Group theoretical basis of some identities for the generalized hypergeometric series

W. A. Beyer, J. D. Louck, and P. R. Stein
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 30 July 1986; accepted for publication 29 October 1986)

It is shown that Thomae's identity between two ${}_3F_2$ hypergeometric series of unit argument together with the trivial invariance under separate permutations of numerator and denominator parameters implies that the symmetric group S_5 is an invariance group of this series. A similar result is proved for the terminating Saalschützian ${}_4F_3$ series, where S_6 is shown to be the invariance group of this series (or S_5 if one parameter is eliminated by using the Saalschütz condition). Here Bailey's identity is realized as a permutation of appropriately defined parameters. Finally, the set of three-term relations between ${}_3F_2$ series of unit argument discovered by Thomae [J. Thomae, J. Reine Angew. Math. **87**, 26 (1879)] and systematized by Whipple [F. J. Whipple, Proc. London Math. Soc. **23**, 104 (1925)] is shown to be transformed into itself under the action of the group $S_6 \times \Lambda$, where Λ is a two-element group. The 12 left cosets of $S_6 \times \Lambda$ with respect to the invariance group S_5 are the structural elements underlying the three-term relations. The symbol manipulator MACSYMA was used to obtain preliminary results.

I. INTRODUCTION

The generalized hypergeometric function—a natural extension of Gauss's function [see Slater[1] (Chap. 2)]—has proved to be of interest, not only as a mathematical object, but also as a tool in physical applications. For example, the functions ${}_3F_2$ and ${}_4F_3$ with unit argument occur, respectively, in the definitions[2] (p. 429) of the Wigner coefficients and the Racah coefficients for SU(2). These functions have a high degree of symmetry in their parameters; this property was investigated systematically by Thomae in 1879,[3] who derived a two-term relation for ${}_3F_2$. This relation was rediscovered by Ramanujan[4] (p. 104), sometime before 1919. Thomae's work was reformulated by Whipple.[5] In addition to numerous papers on generalized hypergeometric series, Bailey[6] wrote an influential monograph in which he gave a two-term relation for the finite form of ${}_4F_3$, which still bears his name. Here the fourth numerator parameter is a negative integer, so that the infinite series becomes a rational function. (For this identity and the Saalschütz condition which must hold, see Sec. II.) Variations of Bailey's identity occur in the definition of the Wigner–Clebsch–Gordan coefficients of SU(3).[7]

For ${}_3F_2$ at least, there also exist three-term relations, known since the time of Thomae. These and the two-term relations mentioned above are (partially) given in tabular form in Slater's book.[1] Slater, however, does not discuss the invariance properties that underlie these relations, nor is the total number of possible relations established.

In the present paper we show that Thomae's two-term relation for ${}_3F_2$ and the invariance of the series to separate permutations of the numerator and denominator parameters may be subsumed under a single invariance group. Thus, we show that under a rescaling of the function [(2.6a) below] and a linear transformation of parameters [(3.4) below], the new function is invariant under all permutations of the new parameters, so that the symmetric group S_5 is the invariance group. A similar result is proved for the terminating Saalschützian ${}_4F_3$ series, where it is Bailey's two-term identity [(2.3) below] that leads to the symmetric group S_6 as the invariance group, or S_5 if one variable is eliminated by the Saalschütz condition [(3.11b) below]. For these two-term cases it is immediate that the number of distinct relations is 5! in each case.

The S_6 symmetry property obtained here for the ${}_4F_3$ series generalizes a result of Wilson.[8] He found that Bailey's identity implies an S_4 symmetry in the four parameters of a certain class of orthogonal polynomials that are functions of a variable t^2, these polynomials being defined in terms of the terminating Saalschützian ${}_4F_3$ series of unit argument. This S_4 symmetry was also found by Biedenharn and Lohe[9] in their generalization of Bailey's identity.

The set of three-term relations for ${}_3F_2$ turns out to be transformed into itself under the group $S_6 \times \Lambda$, where Λ is a two-element group. The 12 cosets of this group with respect to the invariance group S_5 are the key structural elements leading to the determination of all three-term relations. The number of distinct relations is $\binom{12}{3} = 220$; to get this result requires some work.

The symbolic manipulation computer program MACSYMA played a substantial role in obtaining and verifying many preliminary results, which led to direct proofs of many of the theorems.

II. THE BASIC EQUATIONS

A. Notational conventions

In general, ${}_3F_2$ and ${}_4F_3$ are, apart from parameters, functions of a single variable z. In fact, they are power series. In this paper, we shall always take $z = 1$. Here ${}_3F_2(z)$ does not converge at $z = 1$ unless the numerator parameters a,b,c and the denominator parameters d,e satisfy

$$\mathrm{Re}(d + e - a - b - c) > 0.$$

For ${}_4F_3(z=1)$ we consider only the "terminating" case where the numerator parameters are $a,b,c,-n$, with n a non-negative integer. We also require that the Saalschütz condition be fulfilled (see Sec. II B below). In our notation for the generalized hypergeometric series of unit argument, we display only the numerator and denominator parameters [cf. (2.1a) below], which are then the variables for this analysis.

In the sequel we write the ${}_3F_2$ parameters as a five-tuple $\mathbf{a}=(a,b,c,d,e)$. This five-tuple is treated as a column vector when operated on by an appropriate matrix, i.e., a linear transformation. Similarly for ${}_4F_3$, but here the integer parameter n is fixed, and is omitted from the corresponding six-tuple $\mathbf{a}=(a,b,c,d,e,f)$.

We shall use the following notation for permutation matrices. Let e_j denote the $n\times 1$ column vector with 1 in the jth row and 0 elsewhere. Let $i_1,i_2,\ldots,i_n$ be a permutation of $1,2,\ldots,n$. Then we define the symbol $[i_1,i_2,\ldots,i_n]$ to be the $n\times n$ matrix $[e_{i_1},e_{i_2},\ldots,e_{i_n}]$. The group of $n\times n$ permutation matrices will be denoted by P_