fields on S and those on M which satisfy (A1) and (A2). All tensor calculations can, at least formally, be interpreted as concerning fields on the space-time manifold M. However, we are led to concentrate on a certain class of fields on M which are of particular interest, namely, those which are on S [i.e., those which satisfy (A1) and (A2)].

Consider the inner products of the Killing fields:

$$\lambda_{00} = \overset{0}{\xi}{}^a \overset{0}{\xi}{}_a, \quad \lambda_{01} = \overset{0}{\xi}{}^a \overset{1}{\xi}{}_a, \quad \lambda_{11} = \overset{1}{\xi}{}^a \overset{1}{\xi}{}_a. \quad \text{(A3)}$$

It follows immediately from commutativity of the Killing vectors that these three scalar fields are on S. The statement that the 2-flats spanned by $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ are timelike at each point is equivalent to the assertion that the right side of

$$\tau^2 = 2[(\lambda_{01})^2 - \lambda_{00}\lambda_{11}] \quad \text{(A4)}$$

is positive. So (A4) defines a scalar field τ on S. Furthermore, the fields

$$h_{ab} = g_{ab} + 2\tau^{-2}\lambda_{11}\overset{0}{\xi}{}_a\overset{0}{\xi}{}_b + 2\tau^{-2}\lambda_{00}\overset{1}{\xi}{}_a\overset{1}{\xi}{}_b - 4\tau^{-2}\lambda_{01}\overset{0}{\xi}{}_{(a}\overset{1}{\xi}{}_{b)}, \quad \text{(A5)}$$

$$\epsilon_{ab} = 2^{1/2}\tau^{-1}\,\epsilon_{abcd}\overset{0}{\xi}{}^c\overset{1}{\xi}{}^d \quad \text{(A6)}$$

are on S. They are the metric and alternating tensor of S, respectively. The metric, which is positive definite, will be used to raise and lower indices. Note that $\epsilon^{ab}\epsilon_{ab} = 2$. Finally, observe that, if $T^{a\cdots c}{}_{b\ldots d}$ is any tensor field on S, then so is

$$D_p T^{a\cdots c}{}_{b\ldots d} = h_p{}^q h_m{}^a \ldots h_n{}^c h_b{}^r \ldots h_d{}^s \nabla_q T^{m\cdots n}{}_{r\ldots s}. \quad \text{(A7)}$$

Equation (A7) defines the *derivative* D_a on S. Of course, D_a is just the covariant derivative with respect to the metric h_{ab} on S. In particular, we have $D_a h_{bc} = 0$, an equation which can easily be verified using (A5) and (A7).

To summarize, our space-time M, g_{ab} defines an abstract 2-manifold S with positive-definite metric h_{ab} (and, therefore, an alternating tensor and derivative operator). The idea is to find a collection of fields on S which completely characterize the space-time, and to rewrite Einstein's equation as a set of differential equations on these fields. It will turn out that an appropriate collection of fields is given by the metric, the λ's, and two additional constants (to be defined shortly).

It is convenient to first introduce some additional fields. We define the twists of our Killing vectors as follows:

$$\omega^a_{00} = \epsilon^{abcd}\overset{0}{\xi}{}_b\nabla_c\overset{0}{\xi}{}_d,$$
$$\omega^a_{01} = \tfrac{1}{2}\epsilon^{abcd}(\overset{0}{\xi}{}_b\nabla_c\overset{1}{\xi}{}_d + \overset{1}{\xi}{}_b\nabla_c\overset{0}{\xi}{}_d), \quad \text{(A8)}$$
$$\omega^a_{11} = \epsilon^{abcd}\overset{1}{\xi}{}_b\nabla_c\overset{1}{\xi}{}_d.$$

Unfortunately, these three contravariant vector fields on M are not necessarily on S—they may fail to satisfy (A1). To obtain fields on S, we take projections:

$$c_0 = \epsilon^{abcd}\overset{0}{\xi}{}_a\overset{1}{\xi}{}_b\nabla_c\overset{0}{\xi}{}_d, \quad c_1 = \epsilon^{abcd}\overset{0}{\xi}{}_a\overset{1}{\xi}{}_b\nabla_c\overset{1}{\xi}{}_d, \quad \text{(A9)}$$

$$\nu^a_{00} = h^a{}_b\omega^b_{00}, \quad \nu^a_{01} = h^a{}_b\omega^b_{01}, \quad \nu^a_{11} = h^a{}_b\omega^b_{11}. \quad \text{(A10)}$$

It is not difficult to verify, using Einstein's and Killing's equations and the vanishing of the commutator, that c_0 and c_1 must be constants. (The vanishing of both c_0 and c_1 is a necessary and sufficient condition that the 2-flats orthogonal to $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ at each point be integrable.) Einstein's equation implies,

furthermore, that the twists (A8) are curl free. Taking the curls of (A10), and using this fact, we obtain

$$\begin{aligned} D^{[a}\nu_{00}^{b]} &= 2^{-1/2}\tau^{-1}[(c_0)^2\lambda_{01} - c_0c_1\lambda_{00}]\epsilon^{ab}, \\ D^{[a}\nu_{01}^{b]} &= \tfrac{1}{2}2^{-1/2}\tau^{-1}[(c_0)^2\lambda_{11} - (c_1)^2\lambda_{00}]\epsilon^{ab}, \\ D^{[a}\nu_{11}^{b]} &= 2^{-1/2}\tau^{-1}[(c_0c_1\lambda_{11} - (c_1)^2\lambda_{01}]\epsilon^{ab}. \end{aligned} \tag{A11}$$

Although the twists will later play an important role, they do not carry any new information from the space-time M to S. In fact, there are equations, which we now derive, expressing the ν's in terms of the λ's. The vanishing of the commutator of the Killing vectors implies

$$\tfrac{1}{2}D_a\lambda_{01} = \overset{0}{\xi}{}^b\nabla_a\overset{1}{\xi}_b = \overset{1}{\xi}{}^b\nabla_a\overset{0}{\xi}_b. \tag{A12}$$

Substituting into (A12) the expressions[1]

$$\begin{aligned} \nabla_a\overset{0}{\xi}_b &= \tfrac{1}{2}(\lambda_{00})^{-1}\epsilon_{abcd}\overset{0}{\xi}{}^c\omega_{00}^d + (\lambda_{00})^{-1}\overset{0}{\xi}_{[b}D_{a]}\lambda_{00}, \\ \nabla_a\overset{1}{\xi}_b &= \tfrac{1}{2}(\lambda_{11})^{-1}\epsilon_{abcd}\overset{1}{\xi}{}^c\omega_{11}^d + (\lambda_{11})^{-1}\overset{1}{\xi}_{[b}D_{a]}\lambda_{11} \end{aligned} \tag{A13}$$

for the derivative of each Killing vector in terms of its norm and twist, we obtain

$$\begin{aligned} \nu_{00}^a &= 2^{1/2}\tau^{-1}\epsilon^{ab}(-\lambda_{01}D_b\lambda_{00} + \lambda_{00}D_b\lambda_{01}), \\ \nu_{01}^a &= \tfrac{1}{2}2^{1/2}\tau^{-1}\epsilon^{ab}(-\lambda_{11}D_b\lambda_{00} + \lambda_{00}D_b\lambda_{11}), \\ \nu_{11}^a &= 2^{1/2}\tau^{-1}\epsilon^{ab}(-\lambda_{11}D_b\lambda_{01} + \lambda_{01}D_b\lambda_{11}) \end{aligned} \tag{A14}$$

The projected twists, the ν's, are therefore extraneous.[16]

We now derive the first set of Einstein equations. Taking the curls of (A14), and using (A11),

$$\begin{aligned} &\lambda_{01}D^a[\tau^{-1}D_a\lambda_{00}] - \lambda_{00}D^a[\tau^{-1}D_a\lambda_{01}] \\ &\quad = \tau^{-1}[(c_0)^2\lambda_{01} - c_0c_1\lambda_{00}], \\ &\lambda_{11}D^a[\tau^{-1}D_a\lambda_{00}] - \lambda_{00}D^a[\tau^{-1}D_a\lambda_{11}] \\ &\quad = \tau^{-1}[(c_0)^2\lambda_{11} - (c_1)^2\lambda_{00}], \\ &\lambda_{01}D^a[\tau^{-1}D_a\lambda_{11}] - \lambda_{11}D^a[\tau^{-1}D_a\lambda_{01}] \\ &\quad = \tau^{-1}[(c_1)^2\lambda_{01} - c_0c_1\lambda_{11}]. \end{aligned} \tag{A15}$$

Unfortunately, these three equations are linearly dependent, and so cannot be solved immediately for $D^a[\tau^{-1}D_a\lambda_{00}]$, $D^a[\tau^{-1}D_a\lambda_{01}]$ and $D^a[\tau^{-1}D_a\lambda_{11}]$. To obtain a further equation, we proceed directly:

$$\begin{aligned} D^aD_a\lambda_{00} &= h^{ab}\nabla_a(h_b{}^m\nabla_m\lambda_{00}) = 2h^{ab}\nabla_a(\overset{0}{\xi}{}^m\nabla_b\overset{0}{\xi}_m) \\ &= 2h^{ab}\overset{0}{\xi}{}^m\nabla_a\nabla_b\overset{0}{\xi}_m + 2h^{ab}(\nabla_a\overset{0}{\xi}{}^m)(\nabla_b\overset{0}{\xi}_m) \\ &= 2\tau^{-2}\lambda_{00}[(D^a\lambda_{00})(D^a\lambda_{11}) - (D^a\lambda_{01})(D_a\lambda_{01})] \\ &\quad + \tau^{-1}(D_a\tau)(D^a\lambda_{00}) + 2\tau^{-2}(c_0)^2, \end{aligned} \tag{A16}$$

where, in the last step, we have used (A13), $R_{ab} = 0$, and the fact that any Killing vector $\overset{0}{\xi}{}^a$ satisfies

$$\nabla_a\nabla_b\overset{0}{\xi}_c = R_{abcd}\overset{0}{\xi}{}^d. \tag{A17}$$

Equations (A15) and (16) now imply that the three norms, λ_{00}, λ_{01}, and λ_{11} satisfy

$$\begin{aligned} D^a[\tau^{-1}D_a\lambda_{00}] &= 2\tau^{-3}\lambda_{00}[(D^a\lambda_{00})(D_a\lambda_{11}) \\ &\quad - (D^a\lambda_{01})(D_a\lambda_{01})] + 2\tau^{-3}(c_0)^2, \\ D^a[\tau^{-1}D_a\lambda_{01}] &= 2\tau^{-3}\lambda_{01}[(D^a\lambda_{00})(D_a\lambda_{11}) \\ &\quad - (D^a\lambda_{01})(D_a\lambda_{01})] + 2\tau^{-3}c_0c_1, \\ D^a[\tau^{-1}D_a\lambda_{11}] &= 2\tau^{-3}\lambda_{11}[(D^a\lambda_{00})(D_a\lambda_{11}) \\ &\quad - (D^a\lambda_{01}(D_a\lambda_{01})] + 2\tau^{-3}(c_1)^2. \end{aligned} \tag{A18}$$

The first three Einstein equations are (A18). The final one, which relates the curvature of S to the other variables, is derived by commuting D-derivatives. Let k_a be any vector field on S. Then

$$\tfrac{1}{2}\mathcal{R}_{abcd}k^d = D_{[a}D_{b]}k_c = h_{[a}{}^mh_{b]}{}^nh_c{}^p\nabla_m[h_n{}^rh_p{}^s\nabla_rk_s], \tag{A19}$$

where $\mathcal{R}_{abcd}$ is the Riemann tensor[5] of S. Expanding (A19), and eliminating first derivatives of k_a using (A1) and (A2), we have

$$\begin{aligned} \tfrac{1}{2}\mathcal{R}_{abcd}k^d = h_{[a}{}^mh_{b]}{}^nh_c{}^ph^q[&\tfrac{1}{2}R_{mnpq} + \tau^{-2}\lambda_{11}(\nabla_m\overset{0}{\xi}_n)(\nabla_q\overset{0}{\xi}_p) \\ &+ \tau^{-2}\lambda_{00}(\nabla_m\overset{1}{\xi}_n)(\nabla_q\overset{1}{\xi}_p) - \tau^{-2}\lambda_{01}(\nabla_m\overset{0}{\xi}_n)(\nabla_q\overset{1}{\xi}_p) \\ &- \tau^{-2}\lambda_{01}(\nabla_m\overset{1}{\xi}_n)(\nabla_q\overset{0}{\xi}_p) \\ &- \tau^{-2}\lambda_{11}(\nabla_m\overset{0}{\xi}_p)(\nabla_n\overset{0}{\xi}_q) - \tau^{-2}\lambda_{00}(\nabla_m\overset{1}{\xi}_p)(\nabla_n\overset{1}{\xi}_q) \\ &+ \tau^{-2}\lambda_{01}(\nabla_m\overset{0}{\xi}_p)(\nabla_n\overset{1}{\xi}_q) + \tau^{-2}\lambda_{01}(\nabla_m\overset{1}{\xi}_p)(\nabla_n\overset{0}{\xi}_q)]. \end{aligned} \tag{A20}$$

Since k_a is arbitrary (on S), (A20) gives the Riemann tensor of S in terms of that of M. Contracting twice, using (A13) and $R_{ab} = 0$, we obtain the desired result:

$$\begin{aligned} \mathcal{R} = \tau^{-2}&[(D^a\lambda_{00})(D_a\lambda_{11}) - (D^a\lambda_{01})(D_a\lambda_{01})] \\ &+ 6\tau^{-4}[2c_0c_1\lambda_{01} - (c_0)^2\lambda_{11} - (c_1)^2\lambda_{00}]. \end{aligned} \tag{A21}$$

Equations (A18) and (A21) are equivalent to Einstein's. That is to say, a 2-manifold S with (i) a positive-definite metric h_{ab}, (ii) three scalar fields λ_{00}, λ_{01}, and λ_{11}, and (iii) two constants c_0 and c_1 subject to (A18) and (A21), defines a unique space-time with two commuting Killing vectors and with $R_{ab} = 0$, and conversely.

APPENDIX B: WHEN ARE KILLING FIELDS PRESERVED?

Let M, g_{ab} be a source-free solution of Einstein's equation, with commuting Killing fields $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$. Suppose we apply to this metric the transformation (4) with respect to one of the fields, say $\overset{0}{\xi}{}^a$. Then, of course, $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ will still be commuting vector fields on the manifold M, and $\overset{0}{\xi}{}^a$ will be a Killing field for the new metric $\tilde{g}_{ab}$. Under what conditions will $\overset{1}{\xi}{}^a$ also be a Killing field for $\tilde{g}_{ab}$? More generally, what are the necessary and sufficient conditions that the

transformations (4) (with respect to all linear combinations of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$) preserve the presence of two commuting Killing fields? We answer this question here.

Suppose first that neither Killing vector is destroyed by (4), for any linear combination of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$. Apply (4) with respect to $\overset{0}{\xi}{}^a$. Then, in particular, the Lie derivative of

$$\overset{0}{\xi}{}^a\overset{0}{\xi}{}^b\tilde{g}_{ab} = \lambda_{00}[(\cos\theta - \omega_{00}\ \sin\theta)^2 + (\lambda_{00})^2\ \sin^2\theta]^{-1} \tag{B1}$$

by $\overset{1}{\xi}{}^a$ must be zero for all θ. But $\mathcal{L}_{\overset{1}{\xi}}\lambda_{00} = 0$ by (7), and so the quantity

$$\mathcal{L}_{\overset{1}{\xi}}\omega_{00} = \overset{1}{\xi}{}^a\epsilon_{abcd}\overset{0}{\xi}{}^b\nabla^c\overset{0}{\xi}{}^d = -c_0 \tag{B2}$$

must vanish. Similarly for c_1[Eq. (9)]. Thus, a necessary condition that all transformations (4) preserve both Killing fields is $c_0 = c_1 = 0$.

Conversely, we show that, if $c_0 = c_1 = 0$, then $\tilde{g}_{ab}$ necessarily has a pair of commuting Killing vectors.[7] Apply (4) with respect to $\overset{0}{\xi}{}^a$. Taking the Lie derivative of $\tilde{g}_{ab}$ with respect to $\overset{1}{\xi}{}^a$, and using the vanishing of (B2), we have

$$\mathcal{L}_{\overset{1}{\xi}}\tilde{g}_{ab} = 2\tilde{\lambda}\eta_{(a}\mathcal{L}_{\overset{1}{\xi}}\eta_{b)} \tag{B3}$$

Substituting (6), it follows that the vanishing of (B3) (for all θ) is equivalent to

$$\mathcal{L}_{\overset{1}{\xi}}\alpha_a = \mathcal{L}_{\overset{1}{\xi}}\beta_a = 0. \tag{B4}$$

Thus, the new metric $\tilde{g}_{ab}$ will have $\overset{1}{\xi}{}^a$ as a Killing field provided we can choose the vector fields α_a and β_a to satisfy (2), (3), and (B4). There is gauge freedom in α_a and β_a: We can add to either the gradient of a scalar field which has vanishing derivative in the $\overset{0}{\xi}{}^a$-direction. The idea is to use this freedom to satisfy (B4). Suppose a given α_a satisfies (2) but not (B4). Then $\alpha_a + \nabla_a\varphi$ will satisfy both (2) and (B4) provided

$$\overset{0}{\xi}{}^a\nabla_a\varphi = 0, \tag{B5}$$

$$\nabla_b(\overset{1}{\xi}{}^a\nabla_a\varphi) = -\mathcal{L}_{\overset{1}{\xi}}\alpha_b. \tag{B6}$$

Taking the curl of the right side of (B6), using (2), we obtain zero. Hence, $\mathcal{L}_{\overset{1}{\xi}}\alpha_a = \nabla_a k$ for some scalar field k on M. Equations (B5) and (B6) reduce to

$$\overset{0}{\xi}{}^a\nabla_a\varphi = 0, \qquad \overset{1}{\xi}{}^a\nabla_a\varphi = -k. \tag{B7}$$

The integrability condition for (B7) is

$$\mathcal{L}_{\overset{1}{\xi}}(0) - \mathcal{L}_{\overset{0}{\xi}}(-k) = 0. \tag{B8}$$

But (B8) is satisfied, for

$$\begin{aligned}\mathcal{L}_{\overset{0}{\xi}}k &= \overset{0}{\xi}{}^a\nabla_a k = \overset{0}{\xi}{}^a\mathcal{L}_{\overset{1}{\xi}}\alpha_a \\ &= \mathcal{L}_{\overset{1}{\xi}}(\overset{0}{\xi}{}^a\alpha_a) - \alpha_a\mathcal{L}_{\overset{1}{\xi}}\overset{0}{\xi}{}^a = 0,\end{aligned} \tag{B9}$$

where, in the last step, we have used (2) and the vanishing of (B2). A similar argument applies to β_a.

Thus, a necessary and sufficient condition that the transformations (4), for an arbitrary linear combination of $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$, preserve both Killing vectors is $c_0 = c_1 = 0$.

APPENDIX C: GAUGE, PARITY, AND DIMENSION TRANSFORMATION

Consider a pair of fields λ_α and h_{ab} on a 2-manifold S, subject to (17) and (18) (i.e., a solution of Einstein's equation). Define τ by (19), σ by (23), $\overset{0}{A}_\alpha$ by (24) and (25), and $\overset{n}{A}_\alpha$ by (29). At several points in this procedure, we defined a quantity by an expression for its gradient, and so we have the freedom to add to that quantity a constant. That is, not all the fields τ, σ, $\overset{0}{A}_\alpha$ are uniquely determined: There exist gauge transformations. The resulting gauge group (an infinite-dimensional, non-Abelian Lie group) certainly represents a fundamental property of the fields. Furthermore, two alternating tensors were used in Sec. 2, ϵ_{ab} (A6) and $\epsilon_{\alpha\beta\gamma}$ (20). One can reverse the signs of these, with consequent effects on τ, σ, and $\overset{n}{A}_\alpha$. Finally, there is freedom to multiply either of the two metrics h_{ab} [Eq. (A5)] and $G_{\alpha\beta}$ [Eq. (16)] as well as λ_α, by constant factors, giving rise to dimensions associated with τ, σ, and $\overset{n}{A}_\alpha$. In this Appendix we shall derive the formulae for the effects of these transformations. It is interesting that these transformations can be extended to the operator $\mathcal{T}_\alpha$ (that which gives the infinitesimal change in an exact solution of Einstein's equation) of Sec. 3.

We first consider the gauge transformation which arises from (23). This equation defines σ only up to an additive constant. Hence (taking the infinitesimal case), we ask for the effect, to first order in ϵ, of

$$\sigma \to \sigma + \epsilon. \tag{C1}$$

Since τ and λ_α are invariant under (C1), (24) implies that ω_α, and hence $\overset{0}{A}_\alpha$, are invariant. But (27) implies that, under (C1), $\overset{1}{A}_\alpha \to \overset{1}{A}_\alpha + 2\epsilon\overset{0}{A}_\alpha$, to first order in ϵ. Then (28) gives $\overset{2}{A}_\alpha \to \overset{2}{A}_\alpha + 3\epsilon\overset{1}{A}_\alpha$. Proceeding inductively, using (29),

$$\overset{n}{A}_\alpha \to \overset{n}{A}_\alpha + (n+1)\epsilon\overset{n-1}{A}_\alpha. \tag{C2}$$

A second gauge transformation arises from adding a constant to ω_α, defined by (24). This clearly leaves τ, σ, and λ_α invariant, and so has the effect

$$\overset{0}{A}_\alpha \to \overset{0}{A}_\alpha + \overset{0}{P}_\alpha, \tag{C3}$$

where $\overset{0}{P}_\alpha$ is real and constant. Proceeding inductively as before, we have, to first order in $\overset{0}{P}_\alpha$,

$$\overset{n}{A}_\alpha \to \overset{n}{A}_\alpha - \epsilon_{\alpha\mu\nu}\overset{0}{P}{}^\mu\overset{n-1}{A}{}^\nu. \tag{C4}$$

Finally, there is freedom to add a constant to $\overset{n}{A}_\alpha$ [Eq. (29)], leaving the earlier A's invariant:

$$\overset{n}{A}_\alpha \to \overset{n}{A}_\alpha + (n+1)\overset{n}{P}_\alpha, \qquad \overset{m}{A}_\alpha \to \overset{m}{A}_\alpha \quad (m < n). \tag{C5}$$

Using induction on (29), (C5) has the following effect on the remaining $\overset{m}{A}_\alpha$:

$$\overset{m}{A}_\alpha \to \overset{m}{A}_\alpha - \epsilon_{\alpha\mu\nu}\overset{n}{P}^\mu \overset{n-m-1}{A}^\nu, \tag{C6}$$

to first order in $\overset{n}{P}_\alpha$.

These infinitesimal gauge transformations can be dealt with most easily by introducing appropriate operators. The action of (C1) is expressed by the operator $\mathcal{W}$, defined by

$$\begin{aligned} &\mathcal{W}\sigma = 1, \quad \mathcal{W}\tau = 0, \\ &\mathcal{W}\overset{m}{A}_\alpha = (m+1)\overset{m-1}{A}_\alpha \quad m \geq 1, \quad \mathcal{W}\overset{0}{A}_\alpha = 0. \end{aligned} \tag{C7}$$

For each integer $n \geq -1$, define an operator $\overset{n}{\mathcal{W}}_\alpha$ by

$$\begin{aligned} &\overset{n}{\mathcal{W}}_\alpha \sigma = 0, \quad \overset{n}{\mathcal{W}}_\alpha \tau = 0, \\ &\overset{n}{\mathcal{W}}_\alpha \overset{m}{A}_\beta = 0 \quad (m < n), \quad \overset{n}{\mathcal{W}}_\alpha \overset{m}{A}_\beta = (n+1)G_{\alpha\beta} \quad (m = n), \\ &\overset{n}{\mathcal{W}}_\alpha \overset{m}{A}_\beta = \epsilon_{\alpha\beta\mu}\overset{m-n-1}{A}^\mu \quad (m > n), \end{aligned} \tag{C8}$$

Then, for $n = -1$, (C8) represents an infinitesimal rotation in the (indefinite, three-dimensional) vector space of the α's, while the action of (C4) is given by (C8) for $n = 0$, and that of (C5) by (C8) for $n > 0$. The commutator of two infinitesimal gauge transformations is another. These commutators follow immediately from (C7) and (C8):

$$\begin{aligned} &[\mathcal{W}, \overset{n}{\mathcal{W}}_\alpha] = (n+1)\overset{n+1}{\mathcal{W}}_\alpha, \\ &[\overset{n}{\mathcal{W}}_\alpha, \overset{m}{\mathcal{W}}_\beta] = \epsilon_{\alpha\beta\mu}\overset{n+m+1}{\mathcal{W}}{}^\mu. \end{aligned} \tag{C9}$$

Finally, the commutators of the gauge operators and the transformation operator $\mathcal{T}_\alpha$ (45) are.[17]

$$\begin{aligned} &[\mathcal{W}, \mathcal{T}_\alpha] = -\overset{-1}{\mathcal{W}}_\alpha, \\ &[\overset{n}{\mathcal{W}}_\alpha, \mathcal{T}_\beta] = \epsilon_{\alpha\beta\mu}\overset{n-1}{\mathcal{W}}{}^\mu. \end{aligned} \tag{C10}$$

We next consider parity transformations. Reversing the sign of ϵ_{ab} leaves τ and λ_α invariant, but reverses the sign of σ [by (23)] and of ω_α [by (24)]. Hence, the effect on $\overset{0}{A}_\alpha$ is $\overset{0}{A}_\alpha \to -\overset{0}{\bar{A}}_\alpha$, where a bar denotes complex conjugation. Then (29) implies that, for general n,

$$\overset{n}{A}_\alpha \to (-1)^{n+1}\overset{\bar{n}}{A}_\alpha. \tag{C11}$$

Similarly, reversing the sign of $\epsilon_{\alpha\beta\gamma}$ leaves τ, σ, and λ_α invariant, but, by (24) reverses that of ω_α. Hence, for general n,

$$\overset{n}{A}_\alpha \to -\overset{\bar{n}}{A}_\alpha. \tag{C12}$$

Both parities of $\mathcal{T}_\alpha$ are negative.

We next consider the λ-weight. If h_{ab}, λ_α is a solution of (17) and (18), then so is $h_{ab}, c\lambda_\alpha$, where c is any nonzero constant. The λ-weight of a quantity will be defined as the power of c by which it is multiplied under $(h_{ab}, \lambda_\alpha) \to (h_{ab}, c\lambda_\alpha)$. So λ_α and τ [and, by (23) σ] have λ-weight 1, while h_{ab} has λ-weight zero. From (24), ω_α, and hence $\overset{0}{A}_\alpha$, has λ-weight 1. Then (29) implies that $\overset{n}{A}_\alpha$ has λ-weight $(n+1)$. The operator $\mathcal{T}_\alpha$ increases the λ-weight by 1.

Equations (17) and (18) are invariant under constant conformal transformations on $S(h_{ab} \to \Omega^2 h_{ab})$ provided λ_α is left unchanged. Since (23), (24), and (29) are conformally invariant, all the fields except h_{ab} are invariant.

Finally, we consider the "conformal transformation"

$$G_{ab} \to \Omega^2 G_{\alpha\beta}, \tag{C13}$$

where Ω is a real constant. One might think that the G dimension of a quantity is best defined as the power of Ω by which it is multiplied, under (C13). This, however, would be inconvenient, for the G dimension of an indexed field would then depend on the location (raised or lowered) of its indices. To avoid this inconvenience, we define as the G dimension of a field: [the power of Ω by which it is multiplied under (C13)] − (the number of lowered Greek indices) + (the number of raised Greek indices). Then the G dimension is invariant under raising and lowering of indices, and under contraction.[18] The G dimension of $G_{\alpha\beta}$ and $\epsilon_{\alpha\beta\gamma}$ are zero. We are free to choose the G dimension of λ_α arbitrarily, for two distinct choices differ only by a λ-weight transformation. We therefore choose for λ_α G dimension zero, whence all fields have G dimension zero.

---

* Present address: The Enrico Fermi Institute, 933 E. 56 St., Chicago, Ill. 60637.

1 R. Geroch, J. Math. Phys. 12, 918 (1971).

2 This new solution is always defined locally, even in regions in which λ changes sign. However, there may be singularities which arise from global obstructions to integrating (1), (2), and (3). See Ref. 1.

3 The special case of the transformation (4) in which is hypersurface orthogonal (i.e., $\omega = 0$) was earlier obtained by Ehlers, in *Les théories relativistes de la gravitation* (CNRS, Paris, 1959). Unfortunately, hypersurface orthogonality is not preserved by the transformations, and so the possibility of iterating the transformations to obtain further solutions did not arise.

4 The transformation (4) with ξ^a replaced by some constant multiple of ξ^a is identical with (4) with θ replaced by a certain function of θ.

5 Our conventions are as follows: $D_{[a}D_{b]}k_c = \frac{1}{2}\mathcal{R}_{abcd}k^d$, $\mathcal{R}_{ac} = \mathcal{R}_{amc}{}^m$.

6 See, for example, F. A. E. Pirani, in *Brandeis Summer Institute in Theo. Physics, 1964* (Prentice-Hall, Englewood Cliffs, N.J., 1965).

7 Equations (2) and (3) permit the addition of certain gradients to α_a and β_a. This alters the final metric (4), but only by applying to it a diffeomorphism. Consequently, although (4) always has a pair of commuting Killing fields, these fields may not be $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$. It is always possible, on the other hand, to choose α_a and β_a so that $\overset{0}{\xi}{}^a$ and $\overset{1}{\xi}{}^a$ will themselves be the Killing fields. The existence of this possibility is not surprising: any two 4-manifolds, on each of which there is specified a pair of pointwise linearly independent, commuting vector fields, are locally identical.

8 Once (46) has been integrated, the integration of (47) is easy. [See (37).] Hence, we need not be concerned further with (47).

9 For certain special choices for the curve $\gamma(t)$, only a finite number of A's are required for the solution.

10 R. Geroch, J. Math. Phys. 11, 2580 (1970).

11 One could certainly admit curves more general than the piecewise smooth ones. What is the largest admissible class?

12 More precisely, $\gamma(t)$ and $\gamma'(t)$ are equivalent if there is a continuous mapping φ: $[0,1]\times[0,1] \to V$ such that

(i) $\varphi(s,0) = \gamma(0) = \gamma'(0)$ and $\varphi(s,1) = \gamma(1) = \gamma'(1)$ for all s,
(ii) $\varphi(0,t) = \gamma(t)$ and $\varphi(1,t) = \gamma'(t)$ for all t, and if
(iii) the range of φ is a subset of the union of the ranges of γ and γ'. I wish to thank M. MacCallum for this definition.

[13] One would only expect this statement to hold locally, in a suitable sense. An analogous global result may exist, but one would have to eliminate space-times which contain "holes," or which are obtained by making "identifications."

[14] See, for example, R. Geroch, Commun. Math. Phys. **13**, 180 (1969).

[15] See, for example, J. A. Schouten, *Ricci Calculus* (Springer-Verlag, Berlin, 1954), pp. 78.

[16] A number of theorems involving conditions under which commuting Killing fields are hypersurface orthogonal, or orthogonal to each other, follow from (A14). For example: If two Killing fields commute and are everywhere orthogonal to each other, and $R_{ab} = 0$, then each is hypersurface orthogonal.

[17] Compare Eqs. (C10) and (C9). As far as its commutation relations are concerned, $\mathcal{T}_{\omega n}$ behaves exactly like an infinitesimal gauge transformation, $\mathcal{W}_\alpha$ for $n = -2$!

[18] This scheme is also convenient for defining the dimensions of quantities in general relativity. One considers a solution of all his equations, with metric g_{ab}, and asks for a corresponding solution with metric $\Omega^2 g_{ab}$ (Ω = const). Then each field is multiplied by some power of Ω. One defines the dimension of each field (units of seconds) as in the text. This procedure gives the answer one expects physically (e.g., for Maxwell fields, stress-energy, pressure, etc.).

Relativistic Spin-Zero Wave Equation*

Joseph V. Lepore and Robert J. Riddell, Jr.
Lawrence Radiation Laboratory, University of California, Berkeley, California 94720
(Received 5 August 1971)

We have studied the solutions of a wave equation which describes a spin-zero particle in the Coulomb field of a nucleus. An interesting feature of this equation is that the kernel is not of the Fredholm type. The behavior of the momentum space wavefunction for large momentum is not determined solely by the angular momentum state but, as in the cases of the Dirac and Klein–Gordon equations, it depends on the electric charge as well. Our analysis of the asymptotic properties is based on a Mellin transformation of the momentum space equation. This leads to a singular integral equation with a Cauchy-type kernel which may be treated by standard methods. The equation is shown to have unique solutions.

I. INTRODUCTION

When we began a phenomenological analysis of pion alpha-particle scattering sometime ago, we were faced with the problem of choosing a wave equation incorporating two-particle relativistic effects. We wished to describe both electromagnetic and strong interactions. At first, the Klein–Gordon equation appeared to be a likely possibility, but it has no probability interpretation so we were led to consider the problem of Coulomb scattering for two spin-zero particles from the field theoretic point of view.

This led us to the following wave equation for two free particles of mass m_1, m_2 and momenta $\mathbf{p}_1, \mathbf{p}_2$:

$$[(\mathbf{p}_1^2 + m_1^2)^{1/2} + (\mathbf{p}_2^2 + m_2^2)^{1/2}]\psi(\mathbf{p}_1, \mathbf{p}_2) = P_0\psi(\mathbf{p}_1, \mathbf{p}_2),$$

where P_0 is the total energy, and we choose $\hbar = c = 1$. If the Coulomb interaction is included, an additional term,

$$\int V(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_1', \mathbf{p}_2')\psi(\mathbf{p}_1', \mathbf{p}_2')d^3\mathbf{p}_1' d^3\mathbf{p}_2',$$

describing the Coulomb interaction V appears.

This equation is a member of the class derived on the basis of general relativistic principles by Bakamjian and Thomas[1] almost two decades ago. Such an equation has been known[2] even longer. The relationship between this equation and the Klein–Gordon equation has been discussed by Feshbach and Villars.[3] More recently, Zemach[4] has analyzed the relation between this equation and that for the two-body Green's function[5] defined by Schwinger.

The Bakamjian–Thomas equation has been studied by a number of authors[6] during the past few years. They have concerned themselves with the case of short-range interactions. In the following pages we present the theory of the equation for the case of an interaction which is the time component of a vector field. In this case the resulting integral equation is not of the Fredholm type. Its solutions may be shown to behave as p^{-s} for large momentum where the specific value of s depends on the angular momentum state and the strength of the interaction. Because of this behavior, a Mellin transformation of the momentum space wave function seems particularly appropriate. When such a transformation is carried out, the kernel of the new equation is found to have a Cauchy-type singularity. The choice of a contour of integration for the inverse Mellin transformation is made by demanding that the wave function be integrable for large momenta and that the transformed kernel be Hermitian. The integral equation may then be reduced to a Fredholm equation by a standard method which we describe in detail. We are thus able to prove the existence of unique solutions for both bound state and scattering wavefunctions.

II. THE BAKAMJIAN–THOMAS EQUATION FOR COULOMB SCATTERING

We consider two spin-zero fields ϕ, χ of masses m and M, respectively, interacting through a Coulomb interaction. The Schrödinger representation is employed. Free particle states are normalized by

$$\langle \mathbf{p}' | \mathbf{p} \rangle = p_0\delta(\mathbf{p}' - \mathbf{p}), \tag{1}$$

where p_0 is the free particle energy and δ is the Dirac delta function. This implies commutation relations of the form

$$[a(\mathbf{p}), a^+(\mathbf{p}')] = p_0\delta(\mathbf{p} - \mathbf{p}') \tag{2}$$

for the operators $a(\mathbf{p}), b(\mathbf{p})$ and $A(\mathbf{p}), B(\mathbf{p})$ associated with the fields ϕ and χ, respectively. For ϕ we write

$$\phi(\mathbf{r}) = \frac{1}{(2\pi)^{3/2}(2)^{1/2}} \int \frac{d^3\mathbf{p}}{p_0} [a(\mathbf{p})e^{i\mathbf{p}\cdot\mathbf{r}} + b^+(\mathbf{p})e^{-i\mathbf{p}\cdot\mathbf{r}}] \tag{3}$$

and correspondingly for $\chi(\mathbf{r})$. If $\psi(\mathbf{p}, \mathbf{q})$ is a function of the variables $\mathbf{p}$ and $\mathbf{q}$ referring to two different particles, we may introduce a two-particle state vector $|\Psi\rangle$ by writing

$$|\Psi\rangle = \iint \frac{d^3\mathbf{p}}{p_0} \frac{d^3\mathbf{q}}{q_0} \psi(\mathbf{p}, \mathbf{q})a^+(\mathbf{p})A^+(\mathbf{q})|0\rangle, \tag{4}$$

where $|0\rangle$ is the vacuum state vector. One finds

$$\langle \Psi | \Psi \rangle = \iint \frac{d^3\mathbf{p}}{p_0} \frac{d^3\mathbf{q}}{q_0} |\psi(\mathbf{p}, \mathbf{q})|^2. \tag{5}$$

The Schrödinger equation which we seek is just

$$\langle \mathbf{p}, \mathbf{q} | H | \Psi \rangle = i \frac{\partial}{\partial t} \langle \mathbf{p}, \mathbf{q} | \Psi \rangle, \tag{6}$$

where H is the Hamiltonian of the system. The noninteraction part of H contributes $[(\mathbf{p}^2 + m^2)^{1/2} + (\mathbf{q}^2 + M^2)^{1/2}]\psi(\mathbf{p}, \mathbf{q})$ to the left-hand side of this equation, and the Coulomb interaction term is

$$H_c = Ze^2 \int \frac{\rho_m(\mathbf{r})\rho_M(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3\mathbf{r} d^3\mathbf{r}', \tag{7}$$

where ρ_m and ρ_M are the charge densities of the two fields. In terms of the charges e and Ze for the ϕ and χ fields (the particles "a" and "A" have charges e and Ze, respectively),

$$\rho_m = ie(\phi^+\pi^+ - \pi\phi), \tag{8}$$

where π, π^+ are the fields canonically conjugate to ϕ, ϕ^+, respectively. The representation of the $\pi(\mathbf{r})$ field is

$$\pi(\mathbf{r}) = \frac{i}{(2\pi)^{3/2}} \frac{1}{2^{1/2}} \int d^3\mathbf{p}[a^+(\mathbf{p})e^{-i\mathbf{p}\cdot\mathbf{r}} - b(\mathbf{p})e^{i\mathbf{p}\cdot\mathbf{r}}]. \quad (9)$$

Similar expressions describe ρ_M and the χ field. The evaluation of the Coulomb contribution is straightforward. One finds

$$\langle \mathbf{p}, \mathbf{q} | H_c | \Psi \rangle = \times \frac{Ze^2}{8\pi^2} \int \frac{d^3\mathbf{p}'}{p_0'} \frac{d^3\mathbf{q}'}{q_0'} (p_0 + p_0')(q_0 + q_0') \times \frac{\delta(\mathbf{p} + \mathbf{q} - \mathbf{p}' - \mathbf{q}')}{(\mathbf{p} - \mathbf{p}')^2} \psi(\mathbf{p}', \mathbf{q}'). \quad (10)$$

Since we were immediately interested in pion-helium scattering where $m_\pi \ll m_{He}$, we were led to the approximate equation for the case of an infinitely massive field χ. If P_0 now denotes only the energy of the particle of finite mass, we find

$$(\mathbf{p}^2 + m^2)^{1/2}\psi(\mathbf{p}) + \frac{Ze^2}{4\pi^2} \int \frac{d^3\mathbf{p}'}{p_0'} \frac{(p_0 + p_0')}{(\mathbf{p} - \mathbf{p}')^2} \psi(\mathbf{p}') = P_0\psi(\mathbf{p}). \quad (11)$$

This is the equation we will study.

III. SOLUTIONS OF THE BAKAMJIAN–THOMAS COULOMB EQUATION

This section will be devoted to a study of the general properties of Eq. (11)[7] when the interaction is attractive. This restriction will be removed later. A partial-wave decomposition yields

$$(p^2 + m^2)^{1/2}\psi_l(p) + \frac{Ze^2}{2\pi} \int_0^\infty dp' \frac{(p_0 + p_0')}{pp_0'} \times p' Q_l\left(\frac{p^2 + p'^2}{2pp'}\right)\psi_l(p') = P_0\psi_l(p), \quad (12)$$

where Q_l is the Legendre function of the second kind and $\psi_l(p)$ is the new wavefunction. If we try to write Eq. (12) in standard integral equation form, the resulting kernel is not symmetric. To find an equation with a symmetric kernel, one may introduce ψ_l^s by

$$p_0^{1/2}\psi_l^s = \{[P_0 - (p^2 + m^2)^{1/2}]\}^{1/2} p\psi_l(p). \quad (13)$$

The new equation is then

$$\psi_l^s(p) = \frac{Ze^2}{2\pi} \int_0^\infty dp' \frac{(p_0 + p_0')}{(p_0p_0')^{1/2}} \times \frac{Q_l([(p^2 + p'^2)/2pp'])}{[P_0 - (p^2 + m^2)^{1/2}]^{1/2}[P_0 - (p'^2 + m^2)^{1/2}]^{1/2}}. \quad (14)$$

Equation (14) has a kernel which is not of the Fredholm type. To see this, we consider the integral of the square of the kernel (the Fredholm norm):

$$N^2 = \left(\frac{Ze^2}{2\pi}\right)^2 \int_0^\infty dp \int_0^\infty dp' \frac{(p_0 + p_0')^2}{p_0p_0'} \times \frac{Q_l^2([(p^2 + p'^2)/2pp'])}{|[P_0 - (p^2 + m^2)^{1/2}][P_0 - (p'^2 + m^2)^{1/2}]|}. \quad (15)$$

The kernel will be non-Fredholm if the energy P_0 is in the scattering region, because the energy denominators can then vanish. This difficulty is common to scattering integral equations and can be readily removed,[8] so we will ignore it. If we set $p' = \alpha p$ and take account of the symmetry in p and p' we find

$$N^2 = 2\left(\frac{Ze^2}{2\pi}\right)^2 \int_0^\infty \frac{dp}{p} \int_0^1 d\alpha \times \frac{[(1 + m^2/p^2)^{1/2} + (\alpha^2 + m^2/p^2)^{1/2}]\, Q_l^2([(1 + \alpha^2)/2\alpha])}{[(1 + m^2/p^2)(\alpha^2 + m^2/p^2)]^{1/2}|[P_0/p - (1 + m^2/p^2)^{1/2}][P_0/p - (\alpha^2 + m^2/p^2)^{1/2}]|}. \quad (16)$$

As p tends to zero the integral is well behaved; when p becomes large, however, the integral diverges logarithmically. The integrand is positive definite and for any nonzero region in α the logarithmic divergence is present. (Note that when $\alpha \to 1$ no trouble arises since Q_l diverges only logarithmically and is therefore integrable.) It should be noted that if we had considered the interaction appropriate to the time component of a vector meson field of mass μ the argument of the function Q_l would be replaced according to

$$(p^2 + p'^2)/2pp' \to (p^2 + p'^2 + \mu^2)/2pp'; \quad (17)$$

but similar arguments to those above would show the kernel still not to be of the Fredholm type. Since the non-Fredholm nature of the kernel is related to its large momentum behavior, our next task is to study the behavior of the equation for large momenta. For this purpose it is convenient to use the unsymmetric form of the integral equation, Eq. (12). Thus for $p_0 \gg P_0$,

$$\psi_l(p) \cong -\frac{Ze^2}{2\pi} \int_0^\infty dp' \frac{(p_0 + p_0')}{p_0p_0'} \frac{p'}{p} Q_l\left(\frac{p^2 + p'^2}{2pp'}\right)\psi_l(p'). \quad (18)$$

The integral representation for Q_l, i.e.,

$$Q_l(z) = \tfrac{1}{2} \int_{-1}^1 dt \frac{P_l(t)}{z - t}, \quad (19)$$

may then be used to give

$$\psi_l(p) = -\frac{Ze^2}{2\pi} \int_{-1}^1 dt P_l(t)$$

$$\times \int_0^\infty dp' \frac{(p_0 + p_0')}{p_0 p_0'} \frac{p'^2 \psi_l(p')}{(p^2 + p'^2 - 2pp't)}. \quad (20)$$

We now conjecture that solutions of this equation behave as $p^{-\xi}$ for large p. We are thus led to examine the two integrals

$$I_1 = -\frac{Ze^2}{2\pi p} \int_{-1}^{1} dt P_l(t) \int_0^\infty \frac{dp'(p')^{2-\xi}}{(p^2 + p'^2 - 2pp't)}$$

and (21)

$$I_2 = -\frac{Ze^2}{2\pi} \int_{-1}^{1} dt P_l(t) \int_0^\infty \frac{(p')^{2-\xi}}{p_0'(p^2 + p'^2 - 2pp't)}.$$

The integrals may be evaluated by standard contour integration methods. Both have branch points at the origin; I_2 has additional branch points at $\pm im$. We thus find that

$$I_1 = -\frac{Ze^2}{2\pi p} \frac{e^{i\pi\xi}}{2i \sin\pi\xi} \int_0^\pi \sin\theta d\theta P_l(\cos\theta) \times \int_C \frac{dp'(p')^{2-\xi}}{(p' - pe^{i\theta})(p' - pe^{-i\theta})}$$

and (22)

$$I_2 = -\frac{Ze^2}{2\pi} \frac{e^{i\pi\xi}}{2i \sin\pi\xi} \int_0^\pi \sin\theta d\theta P_l(\cos\theta) \times \int_C \frac{dp'(p')^{2-\xi}}{p_0'(p' - pe^{i\theta})(p' - p^{-i\theta})},$$

where we have set

$$t = \cos\theta, \quad (23)$$

and C is a contour from $+\infty$ to $+\infty$ taken around the origin in the counterclockwise direction below and above the branch cut which has been taken along the real axis from the origin to $+\infty$ as shown in Fig. 1. The integrals in Eq. (21) are well defined in the neighborhood of the origin and at ∞ if

$$1 < \xi < 2. \quad (24)$$

The integral I_1 may be evaluated by the method of residues. The integrand has poles at $\arg p' = \theta$, $2\pi - \theta$ (the last value obtains since we may not pass through the branch cut along the real axis; see Fig. 1). One finds for I_1, if θ is less than $\pi/2$,

$$I_1 = \frac{Ze^2}{2} \frac{p^{-\xi}}{\sin\pi\xi} \int_0^\pi d\theta P_l(\cos\theta) \sin[(2 - \xi)(\pi - \theta)]. \quad (25)$$

The integral I_2 may be treated in a similar manner except that account must be taken of the additional branch cuts from im to ∞ and $-im$ to $-\infty$. If θ is again assumed to be less than $\pi/2$ and if I_{2B} represents the contribution from the branch cuts, one has, in the limit $p \gg m$,

$$I_2 = I_{2B} - \frac{Ze^2}{2} \frac{p^{-\xi}}{\sin\pi\xi} \int_0^\pi d\theta P_l(\cos\theta) \sin[(1 - \xi)(\pi - \theta)]. \quad (26)$$

Let us now consider the integral over the branch cuts I_{2B}. Since we are dealing with a square root singularity, it follows that we need only integrate over the portions of the contour which lie in the left plane providing we double the result. Next let us consider an integral $I_{2B'}$ of the integrand over the path C' in Fig. 1, which lies in the left-half plane and which connects the branch points $\pm im$. Clearly

$$I_{2B} + I_{2B'} = 0. \quad (27)$$

The contour for $I_{2B'}$ may be taken along the imaginary axis, except for a small indentation of radius ρ to the left of the origin which allows one to avoid the branch cut to the origin. We thus have

$$I_{2B'} = \frac{Ze^2}{2\pi \sin\pi\xi} \int_0^\pi \sin\theta d\theta P_l(\cos\theta) \left[\int_\rho^m \left(\frac{y^{2-\xi} dy}{(m^2 - y^2)^{1/2}} \times \frac{e^{i\pi\xi/2}}{(y - pe^{i(\theta - \pi/2)})(y + pe^{-i(\theta - \pi/2)})} + \text{c.c.} \right) - \rho^{3-\xi} e^{i\pi\xi} \int_{\pi/2}^{3\pi/2} \times \frac{e^{i(3-\xi)\alpha} d\alpha}{(m^2 + \rho^2 e^{2i\alpha})^{1/2} (\rho e^{i\alpha} - pe^{i\theta})(\rho e^{i\alpha} - pe^{-i\theta})} \right]. \quad (28)$$

The last integral in Eq. (28) vanishes as $\rho \to 0$ for Re $\xi < 3$. Thus one finds in the limit $p \to \infty$,

$$I_{2B'} \approx -\frac{Ze^2}{2\pi p^2 \sin\pi\xi} \int_0^\pi \sin\theta d\theta P_l(\cos\theta) \int_0^m \frac{y^{2-\xi} dy}{(m^2 - y^2)^{1/2}} \times \left[\frac{e^{i\pi\xi/2}}{(1 + (y/p)e^{i(\theta - \pi/2)})(1 - (y/p)e^{-i(\theta - \pi/2)})} + \text{c.c.} \right]. \quad (29)$$

Therefore I_{2B} behaves as p^{-2} so under the restrictions in Eq. (24), I_{2B} does not contribute to the asymptotic behavior. Thus

$$I_2 = -\frac{Ze^2 p^{-\xi}}{2 \sin\pi\xi} \int_0^\pi d\theta P_l(\cos\theta) \sin[(1 - \xi)(\pi - \theta)]. \quad (30)$$

If I_1 and I_2 are now combined, one finds that for self-consistency a solution whose asymptotic form is $p^{-\xi}$ requires that

$$\frac{Ze^2}{2 \sin\pi\xi} \int_0^\pi d\theta P_l(\cos\theta) \{\sin[(2 - \xi)(\pi - \theta)] - \sin[(1 - \xi)(\pi - \theta)]\} = 1. \quad (31)$$

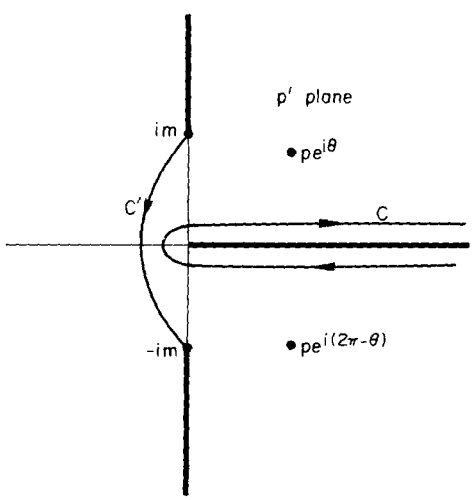


FIG. 1. Structure of the p' plane together with contours used in evaluating integrals.

This may also be written as

$$\frac{(-1)^l Ze^2}{\sin\pi\xi}\int_0^\pi d\theta P_l(\cos\theta)\sin\frac{\theta}{2}\cos(\xi-3/2)\theta = 1. \tag{32}$$

The integral vanishes as $\xi \to 2$, as it should, since otherwise there would be a pole in either I_1 or I_2 in contradiction to the condition (24).

The integrals in Eqs. (31) and (32) may be evaluated in a straightforward manner.[9] When l is even, we find

$$\frac{Ze^2}{2\pi}\sum_{k=0}^{l}\binom{l}{k}B(l-k+\tfrac{1}{2},k+\tfrac{1}{2}) \times\left(\frac{\tan\pi\xi/2}{l-2k+2-\xi}-\frac{\cot\pi\xi/2}{l-2k+1-\xi}\right)=1, \tag{33}$$

and when l is odd,

$$\frac{Ze^2}{2\pi}\sum_{k=0}^{l}\binom{l}{k}B(l-k+\tfrac{1}{2},k+\tfrac{1}{2}) \times\left(\frac{\cot\pi\xi/2}{l-2k+2-\xi}-\frac{\tan\pi\xi/2}{l-2k+1-\xi}\right)=1. \tag{34}$$

In these expressions, the symbol $\binom{l}{k}$ is the usual binomial coefficient and $B(x,y)$ is the beta function of x, y.

The asymptotic behavior just developed strongly suggests that the Bakamjian–Thomas equation be studied by using a Mellin transformation.[10] We now turn to that task, which will verify rigorously that the asymptotic behavior is indeed given by a solution of Eq. (33) or Eq. (34), and, further, will lead to a method for obtaining a unique solution of the singular equation (12).

IV. MELLIN TRANSFORMATION OF THE BAKAMJIAN–THOMAS EQUATION

The Mellin transformation and its inverse are defined by the equations[11]

$$\psi_l(s) = \int_0^\infty \psi_l(p)p^{s-1}dp \tag{35}$$

and

$$\psi_l(p) = \frac{1}{2\pi i}\int_C \psi_l(s)p^{-s}ds, \tag{36}$$

where C goes from $-i\infty$ to $i\infty$. The contour C must be chosen appropriately in order to effect a solution. We note that from Eq. (36) it follows that the asymptotic behavior of $\psi_l(p)$ as $p \to \infty$ is determined by the singularity in $\psi_l(s)$ with the smallest Re(s) to the right of C, while the behavior as $p \to 0$ is determined by the singularity with the largest Re(s) to the left of C. From these relations one finds the transformed integral equation

$$\psi_l(s) = \frac{1}{2\pi i}\int_C K_l(s,s')\psi_l(s')ds', \tag{37}$$

where the kernel $K_l(s,s')$ is given by

$$K_l(s,s') = \frac{Ze^2}{2\pi}\int_0^\infty dp\int_0^\infty dp'\frac{p^{s-1}(p_0+p_0')p'}{(P_0-p_0)p_0'p} \times Q_l\left(\frac{p^2+p'^2}{2pp'}\right)(p')^{-s'}dp'. \tag{38}$$

The conditions for the existence of $K_l(s,s')$ must now be examined. A consideration that the integrals over p and p' be convergent at both limits of integration gives the requirements

$$\begin{aligned} &as\ p\to 0, && \mathrm{Re}(s) > -l,\\ &as\ p\to\infty, && \mathrm{Re}(s) < l+2,\\ &as\ p'\to 0, && \mathrm{Re}(s') < l+3,\\ &as\ p'\to\infty, && \mathrm{Re}(s') > -l+1. \end{aligned} \tag{39}$$

It may be noted that the factor $(P_0 - p_0)^{-1}$ may be expanded in an appropriate manner according to whether $|P_0/p_0|$ is greater or less than 1 and that such an expansion will not alter our conclusions about the domain of existence of K since each successive term is as well behaved at the origin and is better behaved at infinity than the one for which $P_0 = 0$.

Alternatively if the kernel K is divided into two parts, the first of which, $\bar{K}$, is obtained by setting $P_0 = 0$ in K, and the second is simply the difference between K and $\bar{K}$, by such an expansion argument for large p as has just been given one sees that the domain of existence for K is the same as that for $\bar{K}$. We find

$$\bar{K}_l(s,s') = -\frac{Ze^2}{2\pi}\int_0^\infty dp\int_0^\infty dp' p^{s-2} \times\frac{(p_0+p_0')}{p_0p_0'}(p')^{-s+1}Q_l\left(\frac{p^2+p'^2}{2pp'}\right). \tag{40}$$

To carry out the integrals, we again replace Q_l by its integral representation, Eq. (19), to get

$$\bar{K}_l(s,s') = -\frac{Ze^2}{2\pi}\int_0^\pi \sin\theta d\theta P_l(\cos\theta)\int_0^\infty dp\int_0^\infty dp' p^{s-1} \times\left(\frac{1}{p_0}+\frac{1}{p_0'}\right)\frac{(p')^{-s'+2}}{(p'-pe^{i\theta})(p'-pe^{-i\theta})}. \tag{41}$$

Consider now the integral

$$I_1(s,s') = \int_0^\infty dp\int_0^\infty dp'\frac{p^{s-1}(p')^{-s'+2}}{p_0(p'-pe^{i\theta})(p'-pe^{-i\theta})} = \frac{e^{i\pi s'}}{2i\sin\pi s'}\int_0^\infty\frac{dp}{p_0}\int_C\frac{dp'(p')^{-s'+2}}{(p'-pe^{i\theta})(p'-pe^{-i\theta})}. \tag{42}$$

This is one of the terms in Eq. (41). The contour C is the same as that in Eq. (22). We integrate first over p' and then over p to avoid the branch cut associated with p_0 at the first integration. The other term in the integrand of Eq. (41) is treated by integrating first over p and then over p'. Denoting this second term by I_2, we have

$$I_2(s,s') = \int_0^\infty dp\int_0^\infty dp'\frac{(p')^{-s'+2}p^{s-1}}{p_0'(p-p'e^{i\theta})(p-p'e^{-i\theta})} = \frac{e^{-is\pi}}{2i\sin\pi s}\int_0^\infty dp'\frac{(p')^{-s'+2}}{p_0'} \times\int_C dp\frac{p^{s-1}}{(p-p'e^{i\theta})(p-p'e^{-i\theta})}. \tag{43}$$

For I_1, we find

$$I_1(s,s') = \frac{\pi}{\sin\pi s'} \frac{\sin[(2-s')(\pi-\theta)]}{\sin\theta} \int_0^\infty \frac{dp}{p_0} p^{s-s'}. \quad (44)$$

If this integral is to converge, we see that C must be chosen so that

$$\mathrm{Re}(s) < \mathrm{Re}(s') < \mathrm{Re}(s+1). \quad (45)$$

The remaining integration can then be performed to give

$$I_1(s,s') = -\frac{\pi}{2\sin\pi s'} \frac{\sin[(2-s')(\pi-\theta)]}{\sin\theta} \left(\frac{m}{2}\right)^{s-s'} \times B\left(\frac{s'-s}{2}, s'-s+1\right). \quad (46)$$

The second term I_2 may be evaluated in the same way. The result is

$$I_2(s,s') = \frac{\pi}{2\sin\pi s} \frac{\sin[(1-s)(\pi-\theta)]}{\sin\theta} \left(\frac{m}{2}\right)^{s-s'} \times B\left(\frac{s'-s}{2}, s-s'+1\right). \quad (47)$$

In the last two equations, B denotes the beta function. Equations (46) and (47) may now be used to evaluate the expression for $\overline{K}_l(s,s')$:

$$\overline{K}_l(s,s') = \frac{Ze^2}{4}\left(\frac{m}{2}\right)^{s-s'} B\left(\frac{s'-s}{2}, s-s'+1\right) \times \int_0^\pi d\theta P_l(\cos\theta) \times \frac{\sin[(2-s')(\pi-\theta)]}{\sin\pi s'} - \frac{\sin[(1-s)(\pi-\theta)]}{\sin\pi s}. \quad (48)$$

We thus find, when l is even,

$$\overline{K}_l(s,s') = \frac{Ze^2}{4\pi}\left(\frac{m}{2}\right)^{s-s'} B\left(\frac{s'-s}{2}, s-s'+1\right) \times \sum_{k=0}^{l}\binom{l}{k} B(l-k+\tfrac{1}{2}, k+\tfrac{1}{2}) \times \left(\frac{\tan\pi s'/2}{l-2k+2-s'} - \frac{\cot\pi s/2}{l-2k+1-s}\right) \quad (49)$$

and when l is odd,

$$\overline{K}_l(s,s') = \frac{Ze^2}{4\pi}\left(\frac{m}{2}\right)^{s-s'} B\left(\frac{s'-s}{2}, s-s'+1\right) \times \sum_{k=0}^{l}\binom{l}{k} B(l-k+\tfrac{1}{2}, k+\tfrac{1}{2}) \times \left(\frac{\cot\pi s'/2}{l-2k+2-s'} - \frac{\tan\pi s/2}{l-2k+1-s}\right). \quad (50)$$

Poles of the beta functions relate to the conditions of Eq. (45). The reader may note that the even-odd alternative forms for $\overline{K}_l(s,s')$ have terms which produce poles for values of s or s' in the regions which are *not* excluded by the inequalities in Eq. (39). These poles are canceled when the entire series in k is included. For example, when $l = 1$, we find

$$\overline{K}_1(s,s') \propto \cot\pi s'/2[(3-s')^{-1} + (1-s')^{-1}] - \tan\pi s/2[(2-s)^{-1} - s^{-1}]. \quad (51)$$

The poles at $s' = 2$ and $s = 1$ from the cotangent and tangent are thus canceled by the zeroes in the brackets at these values. Hence, $\overline{K}_l(s,s')$ for $l = 1$ is analytic for $-1 < \mathrm{Re}s < 3$, and $0 < \mathrm{Re}s' < 4$.

We are now in a position to begin a determination of the contour of integration C. Firstly, the contour may be taken to run parallel to the imaginary axis from $-i\infty$ to $i\infty$. It is to be noted that the conditions for the existence of $\overline{K}_l(s,s')$ do not at first lead us to an integral equation of the usual type for $\psi_l(s)$, since we have derived an equation which relates $\psi_l(s)$ to values of $\psi_l(s')$, where the set of values of s is different from the set of s' values because of the requirement in Eq. (45). However, we may deform the s' contour by shifting it to the left so that it half encircles the pole contained in the beta function at $s' = s$ or we may increase $\mathrm{Re}(s)$ to $\mathrm{Re}(s')$, again taking the contour to half encircle the pole at $s' = s$. In the neighborhood of this pole,

$$\overline{K}_l(s,s') \cong \frac{R_l(s)}{s'-s}, \quad (52)$$

where $R_l(s)$ is given by

$$R_l(s) = \frac{Ze^2}{2\pi}\sum_{k=0}^{l}\binom{l}{k} B(l-k+\tfrac{1}{2}, k+\tfrac{1}{2}) \times \left[\frac{\tan\pi s/2}{l-2k+2-s} - \frac{\cot\pi s/2}{l-2k+1-s}\right], \quad (53)$$

when l is even, and by

$$R_l(s) = \frac{Ze^2}{2\pi}\sum_{k=0}^{l}\binom{l}{k} B(l-k+\tfrac{1}{2}, k+\tfrac{1}{2}) \times \left(\frac{\cot\pi s/2}{l-2k+2-s} - \frac{\tan\pi s/2}{l-2k+1-s}\right), \quad (54)$$

when l is odd.

We may now write the kernel $\overline{K}_l(s,s')$ as

$$\overline{K}_l(s,s') = \frac{R_l(s)}{s'-s} + \left(\overline{K}_l(s,s') - \frac{R_l(s)}{s'-s}\right), \quad (55)$$

where the kernel

$$\overline{K}_{1l}(s,s') \equiv \overline{K}_l(s,s') - R_l(s)/(s'-s) \quad (56)$$

is not singular at $s' = s$.

This leads to the singular integral equation

$$\psi_l(s) = \frac{R_l(s)\psi_l(s)}{2} + \frac{1}{2\pi i}\int_{s-i\infty}^{s+i\infty} \overline{K}_{1l}(s,s')\psi_l(s')ds' + \frac{P}{2\pi i}\int_{s-i\infty}^{s+i\infty} \frac{R_l(s)}{s'-s}\psi_l(s')ds', \quad (57)$$

where P denotes a principal value integral. There is also an "associate" integral equation to Eq. (57);

$$\psi_l^q(s) = \frac{R_l(s)\psi_l^q(s)}{2} + \frac{1}{2\pi i}\int_{s-i\infty}^{s+i\infty} ds'\, \bar{K}_{1l}^q(s,s')\psi_l^q(s') - \frac{P}{2\pi i}\int_{s-i\infty}^{s+i\infty} ds'\, \frac{R_l(s')}{s'-s}\psi_l^q(s'), \quad (58)$$

where

$$\bar{K}_{1l}^q(s,s') = \bar{K}_l(s's) + R_l(s')/(s'-s).$$

Our method of solution of Eq. (57) consists of first investigating solutions of a singular equation, the "dominant equation." We then derive a new integral equation for the problem which incorporates $\bar{K}_{1l}$. This new equation is of the Fredholm type; its development will be given later in this section.

We shall conclude this section with a qualitative discussion of the solutions of the Bakamjian–Thomas equation. For simplicity, we consider the case when $l = 0$. In this case, Eq. (53) then becomes (when we drop angular momentum subscripts)

$$R(s) = \frac{Ze^2}{2}\left(\frac{\tan\pi s/2}{2-s} - \frac{\cot\pi s/2}{1-s}\right). \quad (59)$$

If we write $f(s)$ for the term involving $\bar{K}_1$, we have

$$\left(1 - \frac{R(s)}{2}\right)\psi(s) - \frac{R(s)P}{2\pi i}\int_{s-i\infty}^{s+i\infty} ds'\, \frac{\psi(s')}{s'-s} = f(s). \quad (60)$$

This equation can be written as $K^0\psi = f$, where K^0 is defined to be the dominant part of the original kernel K.

An equation of this form was first treated by Carlemann and is extensively discussed in the books by Muskhelishvili[12] and Pogorzelski.[13] We follow the discussions given by these authors. First, we introduce the function

$$H(s) = \frac{1}{2\pi i}\int_C \frac{\psi(s')ds'}{s'-s}, \quad (61)$$

where the contour C goes from $-i\infty$ to $i\infty$. We can look at $H(s)$ as a single-valued function in the s plane, cut along C. If we denote the region to the left of the contour by S^+ and that to the right by S^-, we can obtain two functions $H^{\pm}(s)$, analytic in $S^{\pm}$, respectively, according to whether s lies in S^+ or S^-. These two functions can then be analytically continued beyond the cut C. Because of Cauchy's theorem, the contour C can be varied without affecting $H^+(s)$ or $H^-(s)$ unless a singularity in the integrand is encountered on C: i.e., $\psi(s')$ is singular, or C passes through s. If C is chosen to pass through s, we have

$$\begin{aligned}(H(s-\epsilon) + H(s+\epsilon))_{\epsilon\to 0} &\equiv H^+(s) - H^-(s) = \psi(s),\\ (H(s-\epsilon) + H(s+\epsilon))_{\epsilon\to 0} &\equiv H^+(s) + H^-(s)\\ &= \frac{P}{\pi i}\int \frac{\psi(s')ds'}{s'-s}.\end{aligned} \quad (62)$$

These relations reduce Eq. (60) to an algebraic equation[14]:

$$[1 - \tfrac{1}{2}R(s)][H^+(s) - H^-(s)] = f(s) + \tfrac{1}{2}R(s)[H^+(s) + H^-(s)]. \quad (63)$$

To solve this equation, we begin by considering the solution of the equation with $f \equiv 0$ and denote the solutions by $H_0^{\pm}$. We have

$$[1 - R(s)]H_0^+(s) = H_0^-(s) \quad (64)$$

or

$$H_0^+(s)/H_0^-(s) = 1/[1 - R(s)]. \quad (65)$$

Upon taking the logarithm of both sides, one finds

$$\ln H_0^+(s) - \ln H_0^-(s) = -\ln[1 - R(s)]. \quad (66)$$

If one now introduces

$$\ln H_0(s) = -\frac{1}{2\pi i}\int_C \frac{\ln[1 - R(s')]}{s'-s}, \quad (67)$$

this effects a solution of the discontinuity equation (64) for the homogeneous equation.[15] From Eq. (67) one sees that $H_0^{\pm}(s)$ are neither singular nor zero in the regions $S^{\pm}$, respectively.

A solution of the inhomogeneous problem is achieved by using Eq. (64) to replace $1 - R$ in Eq. (63). Thus

$$[H_0^-(s)/H_0^+(s)]H^+(s) = H^-(s) + f(s)$$

or

$$\frac{H^+(s)}{H_0^+(s)} - \frac{H^-(s)}{H_0^-(s)} = \frac{f(s)}{H_0^-(s)}. \quad (68)$$

If we now introduce

$$H(s) \equiv F(s)H_0(s), \quad (69)$$

we obtain

$$F^+(s) - F^-(s) = f(s)/H_0^-(s), \quad (70)$$

which can be formally solved by

$$F(s) = \frac{1}{2\pi i}\int_C \frac{f(s')ds'}{(s'-s)H_0^-(s)} \quad (71)$$

and we see that $F^{\pm}(s)$ are regular in the regions $S^{\pm}$, respectively. The solution of our equation for ψ is then obtained using Eqs. (62) and (69).

We now continue consideration of the choice of contour for our problem. For the $l = 0$ case, we have the conditions

$$0 < \mathrm{Re}(s) < 2, \quad \mathrm{Re}(s) < \mathrm{Re}(s') < \mathrm{Re}(s+1), \quad 1 < \mathrm{Re}(s') < 3. \quad (72)$$

At the various limiting values for s, s', there are singularities in $K_0(s, s')$, of which the pole at $s = s'$ has already been made explicit in the singular integral equation, Eq. (57). We note that $1 - R(s)$ can be given an infinite product representation[16] in the form

$$1 - R(s) = \prod_{n=-\infty}^{\infty}\left(\frac{s-\xi_n}{s-n}\right), \quad (73)$$

Since $R(s)$ has poles at all integers, $R(s) \to 1$ as $|\mathrm{Im}s| \to \infty$, and for each n there is an $s = \xi_n$ such that $R(\xi_n) = 1$.[17] Further, as $|n| \to \infty$, one finds

$$\xi_n - n \sim Ze^2/\pi n, \quad (74)$$

which guarantees that the infinite product converges. Finally, $R(s)$ is symmetric about $s = \frac{3}{2}$, so that

$$R(\tfrac{3}{2} + t) = R(\tfrac{3}{2} - t). \tag{75}$$

Thus the complex s plane shows a pattern of poles and zeroes as indicated in Fig. 2 (for $Ze^2 < 0$).

Let us now note the following facts: The function $H_0^+(s)$ is analytic and nonzero in S^+, while $H_0^-(s)$ in analytic and nonzero in S^-. All three contours C_0, C_1, and C_2 of Fig. 2 satisfy the conditions on s'.[18] We do not consider contours in which $\mathrm{Re}(s) < \frac{3}{2}$, because such contours can either be distorted so that $\mathrm{Re}(s) > \frac{3}{2}$ or there will be a singularity in $\psi(s)$ for $\mathrm{Re}\, s < \frac{3}{2}$. The former case is of no interest, while the latter one would lead to a wavefunction in momentum space which is not square integrable, and is therefore excluded. If we consider C_0, we find from Eq. (67) that

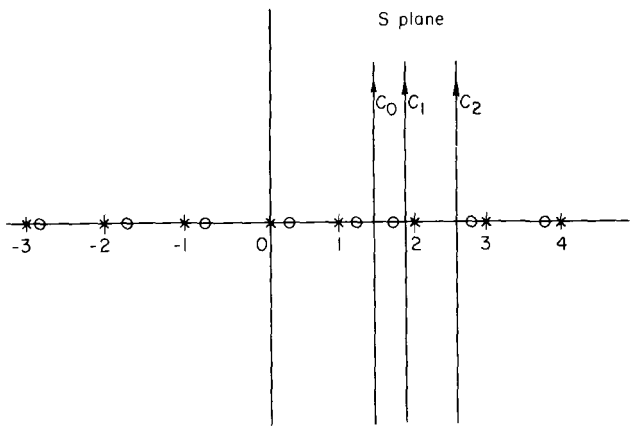


FIG. 2. Structure of $R(s)$; poles are represented by $\times$, zeroes by ○. At first sight C_0, C_1, C_2 represent possible contours of integration but only C_0 is really allowed.

$$H_0^{(0)+}(s) = \prod_{n=2}^{\infty} \left(\frac{s-n}{s-\xi_n}\right)$$
and
$$H_0^{(0)-}(s) = \prod_{n=-\infty}^{1} \left(\frac{s-\xi_n}{s-n}\right), \tag{76}$$

while for C_2

$$H_0^{(2)+}(s) = \prod_{n=3}^{\infty} \left(\frac{s-n}{s-\xi_n}\right)$$
and
$$H_0^{(2)-}(s) = \prod_{n=-\infty}^{2} \left(\frac{s-\xi_n}{s-n}\right). \tag{77}$$

On the other hand, for contour C_1 the integral in Eq. (67) is singular, since the phase of the logarithm does not go to zero as $\mathrm{Im}\, s \to \infty$, so $H_0(s)$ cannot be defined by Eq. (67). One could attempt to use

$$H_0^{(1)+}(s) = (s-2) \prod_{n=3}^{\infty} \left(\frac{s-n}{s-\xi_n}\right)$$
and
$$H_0^{(1)-}(s) = (s-\xi_2) \prod_{n=-\infty}^{1} \left(\frac{s-\xi_n}{s-n}\right), \tag{78}$$

since this separation of $1 - R(s)$ satisfies Eq. (65). In this case, if we consider the "solution" of the homogeneous equation for $\psi(s)$, we see that as $|\mathrm{Im}\, s| \to \infty$, $\psi(s) \to (\xi_2 - 2)$. But this asymptotic behavior is not allowed, since the principle value integral in Eq. (60) is not well defined; in fact, we will show that the contour C_1 is not acceptable. If $Ze^2 > 0$, the relative positions of the poles n and zeros ξ_n in $1 - R(s)$ are reversed. In this case, we obtain a valid solution of the homogeneous equation using C_1, in which

$$H_0^+(s) = \frac{1}{(s-\xi_2)} \prod_{n=3}^{\infty} \left(\frac{s-n}{s-\xi_n}\right) \tag{79}$$
and
$$H_0^-(s) = \frac{1}{(s-2)} \prod_{n=-\infty}^{1} \left(\frac{s-\xi_n}{s-n}\right). \tag{80}$$

Thus the solution for $\psi(s)$ will not be unique, because an arbitrary amount of the solution of the homogeneous equation can always be added to a particular solution. We will return to a further consideration of C_0, C_2 subsequently.

As was seen from Eq. (36), the behavior of $\psi(p)$ for $p \to \infty$ is determined by the properties of $\psi(s)$ in S^-. In this region, $H_0^-(s)$ and $F^-(s)$ are analytic, so it is convenient to express the solutions in terms of them. In S^- it is convenient to represent $H^+(s)$ as

$$H^+(s) = \frac{H_0^-(s)}{1 - R(s)} \left(F^-(s) + \frac{f(s)}{H_0^-(s)}\right). \tag{81}$$

Thus if $1 - R(s)$ vanishes, $\psi(s)$ will have a pole; i.e., at points $s = \xi_n$. By construction we also know that $f(s)$ has a pole at $s = n$, but here $R(s)$ also has a pole which cancels the singularity, so that $\psi(s)$ is regular at n. Thus the asymptotic behavior of $\psi(p)$ will be dominated by the smallest ξ_n, $\xi_n^{\min}$ in the S^- region; i.e.,

$$\psi_l(p) \underset{p\to\infty}{\sim} p^{-\xi_n^{\min}}.$$

We now come to the decisive part of our investigation, the complete solution of Eq. (57). This depends on the existence of solutions of our equations,[19] which in turn can be determined using the Vekua theory of singular integral equations,[20] which we will now briefly recapitulate. Vekua's theorem states that, under certain conditions, each singular integral equation of the form

$$K\phi \equiv A(s)\phi(s) + \frac{P}{\pi i} \int_C ds' \frac{N(s,s')f(s')}{s'-s} = f(s) \tag{82}$$

is equivalent to a Fredholm equation with a completely continuous kernel.[21] The conditions which must be imposed are the Hölder relations:

$$|A(s) - A(s')| < \text{const}\, |s - s'|^n,$$
$$|f(s) - f(s')| < \text{const}\, |s - s'|^n,$$
and
$$|K(s,s') - K(s'',s'')| < \text{const}\, [|s - s''|^n + |s' - s''|^n], \tag{83}$$

where $0 < n \leq 1$. Of central importance in the Vekua theory is the index κ,

$$\kappa \equiv \frac{1}{2\pi i} \Delta_C \ln\left(\frac{A(s) - B(s)}{A(s) + B(s)}\right) = \frac{1}{2\pi} \Delta_C \arg\left(\frac{A(s) - B(s)}{A(s) + B(s)}\right). \tag{84}$$

Here,

$$B(s) \equiv N(s, s) \tag{85}$$

and the notation Δ_C is meant to indicate the total change in phase of $(A - B)/(A + B)$ as we traverse the entire contour C. In our case, $A(s) = 1 - \frac{1}{2}R(s)$ and $B(s) = -\frac{1}{2}R(s)$, and the contours C_0 and C_2 give $\kappa = 0$, while C_1 gives $\kappa = \pm 1$ according to whether $Ze^2 \gtrless 0$. We have seen that only for $Ze^2 > 0$ and the contour C_1 is there a solution of the homogeneous equation $K^0\phi = 0$ in which $\phi(s) \to 0$ as $|\mathrm{Im}s| \to \infty$. In order to effect the reduction of the singular equation to Fredholm form, the "dominant" operation

$$K^0\phi = A(s)\phi(s) + \frac{B(s)P}{\pi i}\int_C ds' \frac{\phi(s')}{s' - s} \tag{86}$$

and its "associate" operation[22]

$$K^{0\prime}\phi = A(s)\phi(s) - \frac{P}{\pi i}\int_C ds' \frac{B(s')\phi(s')}{s' - s} \tag{87}$$

are introduced. If κ is the index of K^0, then κ is said to be the index of the original equation. Since the sign of the imaginary unit has been changed in Eq. (87), $-\kappa$ is the index of $K^{0\prime}$. A theory of Eq. (82) was first developed by Carlemann.[23] If the index of Eq. (82) is κ and

$$\kappa > 0, \tag{88}$$

there are κ linearly independent solutions of the homogeneous equation of the form

$$\Phi(s) = H_0(s)P_\kappa(s), \tag{89}$$

where $P_\kappa(s)$ is a polynomial in s of degree κ. On the other hand, for $\kappa \geq 0$, the associate operation in Eq. (87) has a $\kappa \leq 0$, and there are then no nonvanishing solutions of $K^{0\prime}\phi = 0$ which tend to zero at infinity.

It can be shown by use of the Poincaré–Bertrand transformation[24] that $K^{0\prime}$ is a "regularizing" operator for the kernel K; that is, the kernel $K^{0\prime}K(\equiv \int_C K^{0\prime}(s, s'')K(s'', s')ds'')$ is completely continuous, although K is not. Thus if $\kappa \geq 0$, solutions of the equation $K\phi = f$ can be sought via the regularized equation

$$K^{0\prime}K\phi = K^{0\prime}f, \tag{90}$$

for which the usual Fredholm theorems apply. Since $K^{0\prime}\psi = 0$ has no nontrivial solutions, no extraneous solutions are introduced.

If, on the other hand, κ is negative, one may define

$$\phi = K^{0\prime}\psi \tag{91}$$

and form the equation

$$KK^{0\prime}\psi = f, \tag{92}$$

which may be shown to have a completely continuous kernel. The solution of the original Eq. (82) is then obtained by quadrature from the solution of this equation.

The relevance of the above theorem to our work depends on the following theorems. We first note that for any kernel K and its adjoint, if we have solutions ϕ, ψ such that

$$K\phi = f \tag{93}$$

and

$$K'\psi = 0, \tag{94}$$

then the general relation

$$\int \psi K\phi ds ds' = \int \phi K'\psi ds ds' \tag{95}$$

requires that

$$\int \psi f ds = 0. \tag{96}$$

This is, of course, just the generalization of the familiar property which is known from the theory of Fredholm operators; that is, a necessary and sufficient condition for the solution of an inhomogeneous Fredholm equation is that the driving term be orthogonal to the eigenfunctions of the transposed operator (or Hermitian conjugate operator if orthogonality includes complex conjugation). We now remark that, in analogy to the Fredholm case, the condition (96) is also sufficient to guarantee a solution of Eq. (93): First, suppose that κ is positive or zero. We consider the solution ω of the Fredholm equation

$$(K^{0\prime}K)'\omega = 0 \tag{97}$$

or, equivalently,

$$K'K^0\omega = 0. \tag{98}$$

Since the solutions of Eq. (98) always satisfy Eq. (94), $K^0\omega$ must be a linear combination of the ψ. According to the Fredholm theory, however, a necessary and sufficient condition that there be a solution of an inhomogeneous Fredholm equation is that the inhomogeneous term be orthogonal to all solutions ω of the homogeneous equation, with transposed kernel. Thus a sufficient condition for the solution of Eq. (90) is

$$\int \omega K^{0\prime} f ds = \int f K^0 \omega ds = \int f \sum_i a_i \psi_i ds. \tag{99}$$

Thus, if Eq. (96) holds, there is a solution of Eq. (90) and hence of Eq. (93), and sufficiency is proved.

If κ is negative, we introduce the solutions γ of the transposed Fredholm equation

$$(KK^{0\prime})'\gamma = 0 \tag{100}$$

or, equivalently,

$$K^0K'\gamma = 0. \tag{101}$$

The Fredholm theory shows here that if

$$\int f\gamma ds = 0, \tag{102}$$

one may find a solution of Eq. (92). This allows one to construct ϕ by quadrature [Eq. (91)]. Since the dominant equation for κ negative has no nontrivial solutions, those of Eq. (101) must be linear combinations of those of the homogeneous associate equation (89). Thus the condition of Eq. (96) is sufficient in this case also.

A further theorem has been proved by Vekua: The difference between the number k of linearly independent solutions of the singular equation $K\phi = 0$ and the corresponding number k' for $K'\psi = 0$ is equal to the index κ of the first equation. This can be shown as follows: We assume that $\kappa \geq 0$ without loss of generality, since if $\kappa < 0$ the roles of K and K' can simply be interchanged. Then we know that the equation

$$K\phi = 0 \tag{103}$$

is completely equivalent to

$$K^0{}'K\phi = 0, \tag{104}$$

and therefore the latter also has k linearly independent solutions. From the Fredholm theory we know then that

$$K'K^0\psi = 0 \tag{105}$$

has k linearly independent solutions as well. Since $K^0\psi = 0$ has κ linearly independent solutions, it follows that $k' = k - \kappa$.

Let us now apply the foregoing analysis to our equation. As has been seen, the choice of contour C affects the resulting κ. We note that if κ is positive or zero, except for certain eigenvalues there are no nonzero solutions of the homogeneous adjoint equation. Hence there are no restrictions on the function f as indicated by Eq. (4). If κ is negative, however, f cannot be arbitrary. Thus the contour must be chosen so that κ is positive or zero. Thus, if $Ze^2 > 0$, the path C_1 must be excluded since there will not generally be a solution of the equation. On the other hand, if $Ze^2 < 0$ for physical reasons, C_1 again is excluded since the solutions in this case would not be unique. Thus we are left with the possible contours C_0 and C_2.

Again let us consider $l = 0$. The generalization to arbitrary l is simple. We have seen that the behavior of $\psi(p)$ as $p \to 0$ is determined by the highest singularity in S^+. Thus it is convenient to express the solution of Eq. (60), $\psi(s)$, in terms of H_0^+ and F^+:

$$\psi(s) = R(s)H_0^+(s)F^+(s) + f(s). \tag{106}$$

The singularities of $\psi(s)$ in S^+ are then found either in $R(s)$ or $f(s)$, or in both. Thus there may be poles in $\psi(s)$ at all of the integers to the left of C. If Eq. (37) is used to continue $\psi(s)$, however, it is seen that only the singularity in s associated with $K(s, s')$ produces a singularity in $\psi(s)$, and hence if C_0 is chosen, there will be a pole in $\psi(s)$ at $s = 0$ (and at the negative even integers). On the other hand, if C_2 is chosen, it is convenient to first let $\mathrm{Re}(s) \to \mathrm{Re}(s')$ on C_2, since we have the analytic continuation explicity of $K(s, s')$ for $\mathrm{Re}s \geq 2$, and then obtain a solution of the equation on C_2, and finally use Eq. (37) to continue back to $s = 2$. We thus find a pole at $s = 2$, and

$$\psi(p) \underset{p\to 0}{\sim} p^{-2}. \tag{107}$$

This behavior is not acceptable, however, since the wavefunction would not be normalizable, and so we exclude the path C_2 from further consideration, and we have a unique solution to the singular equation, Eq. (12). Since we are thus restricted to the path C_0 we may ask whether there is some especially appropriate path. It is shown in Appendix A that if $\mathrm{Re}(s) = \mathrm{Re}(s') = \frac{3}{2}$, the kernel of the integral equation satisfies a hermiticity condition, and so this choice seems to be called for.

We shall close this section by noting that there is a maximum value for $-Ze^2$ for which a unique solution of Eq. (12) is possible. Again for $l = 0$, it is easily seen that $R(s)$ is real for $\mathrm{Im}s = 0$, and for $\mathrm{Re}s = \frac{3}{2}$. In the latter case,

$$R(s) = -\frac{Ze^2}{2}\left(\frac{1 - i\tanh(\pi/2)s_I}{1 + i\tanh(\pi/2)s_I}\cdot\frac{1}{\frac{1}{2} - is_I} + \text{c.c.}\right), \tag{108}$$

where $s = \frac{3}{2} + is_I$. As $|s_I| \to \infty$,

$$R(s) \sim -Ze^2/s_I. \tag{109}$$

On the path along which $R(s)$ is real going from $s = 2$ to either $s = \frac{3}{2} \pm i\infty$, $|R(s)|$ is a monotonic decreasing function going from ∞ to 0, and if $Ze^2 < 0$, there will be a point ξ_2 on the path at which[25]

$$R(\xi_2) = 1. \tag{110}$$

This point ξ_2 will only have $\mathrm{Im}(s) = 0$ if

$$R(\tfrac{3}{2}) < 1, \tag{111}$$

that is,

$$|Ze^2| < \tfrac{1}{2}. \tag{112}$$

If this condition is not satisfied, the points ξ_1 and ξ_2 become complex conjugate pairs and the contour C and solution $\psi(s)$ are not unique. The situation is completely analogous to that with the Dirac equation, for which there are too many acceptable solutions also if Ze^2 is too large.[26] The problem raised here is only of mathematical interest, however, since a large Z nucleus would necessitate a form factor to describe its spatial extent and the potential for large p would be cut off, in contradistinction to the point particles dealt with here. Thus we will not pursue this case further.

The Dirac equation may also be dealt with using the Mellin transformation technique. In that case it is found that $R(s)$ is a quadratic function of s and there are only two possible ξ_i. A brief account of the treatment of the Dirac equation is given in Appendix B.

V. MOMENTUM SPACE INTERPRETATION OF THE K_0 KERNEL

In the preceding section, we have provided an analysis of the non-Fredholm Bakamjian–Thomas equation which leads to a unique solution. Although this provides a mathematically satisfactory solution, its significance is probably somewhat obscure. In the present section, we will provide an alternative momentum space treatment which is closely related to the Carlemann approach in Mellin space, but which gives direct insight into the above results.

Since the non-Fredholm behavior of the equation is associated with high momenta, one might try to sepa-

rate the kernel into an asymptotic part and a remainder. Thus we write Eq. (12) as

$$\psi_l(p) = -\frac{Ze^2}{2\pi}\int_0^\infty dp' \frac{(p+p')}{p^2} Q_l\left(\frac{p^2+p'^2}{2pp'}\right) U(p')\psi_l(p') + \frac{Ze^2}{2\pi}\int_0^\infty p'\,\frac{dp'}{p}\left(\frac{(p_0+p_0')}{p_0'(P_0-p_0)} + U(p')\frac{(p+p')}{pp'}\right) Q_l\left(\frac{p^2+p'^2}{2pp'}\right)\psi_l(p'), \tag{113}$$

where $U(p)$ is zero for $p < 1$, and is one for $p \geq 1$. Let us consider the kernel in the first integral to be $K_0(p,p')$, and the balance of the right-hand side as if it were an inhomogeneous term $f(p)$. We thus look for solutions of the equation

$$\psi(p) = \int dp' K_0(p,p')\psi(p') + f(p). \tag{114}$$

The step function U must be introduced because otherwise there would be no solution of Eq. (114) because of the behavior as $p, p' \to 0$.

Clearly the kernel K_0 is of such form that we can write

$$\psi_l(p) = \int_1^\infty \frac{dp'}{p'} F\left(\frac{p}{p'}\right)\psi_l(p') + f(p). \tag{115}$$

This equation can be solved in two ways: If a new variable $x \equiv \ln p$ is introduced, Eq. (115) is converted into a Wiener–Hopf equation which can be solved by known techniques.[27] On the other hand, if a Mellin transformation is carried out as in Sec. IV, we get

$$\psi_l(s) = \frac{1}{2\pi i}\int_C ds' \frac{F(s)\psi_l(s')}{s'-s} + f(s), \tag{116}$$

where $\mathrm{Re}(s'-s) > 0$. This equation can be brought to the form of Eq. (57), and is easily seen to be identical to it. Thus the Carlemann solution of this equation corresponds to finding a solution of the inhomogeneous Eq. (114) to remove the non-Fredholm term $K_0(p,p')$.

In obtaining Eq. (113), we essentially took the asymptotic form of the kernel for $p, p' \to \infty$, and then multiplied the kernel unsymmetrically by $U(p')$. If, on the other hand, we had multiplied by $U(p)$, the only change which would occur would be that in Eq. (116)—$F(s')$ would appear instead of $F(s)$.

VI. CONCLUSION

In this paper we have given arguments which lead to an approximate wave equation for spin-zero particles. This equation has been studied for the case of an interaction which is the time component of a vector field. A simplification was afforded by assuming one of the particles to be infinitely massive. Because of the nature of the interaction the Schrödinger integral equation is singular so that the Fredholm theory does not immediately apply. We have given a simple discussion of the nature of the solutions to be expected for our equation and have then gone on to rigorously show that a unique solution may be achieved if the potential is repulsive or the coupling constant is not too large. We have also gone beyond the considerations of this paper to construct explicit numerical solutions for both bound state and scattering problems for the case when $l = 0$. Because of the length and complexity of this paper and the special techniques which are required to effect a numerical solution, we will report these results elsewhere.

APPENDIX A: HERMITICITY OF THE MELLIN TRANSFORMED EQUATION

In this section we shall show that the kernel in the Mellin transformed integral equation is Hermitian if the contour C is taken to lie along $\mathrm{Re}(s') = \frac{3}{2}$. For this development, we divide the kernel into two parts:

$$K_l(p,p') = \overline{K}_l(p,p') + K_l^R(p,p'), \tag{A1}$$

where [see Eq. (40)]

$$\overline{K}_l(p,p') = -\frac{Ze^2}{2\pi}\frac{p'}{p}\frac{(p_0+p_0')}{p_0p_0'} Q_l\left(\frac{p^2+p'^2}{2pp'}\right) \tag{A2}$$

and

$$K_l^R(p,p') = -\left(1 - \frac{(p_0p_0')^{1/2}}{(P_0-p_0)^{1/2}(P_0-p_0')^{1/2}}\right)\overline{K}_l(p,p'). \tag{A3}$$

This division separates K_l into a part which has a Mellin transform that is regular at $s = s'$, K_l^R, and a singular part $\overline{K}_l$ for which we already have the transform explicity. The singularity in $\overline{K}_l(s,s')$ for $s \to s'$ arises from the asymptotic behavior of $\overline{K}_l(p,p')$ for large p, p'. If we set $p' = \alpha p$ in Eq. (41), we find

$$\overline{K}_l(s,s') = -\frac{Ze^2}{2\pi}\int_0^\pi d\theta\, \sin\theta\, P_l(\cos\theta) \times \int_0^\infty d\alpha \frac{\alpha^{2-s'}}{(\alpha - e^{i\theta})(\alpha - e^{-i\theta})} \times \int_0^\infty dp\, p^{s-s'}[(p^2+m^2)^{-1/2} + (\alpha^2p^2+m^2)^{-1/2}]. \tag{A4}$$

Thus, the integral over p diverges as $p \to \infty$, unless $\mathrm{Re}(s-s') < 0$. If we use the same approach to $K_l^R(s,s')$, however, as $p \to \infty$ the terms involving p_0 and p_0' are now of order p^{-2}, and hence $K_l^R(s,s')$ is regular at $s = s'$.

To investigate the hermiticity condition for the kernel, we convert the transformed integral equation to one in real variables. Thus we set

$$s = s_R + is_I \quad \text{and} \quad s' = s_R + is_I', \tag{A5}$$

so that the integral equation becomes (we do not explicitly exhibit the dependence on s_R, which now becomes a parameter in the equation)

$$\psi_l(s_I) = \tfrac{1}{2}R_l(is_I)\psi_l(s_I) + \frac{P}{2\pi i}\int_{-\infty}^\infty ds_I' \frac{R_l(is_I')\psi_l(s_I')}{s_I'-s_I} + \frac{1}{2\pi}\int_{-\infty}^\infty ds_I'[K_{1l}(is_I,is_I') + K_{1l}{}^R(is_I,is_I')]\psi_l(s_I') \tag{A6}$$

The K^R_{1l} part of the kernel considered as a function of s_I, s'_I can easily be shown to be Hermitian for $s_R = \frac{3}{2}$. We begin by noting that the transform can be written as

$$K^R_{1l}(s_R + is_I, s'_R + is'_I) = \int_0^\infty dp \int_0^\infty dp' \times p^{s_R+is_I-1}(p')^{-s_R-is'_I+2}F_l(pp'), \quad \text{(A7)}$$

where

$$F_l(p,p') = F_l(p',p)^*. \quad \text{(A8)}$$

(We only consider bound states, so $P_0 < m$.) Then we find

$$K^R_1(s_R + is'_I, s_R + is_I)^* = \int_0^\infty dp \int_0^\infty dp' p^{2-s_R+is_I}(p')^{s_R-is'_I-1}F_l(p,p'), \quad \text{(A9)}$$

in which Eqs. (A7) and (A8) have been used, and the dummy variables p, p' have been interchanged. Thus, if $s_R = \frac{3}{2}$, we see that[28] as a function of the real variables s_I, s'_I,

$$K^R_{1l}(\tfrac{3}{2} + is_I, \tfrac{3}{2} + is'_I) = K^R_{1l}(\tfrac{3}{2} + is'_I, \tfrac{3}{2} + is_I)^*, \quad \text{(A10)}$$

so that K^R_{1l} is Hermitian. We must now look more closely at the singular kernel. We consider the case in which l is even. Choosing $s_R = \frac{3}{2}$, we then find from Eq. (49),

$$\begin{aligned}\bar{K}_l(s_I, s'_I) = {} & \frac{Ze^2}{4\pi}\left(\frac{m}{2}\right)^{i(s_I-s'_I)} B\left(\frac{i(s_I-s'_I)}{2}, i(s_I-s'_I)+1\right) \\ & \times \sum_{k=0}^{l}\binom{l}{k} B(l-k+\tfrac{1}{2}, k+\tfrac{1}{2}) \\ & \times\left(\frac{1-i\tanh(\pi s'_I/2)}{1+i\tanh(\pi s'_I/2)}\cdot\frac{1}{l-2k+\frac{1}{2}-is'_I}\right. \\ & \left.+\frac{1+i\tanh(\pi s_I/2)}{1-i\tanh(\pi s_I/2)}\cdot\frac{1}{l-2k+\frac{1}{2}+is_I}\right),\end{aligned} \quad \text{(A11)}$$

in which the second term in the bracket is obtained from the $\cot \pi s/2$ term in Eq. (49) by interchanging k and $l-k$ in the summation over k. From Eq. (116) one can easily see that

$$\bar{K}_l(s_I, s'_I) = \bar{K}_l(s'_I, s_I)^* \quad \text{(A12)}$$

and that $R_l(s_I)$, which is the residue of $\bar{K}_l$ in the pole of the beta function at $s' = s$ is real. Thus we have a Hermitian kernel for the equation. It may be mentioned here that the Fredholm kernels obtained in the Vekua theory do not satisfy the hermiticity requirement. This occurs because of the lack of symmetry in the choice of K_0, for example, not from the singularity at $s' = s$.

APPENDIX B: MOMENTUM SPACE ANALYSIS OF THE DIRAC EQUATION

In this appendix we apply the Mellin transformation technique to the solution of the familiar Dirac equation for a spin-$\frac{1}{2}$ particle in a Coulomb field. The conventional discussion[29] is based on a study of the indicial equation of the differential equation for this problem in coordinate space and involves boundary conditions at the origin.

Let us now consider the nature of the solutions of this equation in momentum space:

$$(\boldsymbol{\alpha}\cdot\mathbf{p} + \beta m)\psi(\mathbf{p}) + \frac{Ze^2}{2\pi^2}\int \frac{d^3\mathbf{p}'}{(\mathbf{p}-\mathbf{p}')^2}\psi(\mathbf{p}') = P_0\psi(\mathbf{p}). \quad \text{(B1)}$$

The usual solutions of the Dirac equation involve the operator

$$k = \beta(\boldsymbol{\sigma}\cdot\mathbf{L} + 1),$$

which has the eigenvalues $\pm(j+\frac{1}{2})$, since it commutes with the Dirac Hamiltonian. If we write

$$\boldsymbol{\alpha} = \rho_1\boldsymbol{\sigma},$$

Eq. (B1) becomes

$$(\rho_1\boldsymbol{\sigma}\cdot\mathbf{p} + \beta m)\psi(\mathbf{p}) + \frac{Ze^2}{2\pi^2}\int \frac{d^3\mathbf{p}'}{(\mathbf{p}-\mathbf{p}')^2}\psi(\mathbf{p}') = P_0\psi(\mathbf{p}). \quad \text{(B2)}$$

We choose

$$\rho_1 = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \quad \beta = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \quad \boldsymbol{\sigma}\cdot\mathbf{p} = \begin{pmatrix}\boldsymbol{\sigma}\cdot\mathbf{p} & 0\\ 0 & \boldsymbol{\sigma}\cdot\mathbf{p}\end{pmatrix}, \quad \text{(B3)}$$

where each of the elements in a matrix is a 2×2 matrix and the corresponding σ's are Pauli spin matrices.

An angular momentum decomposition can be achieved by setting

$$\psi^m_j = \begin{pmatrix} y^m_{j,j-1/2}(\hat{p})F_j(p) \\ y^m_{j,j+1/2}(\hat{p})G_j(p)\end{pmatrix}, \quad \text{(B4)}$$

where y^m_{jl} is an eigenfunction of total and orbital angular momentum of eigenvalues j and l, respectively, and of $j_z = m$, and $\hat{p}$ is a unit vector in the direction of $\mathbf{p}$. In the first place this function is an eigenfunction of k, since[30]

$$(\boldsymbol{\sigma}\cdot\mathbf{L} + 1)y^m_{j,j\pm1/2} = \mp(j+\tfrac{1}{2})y^m_{j,j\pm1/2}. \quad \text{(B5)}$$

Now consider the effect of the operator $\boldsymbol{\sigma}\cdot\mathbf{p}$ on the state $y^m_{j,l}$. We set

$$Y \equiv \boldsymbol{\sigma}\cdot\mathbf{p}\, y^m_{j,l}. \quad \text{(B6)}$$

To characterize this state we note that

$$\begin{aligned}(J^2, \boldsymbol{\sigma}\cdot\mathbf{p}) &= 0,\\ (J_3, \boldsymbol{\sigma}\cdot\mathbf{p}) &= 0,\\ (L^2, \boldsymbol{\sigma}\cdot\mathbf{p}) &= 2\boldsymbol{\sigma}\cdot\mathbf{p} + 2\boldsymbol{\sigma}\cdot\mathbf{p}\,\boldsymbol{\sigma}\cdot\mathbf{L}.\end{aligned} \quad \text{(B7)}$$

Since

$$\boldsymbol{\sigma}\cdot\mathbf{L} = J^2 - L^2 - \tfrac{3}{4}, \quad \text{(B8)}$$

we find that

$$L^2Y = [2j(j+1) - l(l+1) + \tfrac{1}{2}]Y. \tag{B9}$$

If

$$l = j - \tfrac{1}{2},$$

the new eigenvalue is $j + \frac{1}{2}$, and when

$$l = j + \tfrac{1}{2},$$

the new one is

$$l = j - \tfrac{1}{2}.$$

Choosing the conventional Clebsch–Gordon coefficients, we may write

$$y^m_{jj+1/2} = \frac{\boldsymbol{\sigma}\cdot\mathbf{p}}{p}\, y^m_{jj-1/2}$$

and (B10)

$$y^m_{jj-1/2} = \frac{\boldsymbol{\sigma}\cdot\mathbf{p}}{p}\, y^m_{jj+1/2}.$$

If we now substitute Eq. (B4) into the Dirac equation (B2), we have, denoting a unit vector by $\hat{p}$,

$$pG_j(p)y^m_{jj-1/2}(\hat{p}) + mF_j(p)y^m_{jj-1/2}(\hat{p}) + \frac{Ze^2}{2\pi^2}\int \frac{d^3\mathbf{p}'}{(\mathbf{p}-\mathbf{p}')^2} F_j(p')y^m_{jj-1/2}(\hat{p}') = P_0F_j(p)y^m_{jj-1/2}(\hat{p}) \tag{B11}$$

and

$$pF_j(p)y^m_{jj+1/2}(\hat{p}) - mG_j(p)y^m_{jj+1/2}(\hat{p}) + \frac{Ze^2}{2\pi^2}\int \frac{d^3\mathbf{p}'}{(\mathbf{p}-\mathbf{p}')^2} G_j(p')y^m_{jj+1/2}(\hat{p}') = P_0G_j(p)y^m_{jj+1/2}(\hat{p}).$$

To eliminate the angular functions, we multiply by $y^m_{jj-1/2}(\hat{p})^*$ and integrate over solid angle. We use the relations

$$\frac{2l+1}{4\pi} P_l(\hat{p}'\cdot\hat{p}) = \sum_{jm} y^m_{jl}(\hat{p}')^* y^m_{jl}(\hat{p})$$

and (B12)

$$(z-t)^{-1} = \sum_{l=0}^{\infty} (2l+1)Q_l(z)P_l(t)$$

and the normalization condition. The interaction term in the first member of Eq. (B11) is

$$\int d\Omega_p y^m_{jj-1/2}(p)^* \frac{d^3\mathbf{p}'}{(\mathbf{p}-\mathbf{p}')^2} y^m_{jj-1/2}(\hat{p}')F_j(p') = \int_0^\infty \frac{p'}{p}\, dp' d\Omega_p d\Omega_{p'} y^m_{jj-1/2}(\hat{p})^* \sum_{j'l'm'} Q_{l'}\left(\frac{p^2+p'^2}{2pp'}\right) \times y^{m'}_{j'l'}(\hat{p})y^{m'}_{j'l'}(\hat{p}')^* y^m_{jj-1/2}(\hat{p}')F_j(p') = 2\pi \int_0^\infty \frac{p'}{p}\, dp' Q_l\left(\frac{p^2+p'^2}{2pp'}\right)F_j(p'). \tag{B13}$$

We thus obtain the equations

$$pG_j(p) + mF_j(p) + \frac{Ze^2}{\pi}\int_0^\infty \frac{p'}{p}\, dp' Q_{j-1/2}\left(\frac{p^2+p'^2}{2pp'}\right)F_j(p') = P_0F_j(p)$$

and (B14)

$$pF_j(p) - mG_j(p) + \frac{Ze^2}{\pi}\int_0^\infty \frac{p'}{p}\, dp' Q_{j+1/2}\left(\frac{p^2+p'^2}{2pp'}\right)G_j(p') = P_0G_j(p).$$

Equation (B14) may be rewritten as

$$F_j(p) = \frac{Ze^2}{\pi}\int_0^\infty \frac{p'dp'}{p(P_0^2-p_0^2)} \times [(P_0+m)Q_{j-1/2}F_j + pQ_{j+1/2}G_j]$$

and (B15)

$$G_j(p) = \frac{Ze^2}{\pi}\int_0^\infty \frac{p'dp'}{p(P_0^2-p_0^2)} \times [pQ_{j-1/2}F_j + (P_0-m)Q_{j+1/2}].$$

The arguments of the Legendre function have been suppressed.

We now make a Mellin transformation on Eq. (B15):

$$F_j(s) = \int_0^\infty p^{s-1}F_j(p)$$

and (B16)

$$G_j(s) = \int_0^\infty p^{s-1}G_j(p).$$

For convenience we use the same symbols for the functions and their transforms. In matrix notation this leads to the integral equation

$$\psi(s) = \frac{1}{2\pi i}\int_C K(s,s')ds', \tag{B17}$$

where

$$K^j(s,s') = \frac{Ze^2}{\pi}\int_0^\infty dp \int_0^\infty dp' \frac{p^{s-2}(p')^{-s'+1}}{(P_0^2-p_0^2)} \times \begin{vmatrix} (P_0+m)Q_{j-1/2} & pQ_{j+1/2} \\ pQ_{j-1/2} & (P_0-m)Q_{j+1/2} \end{vmatrix}. \tag{B18}$$

The integrals may be evaluated by the methods used in the foregoing paper. If K is conventionally labeled according to the scheme

$$K^j = \begin{vmatrix} K^j_{11} & K^j_{12} \\ K^j_{21} & K^j_{22} \end{vmatrix}, \tag{B19}$$

we find the following: If $-1 < \mathrm{Re}(s-s') < 1$ and $-j+\frac{3}{2} < \mathrm{Re}s' < j+\frac{5}{2}$, we find

$$K^j_{11}(s,s') = \frac{Ze^2}{2}\frac{(P_0+m)\kappa^{s-s'-1}}{\cos\frac{1}{2}\pi(s-s')} C_j(s')S_j(s'), \tag{B20}$$

where we set

$$S_j(s') \equiv \sum_{k=0}^{j-1/2} \binom{j-\frac{1}{2}}{k} \frac{B(j-k, k+\frac{1}{2})}{(j-2k+\frac{3}{2}-s')}, \tag{B21}$$

$$\kappa = (m^2 - P_0^2)^{1/2}$$

and

$$C_j(s') = \frac{(-1)^{j-1/2}[1-(-1)^{j-1/2}\cos\pi s']}{\sin\pi s'}. \tag{B22}$$

Thus, if $j-\frac{1}{2}$ is even, $C_j(s') = \tan\pi s'/2$, and if $j-\frac{1}{2}$ is odd, $C_j(s') = -\cot\pi s'/2$. The other elements of K are easily obtained from $K_{11}^j(s, s')$. Thus

$$K_{12}^j(s, s') = K_{11}^{j+1}(s+1, s')/(P_0+m),$$
$$K_{21}^j(s, s') = K_{11}^j(s+1, s')/(P_0+m),$$
and
$$K_{22}^j(s, s') = \left(\frac{P_0-m}{P_0+m}\right)K_{11}^{j+1}(s, s'). \tag{B23}$$

It is seen that the integral equation is of the singular type since it has a pole at $s = s'$.

In analogy to the $B-T$ equation, the integral equation can be written

$$1 - \frac{R(s)}{2}\psi(s) = \frac{P}{2\pi i}\int_C \frac{R(s)}{s'-s}\psi(s')ds' + \frac{1}{2\pi i}\int_C K_1(s, s')\psi(s')ds', \tag{B24}$$

where K_1 is a regular 2×2 matrix and $R(s)$ is the residue of K at $s' = s$. A singularity in $\psi(s)$ will now occur if the matrix $[1-R(s)]$ is singular; i.e., it has a zero determinant. From Eqs. (B20) and (B23), we find that only K_{12} and K_{21} have poles at $s = s'$, so that

$$R_{12}(s) = \frac{Ze^2}{\pi}C_{j+1}(s)\sum_{k=0}^{j+1/2}\binom{j+\frac{1}{2}}{k}\frac{B(j-k+1, k+\frac{1}{2})}{(j-2k+\frac{5}{2}-s)} \tag{B25}$$

and similarly for $R_{21}(s)$, where $j+1 \to j$. The condition that $\det(1-R) = 0$ is thus

$$1 = (Ze^2/\pi)^2 C_j(s)C_{j+1}(s)S_j(s)S_{j+1}(s) \tag{B26}$$
$$= -(Ze^2/\pi)^2 S_j(s)S_{j+1}(s).$$

We see that $S_j(s)$ has poles at $s = j+\frac{3}{2}, j+\frac{1}{2}, \ldots, -j+\frac{5}{2}$, and hence S_jS_{j+1} can be expressed as a sum of poles times residues, which will now be evaluated.

We first consider the residue of the poles at $s = j+\frac{5}{2}-2k$, where k is an integer. For this purpose we note that[31]

$$S_j(j+\tfrac{5}{2}-2k) = (-1)^{j-1/2} \times \pi[1-(-1)^{j-1/2}\cos\pi(j+\tfrac{5}{2}-2k)]^{-1} \times \int_0^\pi P_{j-1/2}(\cos\theta)\sin[(\pi-\theta)(2k-j-\tfrac{1}{2})]d\theta. \tag{B27}$$

The factor in the square brackets is just 2, and the integral can be written as

$$(-1)^{j+1/2}\int_0^\pi P_{j-1/2}(\cos\theta)\sin(j-2k+\tfrac{1}{2})\theta d\theta. \tag{B28}$$

Further,

$$\sin(j-2k+\tfrac{1}{2})\theta = \sin\theta \sum_{m=0}^{[(j-2k+1/2)/2]} a_m P_{j-1/2-2m}(\cos\theta),$$

so the integral vanishes unless $k = 0$ or $k = j+\frac{1}{2}$. Similarly, one gets zero for all of the residues associated with poles in C_j. Thus the product of $S_j(s)$ $S_{j+1}(s)$ only has poles at $s = j+\frac{5}{2}$ and $s = \frac{3}{2}-j$. At $s = j+\frac{5}{2}$, we find

$$S_j(j+\tfrac{5}{2}) = -\tfrac{1}{2}\pi\Gamma(j+\tfrac{1}{2})\Gamma(\tfrac{1}{2})/\Gamma(j+1), \tag{B29}$$

and at $s = \frac{3}{2}-j$, the result is the same except for a change in sign. We finally obtain

$$1 = \frac{(Ze^2)^2}{(2j+1)}\left(\frac{1}{j+\frac{5}{2}-s} + \frac{1}{j-\frac{3}{2}+s}\right). \tag{B30}$$

From this one easily finds the singular values in s:

$$s_{1,2} = 2 \pm [(j+\tfrac{1}{2})^2 - (Ze^2)^2]^{1/2}. \tag{B31}$$

These are the analog of the well-known result in coordinate space for the indicial equation.

---

* This work was supported by the U.S. Atomic Energy Commission.

1 B. Bakamjian and L. H. Thomas, Phys. Rev. **92**, 1300 (1953). We shall treat only the Coulomb part of the electromagnetic interaction, so that strictly speaking our equation will be an approximate Bakamjian–Thomas equation.

2 R. N. Stuart, Ph.D. thesis (University of California, Berkeley, 1952) (unpublished).

3 H. Feshbach and F. Villars, Rev. Mod. Phys. **30**, 24 (1958).

4 C. Zemach, *Selected Topics in Solid State and Theoretical Physics* (Gordon and Breach, New York, 1968), p. 427.

5 J. Schwinger, Proc. Natl. Acad. Sci. (U.S.) **37**, 452 (1951); M. Gell-Mann and F. Low, Phys. Rev. **84**, 350 (1951); E. E. Salpeter and H. A. Bethe, *ibid.*, 1232 (1951).

6 M. L. Goldberger and K. M. Watson, *Collision Theory* (Wiley, New York, 1964), p. 21; J. Boguta and H. W. Wyld, Jr., Phys. Rev. **164**, 1996 (1967); David Avison, *ibid.* **154**, 1583 (1967); N. D. Son and J. Sucher, *ibid.* **153**, 1496 (1966); A. M. Brett and J. A. Oklowski, Nuovo Cimento **58**, 824 (1968).

7 For simplicity, we shall deal with Eq. (11); the asymptotic properties of Eq. (10) are not affected by the additional factor $(q_0+q_0')/q_0'$.

8 See W. Hunziker, Helv. Phys. Acta **34**, 593 (1961); A. Grossman and T. Wu, J. Math. Phys. **2**, 710 (1961); K. Meetz, *ibid.* **3**, 690 (1962); or M. Scadron, S. Weinberg, and J. Wright, Phys. Rev. **135**, B202 (1964).

9 In carrying out this integration we use the integral representation

$$P_l(\cos\theta) = \pi^{-1}\int_0^\pi (\cos\theta + i\sin\theta\cos t)^l dt$$

[see *Higher Transcendental Functions, Bateman Manuscript Project* (McGraw-Hill, New York, 1953), Vol. I, Eq. (3.7.23)], and note that

$$\cos\theta + i\sin\theta\cos t = \cos^2\tfrac{1}{2}t\cdot e^{i\theta} + \sin^2\tfrac{1}{2}t\cdot e^{-i\theta}.$$

10 E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals* (Oxford U. P., London, 1962), p. 60. It is also possible to make a Fourier transformation instead of the Mellin transformation. In this case one is led to a study of a Wiener–Hopf equation instead of Eq. (57).

11 In subsequent equations we will use the same symbol for a function as for its Mellin transform. The two will be distinguished by their arguments, p or s.

12 N. Muskhelishvili, *Singular Integral Equations* (Noordhoff, Gröningen, 1953).

13 W. Pogorzelski, *Integral Equations and their Applications* (Pergamon, Long Island City, N.Y., 1966).

14 The problem of finding a function with a given discontinuity across a contour is called the "Hilbert problem."

15 In our particular case, $R(s) = 0(1/|s|)$ as $|\mathrm{Im}s| \to \infty$, and thus $\ln H_0(s) \to 0$ also. This implies that $H_0(s) \to 1$ as $|\mathrm{Im}s| \to 0$. Thus,

although we have a solution of the discontinuity equation (64), $H_0(s)$ *cannot* be represented in the form of Eq. (61). We therefore do *not* have a solution of the original homogeneous equation.

[16] E. T. Whitaker and G. N. Watson, *Modern Analysis* (Macmillan, New York, 1943), p. 138.

[17] For $l \neq 0$, as has been seen, $R_l(s)$ has no singularities for $-l+1 < \mathrm{Re}(s) < l+2$. One also sees that, at least for very small Ze^2, there can be no ξ_n near the integers in this region, either. For small Ze^2, the smallest ξ_n above $\frac{3}{2}$ will be near $l+2$. Thus a representation of this form is also available for $l \neq 0$, except that some n values must be excluded.

[18] For $l \neq 0$, C_1 and C_2 would be chosen between the lowest zero-pole pair in $R(s)$ for $\mathrm{Re}(s) > \frac{3}{2}$, and to the right of that pair, respectively. See Footnote 17.

[19] It is to be noted that $H(s)$ must tend to zero at infinity if Eq. (61) is to hold.

[20] See Ref. 12 or 13.

[21] The latter kernel need not be bounded. It is only necessary that, for its kernel K, $\int_C \int_C |K(s,s')|^2 ds ds'$ exists.

[22] The associate to a kernel $K(s,s')$ is given by $K'(s,s') = K(s',s)$.

[23] See Ref. 12 or 13.

[24] See Ref 13.

[25] If $Ze^2 > 0$, the point ξ_2 lies at $\mathrm{Re}\, s > 2$ and $R(s)$ is negative on the entire path discussed here so that for a repulsive potential there is no difficulty.

[26] See K. M. Case, Phys. Rev. 80, 797 (1950).

[27] See, e.g., Titchmarsh, Ref. 10.

[28] Throughout this analysis we have assumed that the integral equation has an integration weight factor of p'^2 so that the complete kernel is $p'^2 K(p,p')$, where $K(p,p')$, is Hermitian. This factor can be modified by a change in the wavefunction of the form $\psi'(p) = p^\alpha \psi(p)$, which then introduces a factor $(p'/p)^\alpha$ in the kernel. If such a change is made, the hermiticity condition becomes $s_R = \frac{3}{2} + \alpha$ and at the same time the conditions of Eq. (39) are also shifted by α.

[29] See, e.g., L. I. Schiff, *Quantum Mechanics* (McGraw-Hill, New York, 1949), pp. 322, ff.

[30] We only consider the case in which the eigenvalue of k is $j + \frac{1}{2}$. The other case can be similarly treated.

[31] This relation has already been used to obtain Eq. (B20), using the integral representation for Q_l.

Nonlinear Perturbation Operators*

L. C. Andrews
Technical Staff, ASW Operation, Magnavox Company, Fort Wayne, Indiana
and
T. Triffet
Center for Applied Mathematics
and
Department of Metallurgy, Mechanics, and Materials Science, Michigan State University, East Lansing, Michigan 48823

(Received 15 January 1971)

A systematic perturbation theory is presented for the analysis of nonlinear boundary-value problems. In particular, those equations are considered whose unperturbed form belongs to the class of linear special-function equations; the nonlinear terms are then regarded as perturbations of special-function operators. Utilizing the coordinate representation of quantum mechanics, a matrix representation is obtained for the perturbed operator, truncation and diagonalization of which will determine the perturbed eigenvalues and eigenvectors. In most cases the method is applicable even when the perturbation term is of the same order of magnitude as the remaining terms, perhaps even when it is larger. To illustrate this point the nonlinear Legendre-like equation $(d/dx)(1-x^2)(du/dx) + \lambda u + \alpha x^2 u^2 = 0$ is solved for the cases when $\alpha = 1$ and $\alpha = 5$. Other examples include the Hartree equations for the helium atom, where a qualitative comparison of the ground-state energy is made with experimental data, and a detailed analysis of the van der Pol equation for $\alpha = 0.5$ and $\alpha = 1$.

1. INTRODUCTION

From the theory of representations it is known that a linear operator L which transforms a Hilbert space into itself gives rise to a matrix representation of that operator. The matrix elements can be defined by the inner product

$$L_{jk} = (\phi_j, L\phi_k), \tag{1.1}$$

where $\{\phi_k\}$ represents a complete set of orthonormal vectors over the space. We shall extend such representations to certain nonlinear forms of special-function operators, the matrix elements of which can be computed by algebraic manipulation instead of by utilizing expressions like (1.1).

The type of nonlinear operator to be considered is represented by the hypergeometric-like form

$$S = \frac{d}{dx}(1-x^2)\frac{d}{dx} - 2(\mu + \nu x)\frac{d}{dx} + \alpha x^2 u_k, \tag{1.2}$$

from which we obtain the nonlinear eigenequation

$$(S + \lambda_k)u_k = 0. \tag{1.3}$$

The term $\alpha x^2 u_k$ is regarded as a perturbation of the linear hypergeometric operator

$$L = \frac{d}{dx}(1-x^2)\frac{d}{dx} - 2(\mu + \nu x)\frac{d}{dx}. \tag{1.4}$$

Thus α will serve as the perturbation parameter, but will be treated as a positive constant of arbitrary magnitude in contrast to the usual restriction that α remain small.[1] The operator S can now be written as

$$S = L + N, \tag{1.5}$$

where $N = \alpha x^2 u_k$, an expression which can itself be regarded as the product of a linear perturbation αx^2 with a nonlinear perturbation u_k. For $\alpha = 0$, of course, Eq. (1.3) becomes a hypergeometric equation with known eigenvalues and eigenvectors. We shall seek solutions of such equations that will reduce to the known solutions in the unperturbed case $\alpha = 0$, and yet will continue to hold when α takes some value large enough to make the nonlinear term at least as important as the linear terms.

In what follows, quantum mechanical examples will be stressed, not because the method to be presented must be limited to such cases, but because they constitute an exceptionally promising field of application. Perturbation techniques have long been used to obtain approximate solutions of the extended, linear special-function forms that usually result from separating variables in the Schrödinger or Klein–Gordon equations. However, as soon as realistic particle interactions are included, these same equations become nonlinear, and little progress has been made in dealing with them systematically. Moreover, it has been widely recognized that any general field equation which incorporates the behavior of the elementary particles must also be nonlinear. Heisenberg, for example, suggested the following equation:

$$i\sigma^\nu \frac{\partial \chi}{\partial x^\nu} + l^2 \sigma^\nu : \chi(\chi^* \sigma_\nu \chi) : = 0, \tag{1.6}$$

where $\chi(x)$ is a local field operator, the σ^ν are the conventional Pauli matrices, l is an arbitrary constant with the dimension of length, and the dots :: in the second term refer to the definition of a product of three field operators at the same space–time point.[2] Einstein's writings leave no doubt that he believed in the existence of such an equation, and de Broglie has for many years invoked a nonlinear wave equation in connection with his double solution theory.[3]

We wish to rewrite the differential equation (1.3) in terms of algebraic operators representing the dynamic variables. This involves transformation of the functions $u_k(x)$ of the differential equation to the vectors $\psi^{(k)}(q)$ of the algebraic equation. Such a transformation can be represented by

$$u_k(x) = \overline{\psi}^{(0)} e^{ipx} \psi^{(k)}(q), \tag{1.7}$$

where x is treated as an unrestricted coordinate variable and p and q, the linear momentum and position operators, are understood to satisfy $qp - pq = i$ (with $\hbar = 1$). Since in general

$$xe^{ipx} = e^{ipx}q - qe^{ipx}, \tag{1.8}$$

it follows that

$$xu_k(x) = \overline{\psi}^{(0)}e^{ipx}q\psi^{(k)}, \tag{1.9}$$

$$-i\frac{d}{dx}u_k(x) = \overline{\psi}^{(0)}e^{ipx}p\psi^{(k)}, \tag{1.10}$$

assuming that $q\psi^{(0)} = 0$. Using these relations it is easily deduced that

$$\frac{d}{dx}(1-x^2)\frac{du_k}{dx} = -\overline{\psi}^{(0)}e^{ipx}p(1-q^2)p\psi^{(k)}, \tag{1.11}$$

$$-2(\mu+\nu x)\frac{du_k}{dx} = -\overline{\psi}^{(0)}e^{ipx}2i(\mu+\nu q)p\psi^{(k)}, \tag{1.12}$$

$$\lambda_k u_k = \overline{\psi}^{(0)}e^{ipx}\lambda^{(k)}\psi^{(k)}; \tag{1.13}$$

but the nonlinear term in (1.3) requires further analysis.

To transform this term we write, parallel to (1.7),

$$u_k(q) = \overline{\psi}^{(0)}e^{ip\otimes q}\psi^{(k)} = \overline{W}(q,\psi^{(k)}), \tag{1.14}$$

where $p\otimes q$ represents the direct product matrix $p_{ik}q_{jl}$, so that $\overline{W}(q,\psi^{(k)})$ will in general be some operator function of q and $\psi^{(k)}$. For most of the applications we have in mind, however, $\overline{W}(q,\psi^{(k)})$ can be expressed in the form $\overline{V}(q)\ \psi^{(k)}$, where, as will be shown later, $\overline{V}(q)$ is automatically absorbed in the process of forming the matrix elements of the nonlinear operator. Hence, the term becomes

$$\begin{aligned}\alpha x^2u_k^2 &= \overline{\psi}^{(0)}e^{ipx}\alpha q^2u_k(q)\psi^{(k)}\\ &= \overline{\psi}^{(0)}e^{ipx}\alpha q^2\overline{V}(q)\psi^{(k)2};\end{aligned} \tag{1.15}$$

and, combining results, (1.3) may now be written as

$$(A-\lambda^{(k)})\psi^{(k)} = 0, \tag{1.16}$$

where

$$A = p(1-q^2)p + 2i(\mu+\nu q)p - \alpha q^2\overline{V}(q)\psi^{(k)}. \tag{1.17}$$

More generally, the algebraic operator A may be taken to have the form

$$A = \Gamma(p,q)B(p,q) + \Delta(p,q)Z(p,q,\psi^{(k)}), \tag{1.18}$$

where B represents an unperturbed self-adjoint linear operator, Γ and Δ are nonsingular linear perturbations, and Z is a nonlinear perturbation. For the above case,

$$\begin{aligned}B &= p(1-q^2)p + 2i(\mu+\nu q)p,\\ \Gamma &= 1,\\ \Delta &= -\alpha q^2,\\ Z &= \overline{V}(q)\psi^{(k)},\end{aligned} \tag{1.19}$$

while two other examples with different features are, firstly,

$$\begin{aligned}&\frac{d^2u_k}{dx^2} + (\lambda_k - x^2)u_k + \alpha x^4u_k^3 = 0,\\ &[p^2+q^2-\alpha q^4\,\overline{V}(q)\,\psi^{(k)2} - \lambda^{(k)}]\psi^{(k)} = 0,\\ &B = p^2+q^2,\\ &\Gamma = 1,\\ &\Delta = -\alpha q^4,\\ &Z = \overline{V}(q)\psi^{(k)2},\end{aligned} \tag{1.20}$$

and secondly,

$$\begin{aligned}&(1-ax^2)\frac{d}{dx}(1-x^2)\frac{du_k}{dx} + 2\alpha(1-x^2)xu_k\frac{du_k}{dx}\\ &\quad + \lambda_k u_k = 0,\\ &[(1-aq^2)p(1-q^2)p - i\alpha(1-q^2)qp\ \overline{V}(q)\psi^{(k)} - \lambda^{(k)}]\\ &\quad\times\psi^{(k)} = 0,\\ &B = p(1-q^2)p,\\ &\Gamma = 1-aq^2,\\ &\Delta = -i\alpha(1-q^2)qp,\\ &Z = \overline{V}(q)\psi^{(k)}.\end{aligned} \tag{1.21}$$

When $\alpha = 0$ and $\Gamma = 1$, Eq. (1.16) reduces to

$$(B-b^{(k)})U^{(k)} = 0, \tag{1.22}$$

where $b^{(k)}$ and $U^{(k)}$ denote the unperturbed eigenvalues and eigenvectors. A factorization method for solving linear differential eigenequations of this type was suggested by Schrödinger in 1940 and later developed in detail by Infeld and Hull.[4] The procedure is to assume that for some self-adjoint operator M, whose eigenvalues m are desired, there exist mutually adjoint linear operators J^+, J^- (usually complex conjugates) satisfying

$$[M,J^+] = MJ^+ - J^+M = J^+, \tag{1.23}$$

$$[M,J^-] = -J^-. \tag{1.24}$$

If, in addition, they satisfy certain other conditions, the operator B is said to admit a factorization. In particular, if a transformation can be found such that B has the form

$$B = p^2 + w(q), \tag{1.25}$$

then Eq. (1.22) can be factorized into the two equations:

$$[K(q,M+1) - ip]U_m^{(j)} = [b^{(j)} - a(m+1)]^{1/2}U_{m+1}^{(j)}, \tag{1.26}$$

$$[K(q,M) + ip]U_m^{(j)} = [b^{(j)} - a(m)]^{1/2}U_{m-1}^{(j)}, \tag{1.27}$$

with

$$J^+ = K(q,M+1) - ip, \tag{1.28}$$

$$J^- = K(q,M) + ip. \tag{1.29}$$

Actually, (1.26) and (1.27) were originally presented as differential equations

$$\left(K(x,m+1)-\frac{d}{dx}\right)u_j^m=[b_j-a(m+1)]^{1/2}u_j^{m+1} \tag{1.30}$$

and

$$\left(K(x,m)+\frac{d}{dx}\right)u_j^m=[b_j-a(m)]^{1/2}u_j^{m-1}, \tag{1.31}$$

and the fact established that a necessary and sufficient condition of factorization is for K and a to satisfy the equation

$$\frac{d}{dx}[K(x,m+1)+K(x,m)]+K^2(x,m+1)-K^2(x,m)+a(m+1)-a(m)=0. \tag{1.32}$$

Miller has also shown that Riccati equations similar to (1.32) are sufficient to determine the four-dimensional Lie algebras $\mathcal{G}(a,b)$ whose representations correspond to a study of all special functions of hypergeometric type.[5,6]

The factorizations (1.26) and (1.27) permit one to obtain the eigenvalues and eigenvectors of the operator B in a simple and elegant manner. However, the step operators J^+, J^- are not unique, as more than one factorization is often possible. Green and Triffet have recently introduced a systematic procedure for determining such operators.[7] They have shown that construction of a sequence of linearly independent operators, which do not commute with the unperturbed operator B, will ultimately lead to an appropriate form. Step operators determined in this fashion will usually not be mutually adjoint; however, they retain their most useful properties and are more general than those defined by the Infeld–Hull factorization technique. Also featured is the basic notational scheme utilized here and an algebraic method for finding matrix representations of linear operators which has been adapted to the present purpose.

2. GENERAL DEVELOPMENT

A nonlinear differential equation of the form

$$(L+\alpha N+\lambda_k)u_k=0, \tag{2.1}$$

where L is a linear differential operator, N a nonlinear differential operator, and α a constant of arbitrary magnitude can be treated as a perturbation problem. To do so, however, it is generally more convenient to utilize "physical boundary conditions" instead of the more common artificially imposed boundary conditions related to the Sturm–Liouville equation. For the special case when $\alpha=0$, Eq. (2.1) becomes linear, so it is natural to regard α as the perturbation parameter.

Writing (2.1) in algebraic form, displaying the quantum mechanical conjugate variables p and q, we arrive at

$$(A-\lambda^{(k)})\psi^{(k)}=(\Gamma B+\Delta Z-\lambda^{(k)})\psi^{(k)}=0, \tag{2.2}$$

where Δ is understood to contain α as a factor. The linear operator L therefore consists of a self-adjoint operator B multiplied by a perturbation Γ (ordinarily unity in the examples to be considered). The nonlinear operator N is decomposed in such a way as to separate out the dependency on the eigenvectors $\psi^{(k)}$; only Z will be a function of $\psi^{(k)}$. Introducing the distinct operators Δ and Z for N is not necessary but merely a convenience for calculating the matrix elements, since Δ may then be interpreted as a linear perturbation of the type treated in Ref. 7.

We seek a matrix representation for A,

$$A_{jk}=\Sigma_n(\Gamma_{jn}B_{nk}+\Delta_{jn}Z_{nk}), \tag{2.3}$$

diagonalization of which by numerical methods will determine the perturbed eigenvalues. Hence, our general procedure will be to find separate matrix representations for each of the operators Γ, B, Δ, and Z. For each choice of a basis which defines the Hilbert space a different, equivalent representation will be obtained; but the most convenient basis to select is the set of normalized eigenvectors belonging to the unperturbed operator B. The representation for B will then be the diagonal matrix

$$B_{jk}=b^{(j)}\delta_{jk}, \tag{2.4}$$

where the $b^{(j)}$ are the eigenvalues of B, and it only remains to determine the matrix representations for Γ, Δ, and Z with respect to this basis.

A. The Operators B, Δ, and Γ

A method of finding such representations for Γ and Δ when B is some special-function operator is also presented by Green and Triffet, together with specific results for a number of important cases.[7] Building on their technique for defining the step operators $J^{(1)}$ and $J^{(2)}$ (generalizations of J^+ and J^-), they assume that Δ, for example, can be represented as a function of these operators and a third operator M, in terms of which B can easily be expressed, but whose eigenvalues $m^{(k)}=m^{(1)}+k-1$ are separated by unity.

Denoting by $\Delta^{(0)}$ the diagonal matrix whose nonzero elements are identical with the center diagonal elements of Δ, and by $\Delta^{(r)}\theta_-^r$ and $\theta_+^r\Delta^{(-r)}$ the matrices whose only nonzero elements are identical with those of Δ in the rth diagonal above and below the center diagonal, respectively, one may write

$$\Delta=\Delta^{(0)}+\sum_{r=1}^{\infty}(\theta_+^r\Delta^{(-r)}+\Delta^{(r)}\theta_-^r),$$
$$\Delta^{(r)}=\Delta^{(r)}(M). \tag{2.5}$$

Since the matrix elements of the unit step operators θ_+ and θ_- will be

$$(\theta_+)_{jk}=\delta_{j\,k+1}, \tag{2.6}$$

$$(\theta_-)_{jk}=\delta_{j+1\,k}, \tag{2.7}$$

the matrix elements of Δ will be given by

$$\Delta_{jk} = \Delta_j^{(0)}\delta_{jk} + \sum_{r=1}^{\infty}(\Delta_k^{(-r)}\delta_{j\,k+r} + \Delta_j^{(r)}\delta_{j+r\,k}),$$

$$\Delta_k^{(r)} = \Delta^{(r)}(m^{(k)}). \tag{2.8}$$

But, supposing that Δ can be represented as some function of $J^{(1)}, J^{(2)}$, and M, we may also write

$$\Delta = E^{(0)} + \sum_{r=1}^{\infty}(J^{(1)r}E^{(-r)} + E^{(r)}J^{(2)r}), \tag{2.9}$$

with

$$E^{(r)} = E^{(r)}(M).$$

It is established by Green and Triffet that

$$J^{(1)r} = \theta_+^r h^{(1)}(M)h^{(1)}(M+1)\cdots h^{(1)}(M+r-1), \tag{2.10}$$

$$J^{(2)r} = h^{(2)}(M)h^{(2)}(M+1)\cdots h^{(2)}(M+r-1)\theta_-^r, \tag{2.11}$$

where

$$h^{(1)}(M-1) = [J^{(1)}\bar{J}^{(1)}]^{1/2} \tag{2.12}$$

and

$$h^{(2)}(M-1) = [\bar{J}^{(2)}J^{(2)}]^{1/2}. \tag{2.13}$$

Thus, comparison of (2.5) with (2.9) yields

$$\Delta^{(0)} = E^{(0)}(M), \tag{2.14}$$

$$\Delta^{(-r)} = E^{(-r)}(M)h^{(1)}(M)h^{(1)}(M+1)\cdots h^{(1)}(M+r-1), \tag{2.15}$$

$$\Delta^{(r)} = E^{(r)}(M)h^{(2)}(M)h^{(2)}(M+1)\cdots h^{(2)}(M+r-1). \tag{2.16}$$

From these relations the matrix representation of appropriate perturbation terms can be obtained. This is explicitly illustrated in Sec. 3.

B. The Operator Z

There is a difficulty inherent in the form of the operator Z that does not occur for the linear operators; it is a function of the unknown perturbed eigenvectors $\psi^{(k)}$. However, this problem can be alleviated by applying a technique characteristic of the standard theory. It is assumed that $\psi^{(k)}$ is a continuous function of the parameter α. When such a parameter does not explicitly appear in the eigenequation it can be introduced and later set equal to 1, so that we can always form the Taylor expansion about $\alpha = 0$:

$$\psi^{(k)}(\alpha) = \phi^{(k)} + \alpha\psi_1^{(k)} + \cdots, \tag{2.17}$$

where $\psi_1^{(k)}$ is the derivative of $\psi^{(k)}$ with respect to α, evaluated at $\alpha = 0$, and $\phi^{(k)}$ is a normalized eigenvector of the operator B. Assuming that Z is analytic, it too may be expanded about $\alpha = 0$:

$$Z(\alpha, \psi^{(k)}) = Y + \alpha Z_1 + \cdots, \tag{2.18}$$

where, by definition,

$$Y = Z(0, \phi^{(k)}) \tag{2.19}$$

and

$$Z_1 = \frac{\partial Z}{\partial \psi^{(k)}}\psi_1^{(k)}|_{\alpha=0}. \tag{2.20}$$

For first-order perturbations we shall henceforth approximate the operator Z by Y. The validity of such an approximation will, of course, depend on the convergence properties of (2.18).

Equation (2.19) makes clear the dependence of Y on the index k. Since this index refers to the kth column in the matrix representation, the matrix elements of Y must be calculated by columns, i.e., for each column, Y will essentially be a different operator. It is this column selection for the matrix elements of Y that constitutes the primary function of the operator $\bar{V}(q)$ defined in the Introduction.

The operator Y can be represented in a manner analogous to Δ. We shall denote by $Y^{(0)}$, $\theta_+^r Y^{(-r)}$, and $Y^{(r)}\theta_-^r$ matrices similar to those defined by the corresponding expressions containing the operator Δ; hence,

$$Y = Y^{(0)} + \sum_{r=1}^{\infty}(\theta_+^r Y^{(-r)} + Y^{(r)}\theta_-^r),$$
$$Y^{(r)} = Y^{(r)}(M), \tag{2.21}$$

with the matrix elements

$$Y_{jk} = Y_j^{(0)}\delta_{jk} + \sum_{r=1}^{\infty}(Y_k^{(-r)}\delta_{j\,k+r} + Y_j^{(r)}\delta_{j+r\,k}),$$
$$Y_j^{(r)} = Y^{(r)}(m^{(j)}). \tag{2.22}$$

Assuming also that Y, like Δ, can be expressed as a function of $J^{(1)}, J^{(2)}$, and M, we obtain

$$Y = E^{(0)} + \sum_{r=1}^{\infty}(J^{(1)r}E^{(-r)} + E^{(r)}J^{(2)r}), \tag{2.23}$$

which, with the aid of Eqs. (2.10) and (2.11), allows us to make the associations

$$Y^{(0)} = E^{(0)}(M), \tag{2.24}$$

$$Y^{(-r)} = E^{(-r)}(M)h^{(1)}(M)h^{(1)}(M+1)\cdots h^{(1)}(M+r-1), \tag{2.25}$$

$$Y^{(r)} = E^{(r)}(M)h^{(2)}(M)h^{(2)}(M+1)\cdots h^{(2)}(M+r-1). \tag{2.26}$$

Because of the presence of the operator $\bar{V}(q)$, the above representation must be repeated for each column of the matrix Y, but only those elements that occur in the given column need be calculated. For example, suppose that the matrix elements in the third column of Y corresponding to $Y = \phi^{(3)}$ are desired. By utilizing Eqs. (2.21)–(2.26), the complete matrix representation for $\phi^{(3)}$ could be obtained:

$$\phi^{(3)} = \begin{bmatrix} \phi_{11}^{(3)} & \phi_{12}^{(3)} & \phi_{13}^{(3)} & \phi_{14}^{(3)} & \cdots \\ \phi_{21}^{(3)} & \phi_{22}^{(3)} & \phi_{23}^{(3)} & \phi_{24}^{(3)} & \cdots \\ \phi_{31}^{(3)} & \phi_{32}^{(3)} & \phi_{33}^{(3)} & \phi_{34}^{(3)} & \cdots \\ \phi_{41}^{(3)} & \phi_{42}^{(3)} & \phi_{43}^{(3)} & \phi_{44}^{(3)} & \cdots \\ & \cdots & & & \end{bmatrix}. \tag{2.27}$$

However, most of this effort is unnecessary since only the elements in the third column will be needed to represent Y.

To help eliminate other unnecessary calculations in computing the matrix elements of Y, we note that whenever the coordinate q has a symmetric matrix representation, then

$$E^{(r)}(M)h^{(2)}(M)h^{(2)}(M+1)\cdots h^{(2)}(M+r-1) = E^{(-r)}(M)h^{(1)}(M)h^{(1)}(M+1)\cdots h^{(1)}(M+r-1). \tag{2.28}$$

This means that none of the coefficients of $J^{(2)r}$ in Eq. (2.23) need be computed; the coefficients of $J^{(1)r}$ will determine all of the matrix elements in a given column.

3. A NONLINEAR LEGENDRE-LIKE EQUATION

When $\mu = \nu = 0$, Eq. (1.3) reduces to the nonlinear Legendre-like equation

$$\left(\frac{d}{dx}(1-x^2)\frac{d}{dx} + \alpha x^2 u_k + \lambda_k\right)u_k = 0, \tag{3.1}$$

which in operational form featuring p and q becomes

$$[p(1-q^2)p - \alpha q^2 \overline{V}(q)\psi^{(k)} - \lambda^{(k)}]\psi^{(k)} = 0. \tag{3.2}$$

Expanding $\psi^{(k)}$ in a Taylor series about $\alpha = 0$ yields

$$\psi^{(k)} = \phi^{(k)} + \alpha\psi_1^{(k)} + \cdots, \tag{3.3}$$

where $\phi^{(k)}$ again represents the normalized eigenvectors of the unperturbed operator. Hence, substitution and retention only of terms linear in α gives

$$[p(1-q^2)p - \alpha q^2 \overline{V}(q)\phi^{(k)} - \lambda^{(k)}]\psi^{(k)} = 0. \tag{3.4}$$

Comparison of this with the standard form [(1.16)–(1.18)] will then establish the following identifications:

$$\begin{aligned} B &= p(1-q^2)p, \\ \Gamma &= 1, \\ \Delta &= -\alpha q^2, \\ Y &= \overline{V}(q)\phi^{(k)}. \end{aligned} \tag{3.5}$$

A. The Operator B

Before the matrix representations of the operators Δ and Y can be constructed, the eigenvalues and eigenvectors of the operator B must be found. The unperturbed form of Eq. (3.4) is given by

$$[p(1-q^2)p - b^{(k)}]U^{(k)} = 0. \tag{3.6}$$

To determine the associated step operators $J^{(1)}$, $J^{(2)}$, it is most convenient to select $J_1 = q$ in the definitive relations provided in Ref. 7:

$$[B, J_k] = \Sigma_j J_j c_{jk}, \tag{3.7}$$

$$J^{(n)} = \Sigma_k J_k \xi_k^{(n)}, \quad n = 1, 2. \tag{3.8}$$

Then

$$\begin{aligned} [B, J_1] &= [p^2, q] - q^2[p^2, q] + 2iq[p, q] \\ &= -2i(1-q^2)p + 2q \\ &= 2(J_1 + J_2) \end{aligned} \tag{3.9}$$

$$\begin{aligned} [B, J_2] &= [-ip(1-q^2) + i(1-q^2)p]p(1-q^2)p \\ &= i[p, q^2]p(1-q^2)p \\ &= 2J_1 B, \end{aligned} \tag{3.10}$$

where $J_2 = -i(1-q^2)p$. The eigenvalues of the matrix c_{jk} are $\Lambda^{(1)} = 1 + (1+4B)^{1/2}$ and $\Lambda^{(2)} = 1 - (1+4B)^{1/2}$. To avoid the square root B may be equated to $M(M+1)$, so that $\Lambda^{(1)} = 2(M+1)$ and $\Lambda^{(2)} = -2M$. The right eigenvectors of c_{jk} are $\xi^{(1)} = (M+1, 1)$ and $\xi^{(2)} = (M, -1)$; hence

$$J^{(1)} = q(M+1) - i(1-q^2)p, \tag{3.11}$$

$$J^{(2)} = qM + i(1-q^2)p. \tag{3.12}$$

Now

$$\begin{aligned} J^{(1)}J^{(2)} &= [q(M+1) - i(1-q^2)p]J^{(2)} \\ &= qJ^{(2)}M - i(1-q^2)pJ^{(2)} \\ &= q^2M^2 + iq(1-q^2)pM - i(1-q^2)pqM \\ &\quad + (1-q^2)p(1-q^2)p \\ &= M^2 - (1-q^2)M(M+1) + (1-q^2)p \\ &\quad \times (1-q^2)p \\ &= M^2, \end{aligned} \tag{3.13}$$

from which it may be concluded that the first eigenvalue of M is $m^{(1)} = 0$.

Therefore $m^{(k)} = k - 1$ and applying the relation $b^{(k)} = m^{(k)}(m^{(k)} + 1)$ gives the eigenvalues

$$b^{(k)} = k(k-1). \tag{3.14}$$

The first eigenvector $U^{(1)}$ satisfies

$$J^{(2)}U^{(1)} = 0, \tag{3.15}$$

which is equivalent to the differential equation

$$(1-x^2)\frac{dv_1}{dx} = 0, \tag{3.16}$$

with v_1 corresponding to the unperturbed eigenfunction. Thus $v_1 = \text{const}$ is a solution of (3.16). Successive eigenvectors may be determined from

$$U^{(k+1)} = J^{(1)k}U^{(1)}, \tag{3.17}$$

which in the form of (3.16) becomes

$$v_{k+1} = \left((k+1)x + (x^2-1)\frac{d}{dx}\right)^k v_1 = C_k \frac{d^k}{dx^k}(x^2-1)^k, \tag{3.18}$$

where the C_k's are constants. Normalization of the eigenfunctions v_k yields

$$v_k(x) = (k-\tfrac{1}{2})^{1/2}P_{k-1}(x), \tag{3.19}$$

the

$$P_k(x) = \frac{1}{2^k k!}\frac{d^k}{dx^k}(x^2-1)^k$$

being the well-known Legendre polynomials. Accordingly, the normalized eigenvectors of (3.6) will be given by

$$\phi^{(k)} = (k-\tfrac{1}{2})^{1/2}P_{k-1}(q). \tag{3.20}$$

Since

$$\bar{J}^{(1)} = (M-1)q + i(1-q^2)p, \tag{3.21}$$

but from (3.11) and (3.12)

$$q = (J^{(1)} + J^{(2)})/(2M+1) \tag{3.22}$$

and

$$-i(1-q^2)p = [J^{(1)}M - J^{(2)}(M+1)]/(2M+1), \tag{3.23}$$

it follows that

$$\begin{aligned}\bar{J}^{(1)} &= (M-1)(J^{(1)} + J^{(2)})/(2M+1) \\ &\quad - [J^{(1)}M - J^{(2)}(M+1)]/(2M+1) \\ &= [J^{(1)}M + J^{(2)}(M-2) - J^{(1)}M + J^{(2)}(M+1)]/ \\ &\quad \times (2M+1)^{-1} \\ &= J^{(2)}(2M-1)/(2M+1).\end{aligned} \tag{3.24}$$

Thus, $\bar{J}^{(2)}$ is also determined and Eqs. (2.12) and (2.13) define

$$h^{(1)}(M-1) = M(2M-1)^{1/2}/(2M+1)^{1/2} \tag{3.25}$$

and

$$h^{(2)}(M-1) = M(2M+1)^{1/2}/(2M-1)^{1/2}. \tag{3.26}$$

Results for the operator B are summarized below for easy reference:

(a) $B = p(1-q^2)p$, $\eta = 1$ (see Ref. 7);

(b) $B = M(M+1)$;

(c) $J_1 = q = (J^{(1)} + J^{(2)})/(2M+1)$,
$J_2 = -i(1-q^2)p = [J^{(1)}M - J^{(2)}(M+1)]/(2M+1)$;

(d) $m^{(k)} = k-1, \quad b^{(k)} = k(k-1)$;

(e) $\phi^{(k)} = (k-\tfrac{1}{2})^{1/2}P_{k-1}(q)$;

(f) $h^{(1)}(M) = (M+1)(2M+1)^{1/2}/(2M+3)^{1/2}$;
$h^{(2)}(M) = (M+1)(2M+3)^{1/2}/(2M+1)^{1/2}$;

(g) $J^{(2)}J^{(1)} = h^{(2)}(M)h^{(1)}(M) = (M+1)^2$.

B. Matrix Representation for Δ

The matrix representation for $\Delta = -\alpha q^2$ can readily be obtained from the expression

$$\begin{aligned}q^2 &= M^2[(2M+1)(2M-1)]^{-1} + (M+1)^2 \\ &\times [(2M+1)(2M+3)]^{-1} + J^{(1)2}[(2M+1)(2M+3)]^{-1} \\ &+ [(2M+3)(2M+5)]^{-1}J^{(2)2},\end{aligned} \tag{3.27}$$

derived by employing the above list of properties. From Eq. (2.9)

$$\Delta = E^{(0)} + \sum_{r=1}^{\infty}(J^{(1)r}E^{(-r)} + E^{(r)}J^{(2)r}), \tag{3.28}$$

so that the following may also be identified:

$$\begin{aligned}E^{(0)}(M) &= -\alpha\{M^2[(2M+1)(2M-1)]^{-1} \\ &\quad + (M+1)^2[(2M+1)(2M+3)]^{-1}\},\end{aligned} \tag{3.29}$$

$$E^{(-2)}(M) = -\alpha[(2M+1)(2M+3)]^{-1}, \tag{3.30}$$

$$E^{(2)}(M) = -\alpha[(2M+3)(2M+5)]^{-1}, \tag{3.31}$$

all other $E^{(r)}(M)$ being zero. Thus, Eqs. (2.50)–(2.52) yield

$$\Delta^{(0)}(M) = -\alpha(2M^2+2M-1)[(2M-1)(2M+3)]^{-1}, \tag{3.32}$$

and

$$\begin{aligned}\Delta^{(-2)}(M) &= \Delta^{(2)}(M) \\ &= -\alpha(M+1)(M+2)[(2M+1)(2M+3)^2 \\ &\quad \times (2M+5)]^{-1/2}.\end{aligned} \tag{3.33}$$

Consequently, since $m^{(k)} = k-1$, the matrix elements are

$$\Delta_{jk} = \Delta^{(0)}(m^{(j)})\delta_{jk} + \Delta^{(-2)}(m^{(k)})\delta_{j\,k+2} + \Delta^{(2)}(m^{(j)})\delta_{j+2\,k} \tag{3.34}$$

or

$$\Delta_{kk} = -\tfrac{1}{2}\alpha(k^2-k-\tfrac{1}{2})/[(k+\tfrac{1}{2})(k-\tfrac{3}{2})], \tag{3.35}$$

$$\begin{aligned}\Delta_{k\,k+2} &= \Delta_{k+2\,k} \\ &= -\tfrac{1}{4}\alpha k(k+1)/[(k+\tfrac{1}{2})^2(k-\tfrac{1}{2})(k+\tfrac{3}{2})^{1/2}.\end{aligned} \tag{3.36}$$

with all other $\Delta_{jk} = 0$.

C. Matrix Representation for Y

By utilizing (3.22) the normalized eigenvectors of the operator B can now be expressed as functions of $J^{(1)}$, $J^{(2)}$, and M:

$$\phi^{(1)} = (\tfrac{1}{2})^{1/2}, \tag{3.37}$$

$$\phi^{(2)} = (\tfrac{3}{2})^{1/2} q$$
$$= (\tfrac{3}{2})^{1/2}(J^{(1)} + J^{(2)})/(2M+1), \quad (3.38)$$

$$\phi^{(3)} = \tfrac{1}{2}(\tfrac{5}{2})^{1/2}(3q^2 - 1)$$
$$= \tfrac{1}{2}(\tfrac{5}{2})^{1/2}\{3[(J^{(1)} + J^{(2)})/(2M+1)]^2 - 1\}, \quad (3.39)$$
$$\cdots$$

Keeping in mind that the matrix elements for Y must be calculated by columns as determined by $\overline{V}(q)$, we have for the first column, where $Y = \phi^{(1)}$,

$$E^{(0)}(M) = (\tfrac{1}{2})^{1/2}. \quad (3.40)$$

Therefore

$$E^{(0)}(0) = (\tfrac{1}{2})^{1/2}, \quad (3.41)$$

and the only nonzero matrix element in this column is

$$Y_{11} = (\tfrac{1}{2})^{1/2}. \quad (3.42)$$

For the second column, where $Y = \phi^{(2)}$,

$$E^{(0)}(M) = 0$$

and

$$E^{(-1)}(M) = (\tfrac{3}{2})^{1/2}/(2M+1), \quad (3.43)$$

so that by Eqs. (2.25)–(2.28) the required matrix elements must be

$$Y_{12} = E^{(-1)}(0)h^{(1)}(0)$$
$$= (\tfrac{1}{2})^{1/2}, \quad (3.44)$$

$$Y_{22} = E^{(0)}(1)$$
$$= 0, \quad (3.45)$$

$$Y_{32} = E^{(-1)}(1)h^{(1)}(1)$$
$$= \tfrac{2}{5}(\tfrac{5}{2})^{1/2}. \quad (3.46)$$

For the third column,

$$Y = \phi^{(3)}$$
$$= \tfrac{3}{2}(\tfrac{5}{2})^{1/2}\{M^2[(2M+1)(2M-1)]^{-1} + (M+1)^2 \times [(2M+1)(2M+3)]^{-1} - \tfrac{1}{3} + J^{(1)2} \times [(2M+1)(2M+3)]^{-1} + [(2M+3)(2M+5)]^{-1} J^{(2)2}\}. \quad (3.47)$$

Thus

$$E^{(0)}(M) = \tfrac{3}{2}(\tfrac{5}{2})^{1/2}\{M^2[(2M+1)(2M-1)]^{-1} + (M+1)^2[(2M+1)(2M+3)]^{-1} - \tfrac{1}{3}\},$$
$$E^{(-2)}(M) = \tfrac{3}{2}(\tfrac{5}{2})^{1/2}[(2M+1)(2M+3)]^{-1}, \quad (3.48)$$

and the nonzero matrix elements become

$$Y_{13} = E^{(-2)}(0)h^{(1)}(0)h^{(1)}(1)$$
$$= (\tfrac{1}{2})^{1/2}, \quad (3.49)$$

$$Y_{33} = E^{(0)}(2)$$
$$= \tfrac{2}{7}(\tfrac{5}{2})^{1/2}, \quad (3.50)$$

$$Y_{53} = E^{(-2)}(2)h^{(1)}(2)h^{(1)}(3)$$
$$= \tfrac{6}{7}(\tfrac{1}{2})^{1/2}. \quad (3.51)$$

Following the same procedure for $\phi^{(4)}, \phi^{(5)}, \cdots$, will determine as many columns of Y as may be needed.

D. Discussion of Results

The first few matrix elements calculated from the above relations for B, Δ, and Y as they appear in the nonlinear Legendre-like equation (3.4) are displayed below:

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & 2 & 0 & 0 & \cdots \\ 0 & 0 & 6 & 0 & \cdots \\ 0 & 0 & 0 & 12 & \cdots \\ & & \cdots & & \end{bmatrix}, \quad (3.52)$$

$$\Delta = -\alpha \begin{bmatrix} \frac{1}{3} & 0 & 2/3\sqrt{5} & \cdots \\ 0 & \frac{3}{5} & 0 & \cdots \\ 2/3\sqrt{5} & 0 & \frac{11}{21} & \cdots \\ & & \cdots & \end{bmatrix}, \quad (3.53)$$

$$Y = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 2/\sqrt{10} & 2\sqrt{5}/7\sqrt{2} & \cdots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 6/7\sqrt{2} & \cdots \\ & & \cdots & \end{bmatrix}. \quad (3.54)$$

Combining these results gives the desired representation for A:

$$A_{jk} = b^{(j)}\delta_{jk} + \Sigma_n \Delta_{jn} Y_{nk}. \quad (3.55)$$

A modified version of the FORTRAN program described by Green and Triffet[7] was used to find the eigenvalues and eigenvectors of the resulting nonsymmetric matrix. The eigenvalues corresponding to $\alpha = 1$, computed by truncating A to 10 rows and columns, are listed in Table I, along with comparable values computed by the standard perturbation method.[1]

The reason that every other value agrees exactly is that the nondiagonal elements of A are zero for every other row; hence, the diagonal element is itself an eigenvalue. This is a peculiarity of the particular matrix being considered. The difference in the other eigenvalues can be accounted for by the fact that standard perturbation theory presumes that the matrix A is essentially a diagonal matrix from the beginning, i.e., that all nondiagonal elements are either zero or negligible in comparison with the diagonal elements. For such a matrix the eigenvalues can be approximated by the diagonal elements themselves. To assure that the nondiagonal elements will be small, conventional theory requires that the perturbation parameter α be small, thus forcing the desired condition.

Table I. A comparison of eigenvalues for Eq. (3. 1) obtained by standard and matrix perturbation methods with $\alpha = 1$.

Standard method		Matrix method	
k	$\lambda^{(k)}$	k	$\lambda^{(k)}$
1	− 0. 235 70	1	− 0. 249 90
2	2. 000 0	2	2. 000 0
3	5. 397 7	3	5. 394 3
4	12. 000	4	12. 000
5	19. 631	5	19. 638
6	30. 000	6	30. 000
7	41. 703	7	41. 709
8	56. 000	8	56. 000
9	71. 744	9	71. 750
10	90. 000	10	90. 000

Table II. Eigenvalues for the matrix given in Eq. (3. 52)–(3. 55) with $\alpha = 5$.

k	$\lambda^{(k)}$	k	$\lambda^{(k)}$
1	− 1. 6896	14	182. 00
2	2. 0000	15	209. 04
3	3. 0686	16	240. 00
4	12. 000	17	271. 08
5	18. 309	18	306. 00
6	30. 000	19	341. 07
7	40. 591	20	380. 00
8	56. 000	21	418. 97
9	70. 758	22	462. 00
10	90. 000	23	504. 78
11	108. 88	24	552. 00
12	132. 00	25	598. 51
13	154. 97		

When $\alpha = 1$, the nondiagonal elements of the matrix (3. 55) are relatively small so that the two methods of computing eigenvalues should give nearly the same results. As α increases in magnitude, however, a great deal more error is introduced by neglecting the nondiagonal elements of A.

To present a qualitative comparison of the two methods we shall examine how accurately the eigenvalues and eigenfunctions satisfy the original equation. The first eigenvalue and eigenfunction computed by the standard theory are

$$\lambda_1 = -0.23570 \tag{3.56}$$

and

$$u_1(x) = v_1(x) + 0.0351\, v_3(x), \tag{3.57}$$

where $v_k(x)$ is defined by Eq. (3. 19). Substituting these expressions into Eq. (3. 1) yields the inequality

$$\left|\frac{d}{dx}(1-x^2)\frac{du_1}{dx} + \lambda_1 u_1 + x^2 u_1^2\right| \leq 0.069, \quad |x| \leq 1, \tag{3.58}$$

whereas an exact solution would reduce the right-hand size to zero. The corresponding eigenvalue and eigenfunction for the matrix method are

$$\lambda_1 = -0.24990 \tag{3.59}$$

and

$$u_1(x) \cong v_1(x) + 0.0374\, v_3(x) + 0.0008\, v_5(x) + 0.0001\, v_7(x). \tag{3.60}$$

Substituting these values into Eq. (3. 1) yields

$$\left|\frac{d}{dx}(1-x^2)\frac{du_1}{dx} + \lambda_1 u_1 + x^2 u_1^2\right| \leq 0.004, \quad |x| \leq 1. \tag{3.61}$$

Equation (3. 57) has been limited to the sum of two unperturbed eigenfunctions because that is all the standard theory predicts. On the other hand, the matrix method generates an infinite series of terms, the first four of which have been used for this calculation.

To provide an example where the perturbation term dominates the remaining linear terms, we have chosen the case where $\alpha = 5$. Table II contains the eigenvalues obtained by truncating A to 25 rows and columns. Substituting the first eigenvalue and eigenfunction computed into Eq. (3. 1) gives

$$\left|\frac{d}{dx}(1-x^2)\frac{du_1}{dx} + \lambda_1 u_1 + 5x^2 u_1^2\right| \leq 0.44, \quad |x| \leq 1. \tag{3.62}$$

Though this represents substantially more error than when $\alpha = 1$, it still may be considered a fair approximation for a first-order perturbation. No attempt at comparison with standard perturbation theory is made, since the latter is not expected to give even a qualitative solution for such a large perturbation.

In general the amount of error will increase with an increase in the magnitude of α. However, for the nonlinear theory a significant amount of this error can be introduced when we approximate equations like (3. 2) with equations like (3. 4). No similar approximation occurs in the linear perturbation theory,[7] and it is this fact that most clearly distinguishes the one from the other.

4. GENERAL NONLINEAR SPECIAL-FUNCTION OPERATORS

A. A Nonlinear Hypergeometric-like Equation

Consider again the nonlinear hypergeometric-like equation introduced in Sec. 1,

$$\left(\frac{d}{dx}(1-x^2)\frac{d}{dx} - 2(\mu + \nu x)\frac{d}{dx} + \alpha x^2 u_k + \lambda_k\right) u_k = 0, \tag{4.1}$$

with the first-order algebraic form

$$[p(1-q^2)p + 2i(\mu + \nu q)p - \alpha q^2 \overline{V}(q)\phi^{(k)} - \lambda^{(k)}] \times \psi^{(k)} = 0, \tag{4.2}$$

so that

$$\begin{aligned} B &= p(1-q^2)p + 2i(\mu + \nu q)p, \\ \Gamma &= 1, \\ \Delta &= -\alpha q^2, \\ Y &= \overline{V}(q)\phi^{(k)}. \end{aligned} \tag{4.3}$$

Since most of the special-function equations can be obtained from the hypergeometric equation, it is appropriate to feature this case as a general

example. The results may then be applied to many other equations through appropriate transformations, and different kinds of nonlinearities may be studied separately. There is, in fact, no essential reason beyond computational convenience why considerations need be restricted to equations of special-function type.

By procedures similar to those of Sec. 6 the following relations may be obtained for the unperturbed operator B:

(a) $B = p(1-q^2)p + 2i(\mu + \nu q)p$,
$\eta = (1-q)^{-(\mu+\nu)}(1+q)^{-(\nu-\mu)}$;

(b) $B = (M-\nu)(M+\nu+1)$;

(c) $J_1 = q = (J^{(1)} + J^{(2)})/(2M+1) - \mu\nu/[M(M+1)]$,
$J_2 = -i(1-q^2)p$
$= [J^{(1)}(M-\nu) - J^{(2)}(M+\nu+1)]/(2M+1)$
$+ \mu(M-\nu)(M+\nu+1)/[M(M+1)]$;

(d) $b^{(k)} = (m^{(1)} + k + \nu)(m^{(1)} + k - \nu - 1)$,
$m^{(1)2} + \mu^2\nu^2/m^{(1)2} = \mu^2 + \nu^2$;

(e) $J^{(1)}J^{(2)} = M^2 + \mu^2\nu^2/M^2 - \mu^2 - \nu^2$;

(f) $\overline{J}^{(1)} = J^{(2)}(2M-1)/(2M+1)$.

Details of computing the matrix elements of Δ and Y will be omitted since they also are analogous to those given in Sec. 3. To represent $\Delta = -\alpha q^2$, however, note that in this case

$$q^2 = R_1(M) + J^{(1)2}R_2(M) + R_2(M+1)J^{(2)2} + J^{(1)}R_3(M) + R_4(M)J^{(2)}, \tag{4.4}$$

where the following have been defined for notational convenience:

$$R_1(M) = J^{(1)}J^{(2)}/[(2M+1)(2M-1)] + h^{(2)}(M) \times h^{(1)}(M)\,[(2M+1)(2M+3)] + \mu^2\nu^2/[M^2(M+1)^2], \tag{4.5}$$

$$R_2(M) = [(2M+1)(2M+3)]^{-1}, \tag{4.6}$$

$$R_3(M) = -\mu\nu\{[M(M+1)]^{-1} + [(M+1)(M+2)]^{-1}\} \times (2M+1)^{-1}, \tag{4.7}$$

$$R_4(M) = -\mu\nu\{[M(M+1)]^{-1} + [(M+1)(M+2)]^{-1}\} \times (2M+3)^{-1}. \tag{4.8}$$

Thus we identify

$$\Delta^{(0)} = -\alpha R_1(M), \tag{4.9}$$

$$\Delta^{(1)} = \Delta^{(-1)} = -\alpha R_3(M)h^{(1)}(M), \tag{4.10}$$

$$\Delta^{(2)} = \Delta^{(-2)} = -\alpha R_2(M)h^{(1)}(M)h^{(1)}(M+1), \tag{4.11}$$

and by setting $M = m^{(k)}$, $m^{(k)} = m^{(1)} + k - 1$, calculate the matrix elements:

$$\Delta_{kk} = -\alpha R_1(m^{(k)}), \tag{4.12}$$

$$\Delta_{k\,k+1} = \Delta_{k+1\,k} = -\alpha R_3(m^{(k)})[m^{(k+1)2} + \mu^2\nu^2/m^{(k+1)2} - \mu^2 - \nu^2]^{1/2}[(2m^{(k)}+1)/(2m^{(k)}+3)]^{1/2}, \tag{4.13}$$

$$\Delta_{k\,k+2} = \Delta_{k+2\,k}, = -\alpha R_2(m^{(k)})[m^{(k+1)2} + \mu^2\nu^2/m^{(k+1)2} - \mu^2 - \nu^2]^{1/2}[m^{(k+2)2} + \mu^2\nu^2/m^{(k+2)2} - \mu^2 - \nu^2]^{1/2}[(2m^{(k)}+1)/(2m^{(k)}+5)]^{1/2}. \tag{4.14}$$

All elements not listed are to be taken as zero.

To represent the operator Y we begin by assuming the normalized eigenvectors of B to have the form of polynomials:

$$\phi^{(k)} = \sum_{n=0}^{N} C_n^{(k)} q^n, \tag{4.15}$$

where the $C_n^{(k)}$ are constants. When this is not the case the eigenvectors can always be approximated by polynomials through the use of a truncated Taylor series. For the first column,

$$Y = \phi^{(1)} = E^{(0)}(M) = C_0^{(1)}; \tag{4.16}$$

so the only nonzero element in this column is

$$Y_{11} = C_0^{(1)}. \tag{4.17}$$

The second column has

$$Y = \phi^{(2)} = C_0^{(2)} + C_1^{(2)}q = C_0^{(2)} + C_1^{(2)}\{-\mu\nu/[M(M+1)] + J^{(1)}(2M+1)^{-1} + (2M+3)^{-1}J^{(2)}\}, \tag{4.18}$$

from which we identify

$$E^{(0)}(M) = C_0^{(2)} - C_1^{(2)}\{\mu\nu/[M(M+1)]\}, \quad E^{(-1)}(M) = C_1^{(2)}(2M+1)^{-1}. \tag{4.19}$$

Thus, the matrix elements are

$$Y_{12} = E^{(-1)}(m^{(1)})h^{(1)}(m^{(1)}) = C_1^{(2)}[m^{(2)2} + \mu^2\nu^2/m^{(2)2} - \mu^2 - \nu^2]^{1/2} \times [(2m^{(1)}+1)(2m^{(1)}+3)]^{-1/2}, \tag{4.20}$$

$$Y_{22} = E^{(0)}(m^{(2)}) = C_0^{(2)} - C_1^{(2)}(\mu\nu/m^{(2)}m^{(3)}), \tag{4.21}$$

$$Y_{32} = E^{(-1)}(m^{(2)})h^{(1)}(m^{(2)}) = C_1^{(2)}[m^{(3)2} + \mu^2\nu^2/m^{(3)2} - \mu^2 - \nu^2]^{1/2} \times [(2m^{(2)}+1)(2m^{(2)}+3)]^{-1/2}. \tag{4.22}$$

Subsequent columns can be computed in the same manner.

B. Other Nonlinearities

Theoretically, matrix representations can be found for any nonlinear perturbation for which a Taylor series expansion like (2.18) exists. Listed below are a few examples specially selected to illustrate other forms of nonlinearities. In addition to the first few matrix elements of the perturbations, all essential operator relationships are provided.

1. Nonlinear Legendre-like Harmonics

$$\frac{d}{dx}(1-x^2)\frac{du_k}{dx} + 2\alpha(1-x^2)\frac{du_k}{dx}u_k + \lambda_k u_k = 0,$$
$$A = p(1-q^2)p - i\alpha(1-q^2)p\bar{V}(q)\phi^{(k)},$$
$$B = p(1-q^2)p = M(M+1),$$
$$\Gamma = 1,$$
$$\Delta = -i\alpha(1-q^2)p,$$
$$Y = \bar{V}(q)\phi^{(k)},$$
$$q = (J^{(1)} + J^{(2)})/(2M+1),$$
$$-i(1-q^2)p = [J^{(1)}M - J^{(2)}(M+1)]/(2M+1),$$
$$J^{(1)}J^{(2)} = M^2, \bar{J}^{(1)} = J^{(2)}(2M-1)/(2M+1),$$
$$b^{(k)} = k(k-1), \quad m^{(k)} = k-1,$$
$$\phi^{(k)} = (k-\tfrac{1}{2})^{1/2}P_{k-1}(q),$$

where $P_j(q)$ is the jth Legendre polynomial,

$$\Delta_{k+1\,k} = \alpha k(k-1)/[(2k-1)(2k+1)]^{1/2},$$
$$\Delta_{k\,k+1} = -\alpha k(k+1)/[(2k-1)(2k+1)]^{1/2}.$$

The matrix elements of Y are exactly those given in Sec. 3 for the nonlinear Legendre-like equation discussed there.

2. Nonlinear Hermite-like Harmonics

$$\frac{d^2u_k}{dx^2} - 2x\frac{du_k}{dx} + \alpha x u_k^3 + \lambda_k u_k = 0,$$
$$A = p^2 + 2iqp - \alpha q\bar{V}(q)\phi^{(k)2},$$
$$B = p^2 + 2iqp = 2M,$$
$$\Gamma = 1,$$
$$\Delta = -\alpha q,$$
$$Y = \bar{V}(q)\phi^{(k)2},$$
$$q = \tfrac{1}{2}(J^{(1)} + J^{(2)}),$$
$$ip = J^{(2)},$$
$$J^{(1)}J^{(2)} = 2M, \quad \bar{J}^{(1)} = J^{(2)},$$
$$b^{(k)} = 2(k-1), \quad m^{(k)} = k-1,$$
$$\phi^{(k)} = [2^{k-1}\pi^{1/2}(k-1)!]^{-1/2}H_{k-1}(q),$$

where $H_j(q)$ is the jth Hermite polynomial,

$$\Delta_{k+1\,k} = \Delta_{k\,k+1} = -\alpha(\tfrac{1}{2}k)^{1/2},$$
$$Y_{11} = (1/\pi)^{1/2}, \quad Y_{12} = 0, \quad \cdots,$$
$$Y_{21} = 0, \quad Y_{22} = (9/\pi)^{1/2}, \quad \cdots,$$
$$Y_{31} = 0, \quad Y_{32} = 0, \quad \cdots,$$
$$\cdots, \quad \cdots, \quad \cdots,$$

It should be noted that the unperturbed operator $B = p^2 + 2iqp$ is related by a similarity transformation to the familiar harmonic oscillator operator $B' = p^2 + q^2$.

3. Nonlinear Laguerre-like Harmonics

$$x\frac{d^2u_k}{dx^2} + (1-x)\frac{du_k}{dx} + \alpha x u_k \sin(u_k) + \lambda_k u_k = 0,$$
$$A = qp^2 - i(1-q)p - \alpha q\bar{V}(q)\sin(\phi^{(k)}),$$
$$B = qp^2 - i(1-q)p = M,$$
$$\Gamma = 1,$$
$$\Delta = -\alpha q,$$
$$Y = \bar{V}(q)\sin(\phi^{(k)}),$$
$$q = (J^{(1)} + J^{(2)}) + 2M + 1,$$
$$J^{(1)}J^{(2)} = M, \quad \bar{J}^{(1)} = J^{(2)},$$
$$b^{(k)} = k-1, \quad m^{(k)} = k-1,$$
$$\phi^{(k)} = L_{k-1}(q),$$

where $L_j(q)$ is the jth Laguerre polynomial,

$$\Delta_{kk} = -\alpha(2k-1),$$
$$\Delta_{k+1\,k} = \Delta_{k\,k+1} = -\alpha k^{1/2},$$
$$Y_{11} = \sin(1), \quad Y_{12} = \tfrac{1}{2}\sin(1), \quad \cdots,$$
$$Y_{21} = 0, \quad Y_{22} = 0 \quad \cdots,$$
$$Y_{31} = 0, \quad Y_{32} = -\tfrac{1}{2}\cos(1), \quad \cdots,$$
$$\cdots, \quad \cdots, \quad \cdots.$$

5. OTHER EXAMPLES

A. The Hartree Equations for the Helium Atom

The Hartree equations for the helium atom are classic examples of nonlinear equations arising because of interactions between particles. They follow from applying the theory of the self-consistent field to the Hamiltonian

$$H = -(\nabla_1^2 + \nabla_2^2) - \frac{4}{r_1} - \frac{4}{r_2} + \frac{2\alpha}{r_{12}} \tag{5.1}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the coordinates of the two electrons relative to the nucleus, $r_{12} = r_{21}$ represents the distance between the electrons, and α is a parameter related to the strength of the Coulomb repulsion between the electrons. Spin-dependent forces are neglected.

The time-independent wavefunction is written as

$$\psi(\mathbf{r}_1, \mathbf{r}_2) = u^{(1)}(\mathbf{r}_1)u^{(2)}(\mathbf{r}_2), \tag{5.2}$$

where

$$u^{(j)}(\mathbf{r}_j) \sim R_{nl}^{(j)}(r_j)Y_{lm}^{(j)}(\theta_j, \phi_j), \quad j = 1, 2 \tag{5.3}$$

designates the one-electron orbitals; r, θ, and ϕ are the standard spherical coordinates; the $Y_{lm}^{(j)}$ are the spherical harmonic functions; and the $R_{nl}^{(j)}$ are radial functions. This leads to

$$\left(-\frac{d^2}{dr_j^2}+\frac{l(l+1)}{r_j^2}-\frac{4}{r_j}+\frac{2\alpha}{r_j}S(r,n,l)\right)P_{nl}^{(j)} = E_{nl}^{(j)}P_{nl}^{(j)}, \tag{5.4}$$

for $j = 1, 2$,

$$P_{nl}^{(j)}(r_j) = r_j R_{nl}^{(j)}(r_j), \tag{5.5}$$

and

$$S(r_j,n,l) = \int_0^{r_j}[P_{nl}^{(i)}(r_i)]^2 dr_i + \int_{r_j}^{\infty}[P_{nl}^{(i)}(r_i)]^2 \times\left(\frac{r_j}{r_i}\right)dr_i, \tag{5.6}$$

where $i = 1, 2$, but $i \neq j$.

When $\alpha = 0$, Eq. (5.4) becomes a hydrogenic radial equation with the operational form

$$[p^2 + l(l+1)q^{-2} - 4q^{-1} - b^{(k)}]U^{(k)} = 0. \tag{5.7}$$

The operator $B = p^2 + l(l+1)q^{-2} - 4q^{-1}$ does not lend itself to the calculation of the step operators $J^{(1)}$ and $J^{(2)}$ as they have been defined; this is a consequence of the fact that the discrete eigenvalues of B are bounded above.

We consider the transformation

$$Q = (-b^{(k)})^{1/2}q, \quad P = (-b^{(k)})^{-1/2}p, \tag{5.8}$$

$QP - PQ = i$, which transforms the operator B to

$$B' = QP^2 + Q + l(l+1)Q^{-1}, \tag{5.9}$$

with the associated eigenequation

$$B'U^{(k)} = 4(-b^{(k)})^{-1/2}U^{(k)}. \tag{5.10}$$

The eigenvectors of B' in the coordinate Q are the same as for B in the coordinate q.

Following the procedure of Sec. 2 the step operators of B' are determined to be

$$J^{(1)} = Q - iQP - M \tag{5.11}$$

and

$$J^{(2)} = Q + iQP - M, \tag{5.12}$$

from which it follows that:

$$\begin{aligned} &B' = 2M, \\ &J_1 = Q = \tfrac{1}{2}(J^{(1)} + J^{(2)}) + M, \\ &J_2 = -iQP = \tfrac{1}{2}(J^{(1)} - J^{(2)}), \\ &J^{(1)}J^{(2)} = M(M-1) - l(l+1), \bar{J}^{(1)} = J^{(2)}, \\ &b'^{(k)} = 4(-b^{(k)})^{-1/2} = 2(k+l), \\ &U^{(k)} = (2Q)^{l+1}e^{-Q}L_{n+l}^{2l+1}(2Q) \end{aligned} \tag{5.13}$$

where $n = k + l$ and the $L_j^m(x)$ are generalized Laguerre polynomials. Thus, the eigenvalues and normalized eigenvectors of B are given by

$$b^{(k)} = -4/n^2, \quad n = 1, 2, \cdots, \tag{5.14}$$

$$\phi^{(k)} = \left(\frac{2(n-l-1)!}{n^2[(n+l)!]^3}\right)^{1/2}\left(\frac{4q}{n}\right)^{l+1}e^{-2q/n}L_{n+l}^{2l+1}\left(\frac{4q}{n}\right). \tag{5.15}$$

The perturbing operators Δ and Y for Eq. (5.4) also have similar forms, defined by

$$\Delta = 2\alpha q^{-1} \tag{5.16}$$

and

$$Y = \bar{V}(q)S(q,n,l). \tag{5.17}$$

Because the discrete eigenvalues $b^{(k)}$ have the upper limit zero, the discrete eigenvectors $\phi^{(k)}$ of the operator B do not form a complete system of orthonormal vectors and, accordingly, do not provide a suitable basis for a matrix representation. The eigenvectors associated with the operator B' do form a complete orthonormal system, however, and this set may be utilized for matrix representations, even though the operator B will no longer be diagonal.

To calculate the ground state energy of the helium atom we seek to diagonalize the matrix

$$A = B + \Delta Y, \tag{5.18}$$

where B represents the energy of one electron due only to the Coulomb potential of the nucleus, and ΔY expresses the energy of interaction of the electrons. To the first eigenvalue of A must be added the energy $b^{(1)}$ of the other electron due to the potential of the nucleus alone. Our matrix representation will be based upon the condition that $l = 0$, even though this is not entirely correct whenever $n > 1$; there is a degeneracy in the l quantum number but the major contribution to the energy occurs for $l = 0$. Also, for $l = 0$ the operator Y can be approximated by[8]

$$Y = 1 - (1 + 2q/n)e^{-4q/n}. \tag{5.19}$$

Because the form of the perturbation does not lend itself to the algebraic method of computing matrix elements, we have in this case utilized the analytical technique discussed in Sec. 6. Matrix representations for B and ΔY can then be defined by

$$B_{jk} = \int_0^{\infty}\Phi^{(j)}B\Phi^{(k)}dq \tag{5.20}$$

and

$$(\Delta Y)_{jk} = 2\alpha\int_0^{\infty}\Phi^{(j)}q^{-1}S(q,k,0)\Phi^{(k)}dq, \tag{5.21}$$

where $\Phi^{(k)}$ represents the normalized eigenvectors of the operator B'. To be explicit, with $\alpha = 1$ the first matrix elements of (5.20) and (5.21) will be

$$B_{11} = -4\int_0^{\infty}e^{-2q}(q^2+2q)dq = -3 \tag{5.22}$$

and

$$(\Delta Y)_{11} = 8\int_0^{\infty}[e^{-2q}q - e^{-6q}(2q^2+q)]dq \cong 1.6296. \tag{5.23}$$

Continued calculation will yield the representations

$$B = \begin{bmatrix} -3.0000 & 1.1550 \\ 1.1550 & -1.6667 \end{bmatrix} \tag{5.24}$$

and

$$\Delta Y = \begin{bmatrix} 1.6296 & -0.3512 \\ -0.6842 & 1.0000 \end{bmatrix}. \tag{5.25}$$

The first eigenvalue of $A = B + \Delta Y$ is

$$E_{10}^{(1)} = -1.724 \text{ Ry}, \tag{5.26}$$

so that the total ground state energy becomes

$$E_{10} = E_{10}^{(1)} + b^{(1)} = -5.724 \text{ Ry}. \tag{5.27}$$

Compared with the experimental value of -5.807 Ry, this represents an error of 0.083 Ry. If larger matrix representations are desired, the approximation (5.19) probably should not be used, even though the calculation of Y from expressions like (5.6) will be tedious.

Millman and Keller[9] have computed the ground state energy for the helium atom by applying a modified form of the standard perturbation method to the Hartree equations. They obtained a value of -5.500 Ry, representing an error of 0.307 Ry with the experimental value. In addition, they made a comparison of results with those achieved by applying conventional perturbation techniques to the Schrödinger equation and showed that the eigenvalues agree up to first-order terms, but that the eigenvectors agree only up to zero-order terms.

By the present method, however, it can be shown that the matrix representation for the radial Schrödinger equation, whose Hamiltonian is given by (5.1), becomes

$$H_r = 2B + \Delta Y, \tag{5.28}$$

where B, Δ, and Y are the same as defined for the Hartree equations and, again, $l = 0$. Clearly, the eigenvalues of (5.28) will not be the same as those for the Hartree equations, since the sum of the eigenvalues of two matrices is not the same, in general, as the eigenvalues of the sum of those matrices.

B. The van der Pol Equation

An equation of frequent occurrence in the field of nonlinear oscillations is the van der Pol equation

$$u'' - \alpha(1 - u^2)u' + \lambda u = 0, \quad \lambda > 0, \tag{5.29}$$

where the primes indicate derivatives with respect to time. In nonlinear oscillation theory such equations are ordinarily treated as initial-value problems, where one seeks periodic solutions for a fixed value of λ, rather than as eigenvalue problems featuring definite boundary conditions. Because of this distinction, the matrix perturbation method featured here must be applied in a slightly different way.

To be precise, we look for solutions of (5.29) with $\lambda = 1$ satisfying

$$u(t + 2\pi/\omega) = u(t) \tag{5.30}$$

and subject to the initial conditions

$$u(0) = 2, \quad u'(0) = 0. \tag{5.31}$$

The existence of such solutions is guaranteed by the theorem of Liénard[1]; and stability requirements indicate that the amplitude of $u(t)$ varies between $+2$ and -2.

The unperturbed solution of (5.29), subject to the initial conditions (5.31), is given by

$$u(t) = 2\sin(\omega_0 t + \theta), \tag{5.32}$$

where $\omega_0 = \sqrt{\lambda} = 1$ is the angular frequency and $\theta = \frac{1}{2}\pi$ is the phase angle. The effect of the nonlinear term in the van der Pol equation is to change the angular frequency ω_0 to a new value ω, but this can be calculated by expanding $\sqrt{\lambda}$ in a Taylor series about $\alpha = 0$,

$$\sqrt{\lambda} = \omega_0 + \alpha\rho_1 + \cdots, \tag{5.33}$$

then setting $\lambda = 1$ to obtain

$$\omega \cong 1 - \alpha\rho_1, \tag{5.34}$$

where $\alpha\rho_1$ represents the first-order perturbation in the square root of the eigenvalues.

In order to acquire a complete set of eigenfunctions, from which appropriate matrix representations can be derived, we reformulate (5.29) as a boundary-value problem. Choosing $0 \le t \le \pi$ as the fundamental domain of the Hilbert space, the eigenvalues and normalized eigenfunctions of the unperturbed equation become

$$b_k = k^2 \tag{5.35}$$

and

$$v_k(t) = C\sin kt, \quad C = (2/\pi)^{1/2}. \tag{5.36}$$

As a first-order perturbation, the nonlinear operator of (5.29) may be identified as

$$N^0 = (1 - v_k^2)\frac{d}{dt}. \tag{5.37}$$

To compute the matrix elements of N^0 it is again most convenient to apply the method discussed in Sec. 6, writing

$$N_{jk}^0 = kC^2 \int_0^\pi [\sin jt(1 - C^2\sin^2 kt)\cos kt]dt. \tag{5.38}$$

Upon evaluation this yields

$$N_{kk}^0 = N_{3k\,k}^0 = 0, \tag{5.39}$$

and for $j \neq k$ or $3k$,

$$N_{jk}^0 = \frac{jkC^2}{4}[1 - (-1)^{j+k}]\left(\frac{4 - C^2}{j^2 - k^2} + \frac{C^2}{j^2 - 9k^2}\right), \tag{5.40}$$

Table III. Approximations of $u(t)$ and $u'(t)$ for the van der Pol equation obtained by numerical and matrix techniques.

$\alpha = 0.5$					$\alpha = 1.0$				
	Numerical method		Matrix method			Numerical method		Matrix method	
t	$u(t)$	$u'(t)$	$u(t)$	$u'(t)$	t	$u(t)$	$u'(t)$	$u(t)$	$u'(t)$
0.0	2.00	0.00	2.00	0.00	0.0	2.00	0.00	2.00	0.00
0.1	1.99	−0.19	1.99	−0.12	0.1	1.99	−0.17	2.00	−0.08
0.2	1.96	−0.34	1.98	−0.25	0.2	1.97	−0.30	1.98	−0.15
0.3	1.92	−0.48	1.94	−0.37	0.3	1.93	−0.40	1.97	−0.23
0.4	1.87	−0.60	1.90	−0.50	0.4	1.89	−0.47	1.94	−0.31
0.5	1.80	−0.71	1.84	−0.62	0.5	1.84	−0.53	1.90	−0.38
0.6	1.73	−0.80	1.78	−0.75	0.6	1.78	−0.59	1.86	−0.46
0.7	1.64	−0.89	1.69	−0.88	0.7	1.72	−0.64	1.81	−0.53
0.8	1.55	−0.98	1.60	−1.00	0.8	1.65	−0.68	1.76	−0.61
0.9	1.45	−1.07	1.49	−1.13	0.9	1.58	−0.73	1.69	−0.70
1.0	1.33	−1.15	1.38	−1.25	1.0	1.51	−0.78	1.62	−0.79
1.1	1.21	−1.24	1.24	−1.37	1.1	1.43	−0.83	1.53	−0.88
1.2	1.09	−1.34	1.10	−1.49	1.2	1.34	−0.89	1.44	−0.99
1.3	0.95	−1.44	0.95	−1.60	1.3	1.25	−0.96	1.33	−1.10
1.4	0.80	−1.54	0.78	−1.71	1.4	1.15	−1.04	1.22	−1.23
1.5	0.64	−1.65	0.60	−1.81	1.5	1.04	−1.12	1.09	−1.36
1.6	0.47	−1.77	0.42	−1.90	1.6	0.92	−1.23	0.95	−1.49
1.7	0.28	−1.88	0.23	−1.97	1.7	0.80	−1.35	0.79	−1.63
1.8	0.09	−2.00	0.03	−2.02	1.8	0.65	−1.49	0.62	−1.77
1.9	−0.11	−2.10	−0.18	−2.06	1.9	0.50	−1.65	0.44	−1.90
2.0	−0.33	−2.18	−0.38	−2.07	2.0	0.32	−1.83	0.24	−2.01
2.1	−0.55	−2.22	−0.59	−2.05	2.1	0.13	−2.04	0.04	−2.11
2.2	−0.77	−2.22	−0.79	−2.01	2.2	−0.08	−2.25	−0.18	−2.18
2.3	−0.99	−2.15	−0.99	−1.93	2.3	−0.32	−2.46	−0.40	−2.21
2.4	−1.20	−2.02	−1.18	−1.82	2.4	−0.58	−2.62	−0.62	−2.21
2.5	−1.39	−1.83	−1.35	−1.68	2.5	−0.84	−2.68	−0.84	−2.17
2.6	−1.56	−1.58	−1.51	−1.51	2.6	−1.11	−2.59	−1.05	−2.08
2.7	−1.71	−1.29	−1.66	−1.31	2.7	−1.35	−2.34	−1.25	−1.95
2.8	−1.82	−1.00	−1.77	−1.08	2.8	−1.57	−1.95	−1.44	−1.76
2.9	−1.91	−0.71	−1.87	−0.83	2.9	−1.74	−1.48	−1.61	−1.54
3.0	−1.96	−0.43	−1.94	−0.57	3.0	−1.87	−1.02	−1.75	−1.27
3.1	−1.99	−0.19	−1.98	−0.29	3.1	−1.95	−0.62	−1.86	−0.97
3.2	−2.00	0.02	−2.00	−0.01	3.2	−1.99	−0.29	−1.94	−0.65
3.3	−1.99	0.20			3.3	−2.00	−0.04	−1.99	−0.30
3.4	−1.96	0.36			3.4	−2.00	0.14	−2.00	0.05
3.5	−1.92	0.50			3.5	−1.98	0.28		

where $C^2 = 2/\pi$.

Diagonalization of the matrix $A_{jk} = b_j \delta_{jk} + \alpha N^0_{jk}$ will produce an infinite set of eigenvalues and associated eigenvectors, from which solutions of the van der Pol equation may be extracted. The first eigenvalues are

$$\lambda^{(1)} = 1.0381 \tag{5.41}$$

and

$$\lambda^{(1)} = 1.1502 \tag{5.42}$$

for $\alpha = 0.5$ and $\alpha = 1$, respectively. From these we determine

$$\alpha\rho_1 = 0.019 \tag{5.43}$$

and

$$\alpha\rho_1 = 0.072. \tag{5.44}$$

Hence, by applying (5.34), the perturbed angular frequencies become

$$\omega = 0.981, \quad \alpha = 0.5 \tag{5.45}$$

and

$$\omega = 0.928, \quad \alpha = 1. \tag{5.46}$$

The desired solution can be expressed in the form

$$u(t) = D\Sigma_k c_k \sin(k\omega t + \theta), \tag{5.47}$$

where the c_k have been determined along with the eigenvalues, and D and θ designate constants to be evaluated from the initial conditions (5.31) and the additional boundary condition

$$u(\pi/\omega) = -2. \tag{5.48}$$

This condition, consistent with stability requirements, is necessary for the complete formulation of a two-point boundary-value problem. The solutions corresponding to $\alpha = 0.5$ and $\alpha = 1$ are given by

$$u(t) \cong 0.2083 + 1.9837 \cos\omega t - 0.2083 \cos 2\omega t + 0.0163 \cos 3\omega t \tag{5.49}$$

and

$$u(t) \cong 0.4011 + 1.9367 \cos\omega t - 0.4011 \cos 2\omega t + 0.0633 \cos 3\omega t. \tag{5.50}$$

These solutions provide values of $u(t)$ and $u'(t)$ which are at least as good as those obtained through the use of time-consuming numerical techniques.[10] The comparison is presented for half a period in Table III; by periodic extension the entire solution can be constructed.

It should be noted that the results for the matrix method were calculated from a first-order per-

turbation, whereas for the standard method the second-order perturbation must be included before a change in angular frequency is detected.

6. FINAL REMARKS

For applications where the nonlinear term is no larger than the linear terms ($\alpha \leqq 1$), first-order corrections obtained by the present method are probably sufficient. However, when the perturbing term becomes larger ($\alpha > 1$), higher-order corrections to the unperturbed solution may be necessary in the interests of accuracy. A second-order perturbation will appear in the form of another operator added to A:

$$A = B + \Delta(Y + \Omega), \tag{6.1}$$

where B still designates the unperturbed linear operator, ΔY is the first-order nonlinear perturbation of B, and $\Delta\Omega$ represents the second-order perturbation. Here Γ is assumed to be unity, since it plays no important part in the following discussion.

To obtain the explicit form of the operator Ω from which its matrix elements can be calculated, it is desirable to reconsider the general nonlinear Eq. (2.2),

$$(B + \Delta Z - \lambda^{(k)})\psi^{(k)} = 0. \tag{6.2}$$

As before, the operator Z may be expanded in a Taylor series about $\alpha = 0$:

$$Z = Y + \alpha Z_1 + \alpha^2 Z_2 + \cdots, \tag{6.3}$$

where again

$$Z_1 = \frac{\partial Z}{\partial \psi^{(k)}} \psi_1^{(k)} \Big|_{\alpha=0}, \cdots. \tag{6.4}$$

Eq. (6.2) will then take the form

$$[B + \Delta(Y + \alpha Z_1 + \alpha^2 Z_2 + \cdots) - \lambda^{(k)}]\psi^{(k)} = 0 \tag{6.5}$$

or, retaining both Y and the new term αZ_1,

$$[B + \Delta(Y + \alpha Z_1) - \lambda^{(k)}]\psi^{(k)} = 0, \tag{6.6}$$

so that $\Omega = \alpha Z_1$.

Assuming the first-order perturbation problem to be solved, $\psi_1^{(k)}$ will be known; hence, Z_1 will also be known and we write

$$\Omega = \sum_{l \neq k} \gamma_l^{(k)} \phi^{(l)} \frac{\partial Z}{\partial \psi^{(k)}} \Big|_{\alpha=0}, \tag{6.7}$$

where $\gamma_l^{(k)}$ represents the vector components of $\psi^{(k)}$ (normalized in such a way that $\gamma_k^{(k)} = 1$), and use of the expansion

$$\psi^{(k)} = \phi^{(k)} + \alpha\psi_1^{(k)} = \phi^{(k)} + \sum_{l \neq k} \gamma_l^{(k)} \phi^{(l)} \tag{6.8}$$

for the first-order terms is implied. A matrix representation for Ω can now be obtained in much the same way as that for Y, and A expressed as

$$A_{jk} = b^{(j)}\delta_{jk} + \Sigma_n \Delta_{jn}(Y_{nk} + \Omega_{nk}). \tag{6.9}$$

There is, of course, one form of Z which simplifies the computations required to obtain the matrix elements of Ω. This occurs when $Z = \bar{V}(q)\psi^{(k)}$, so that

$$\Omega = \sum_{l \neq k} \gamma_l^{(k)} \bar{V}(q) \phi^{(l)}; \tag{6.10}$$

in particular,

$$\Omega_{1k} = 0,$$
$$\Omega_{j1} = \gamma_j^{(1)}/\sqrt{2}, \quad j \neq 1,$$
$$\Omega_{kk} = \sum_{l \neq k} \gamma_l^{(k)} Y_{lk}. \tag{6.11}$$

Some of the matrix elements of Ω for the nonlinear Legendre-like equation discussed in Sec. 3 are displayed below; for the case when $\alpha = 1$:

$$\Omega = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 0 & -0.0592 & 0 & \cdots \\ 0.0264 & 0 & -0.0303 & \cdots \\ & \cdots & & \end{bmatrix}. \tag{6.12}$$

This leads to a first eigenvalue of

$$\lambda_1 = -0.2579$$

and a first eigenfunction

$$u_1(x) = v_1(x) + 0.0396\, v_3(x),$$

compared with

$$\lambda_1 = -0.2495$$

and

$$u_1(x) = v_1(x) + 0.0373\, v_3(x)$$

for a first-order perturbation and 3×3 matrix representation. Substituting these values into Eq. (3.1) gives

$$\left| \frac{d}{dx}(1 - x^2)\frac{du}{dx} 1 + \lambda_1 u_1 + x^2 u_1^2 \right| \leq 0.042, \quad |x| \leq 1$$

and

$$\left| \frac{d}{dx}(1 - x^2)\frac{du}{dx} 1 + \lambda_1 u_1 + x^2 u_1^2 \right| \leq 0.018, \quad |x| \leq 1$$

for first-order and second-order perturbations, respectively.

Following similar procedures, one can in principle obtain more and more accurate results by defining third-, fourth-, and higher-order perturbations. In practice, however, calculating the matrix elements in such cases may well take too long to be feasible.

It was shown in Sec. 5 that there are certain differential equations for which the algebraic tech-

nique of computing matrix elements is not particularly convenient. In these cases the usual integral method was applied. To understand the latter approach and its relation to the former, consider again the nonlinear Eq. (2. 1),

$$(L + \alpha N + \lambda)u = 0.$$

Assuming that $u = u(x)$ is also a continuous function of α, the Taylor expansion about $\alpha = 0$ may be formed,

$$u = u_0 + \alpha T_1 + \cdots , \tag{6. 13}$$

where u_0 represents the unperturbed solution and αT_1 is the first-order perturbation. Thus it follows that

$$N(x, u, \cdots) = N^0 + \alpha \frac{\partial N}{\partial u} T_1 + \cdots , \tag{6. 14}$$

defining $N^0 = N(x, u_0, \cdots)$. Substituting and retaining only the linear terms in α then reduces the basic equation to

$$(L + \alpha N^0 + \lambda)u = 0. \tag{6. 15}$$

Suppose that when $\alpha = 0$ this equation has the eigenvalues and normalized eigenfunctions

$$\lambda_0 = b_k \tag{6. 16}$$

and

$$u_0 = v_k(x). \tag{6. 17}$$

Expanding the perturbed eigenfunctions $u(x)$ in terms of the unperturbed eigenfunctions $v_k(x)$ yields

$$u = \sum_k c_k v_k , \tag{6. 18}$$

and substitution into (6. 15) gives

$$\sum_k (L + \alpha N^0 + \lambda)c_k v_k = 0. \tag{6. 19}$$

Multiplying (6. 19) by v_j and forming an inner product over the fundamental domain we get

$$-\sum_k [(v_j, Lv_k) + \alpha(v_j, N^0 v_k)]c_k = \lambda c_k \delta_{jk} \tag{6. 20}$$

or

$$-\sum_k (L_{jk} + \alpha N^0_{jk})c_k = \lambda c_j , \tag{6. 21}$$

where

$$L_{jk} = (v_j, Lv_k) \tag{6. 22}$$

and

$$N^0_{jk} = (v_j, N^0 v_k). \tag{6. 23}$$

However, L has the diagonal representation

$$L_{jk} = -b_j \delta_{jk} , \tag{6. 24}$$

so that (6. 21) can be written

$$\sum_k (b_j \delta_{jk} - \alpha N^0_{jk})c_k = \lambda c_j , \tag{6. 25}$$

and completing the diagonalization of the matrix $A_{jk} = b_j \delta_{jk} - \alpha N^0_{jk}$ will yield the perturbed eigenvalues λ. The matrix A referred to here is the same as that utilized for the development of the algebraic method. It is the form of the differential equation that will generally determine which method should be adopted for the computation of the matrix elements.

A comparison with standard perturbation theory can easily be made at this point. If, instead of multiplying (6. 19) by v_j, one multiplies by v_k and forms the inner product, (6. 20) will become

$$\sum_k [(v_k, Lv_k) + \alpha(v_k, N^0 v_k) + \lambda]c_k = 0 \tag{6. 26}$$

or

$$\sum_k [b_k - \alpha N^0_{kk} - \lambda]c_k = 0. \tag{6. 27}$$

Thus

$$\lambda = b_k - \alpha N^0_{kk} , \tag{6. 28}$$

which clearly defines the perturbed eigenvalues as the diagonal elements of the matrix A.

ACKNOWLEDGMENT

We would like to thank Professor H. S. Green of the Department of Mathematical Physics in the University of Adelaide, South Australia for reviewing and commenting on the present work.

---

* Based on a Ph.D. thesis submitted to the Department of Metallurgy, Mechanics, and Materials Science at Michigan State University by L. C. Andrews.

1 T. L. Saaty and J. Bram, *Nonlinear Mathematics* (McGraw-Hill, New York, 1964)

2 W. Heisenberg, *Introduction to the Unified Field Theory of Elementary Particles* (Interscience, New York, 1966).

3 L. de Broglie, *Introduction to the Vigier Theory of Elementary Particles*, translated by A. J. Knodel (Elsevier, New York, 1963).

4 L. Infeld and T. E. Hull, Rev. Mod. Phys. **23**, 21 (1951).

5 W. Miller, Jr., *Lie Theory and Special Functions* (Academic, New York, 1968).

6 J. D. Talman, *Special Functions* (Benjamin, New York, 1968).

7 H. S. Green and T. Triffet, J. Math. Phys. **10**, 1069 (1969).

8 J. C. Slater, *Quantum Theory of Atomic Structure* (McGraw-Hill, New York, 1960).

9 M. H. Millman and J. B. Keller, J. Math. Phys. **10**, 342 (1969).

10 H. T. Davis, *Introduction to Nonlinear Differential and Integral Equations* (Dover, New York, 1962).

Errata: On Green's Functions to the Bethe-Salpeter Equation

[J. Math. Phys. 11, 715 (1970)]

Justin C. Huang and Brian DeFacio

Department of Physics, University of Missouri–Columbia, Columbia, Missouri 65201

(Received 8 September 1971)

Figures 1 and 4 are reversed; however, the equations in the text describing choices of poles are all correct and explain which poles and contours were used in the analysis.

Figure 6 is incorrect for the contour used for the integral $L(-, +)$ and should close over the upper half-plane and enclose the top half of the branch cut and the branch point at $k = im$.

Equation (3.12) on page 717 should read as

$$N(R, \pm T) = \int_0^m \frac{k \exp[\pm i(m^2 - k^2)^{1/2} T]}{k^2 + q^2}\, dk. \qquad (3.12)$$

The conclusions of the paper remain unaltered after the above corrections.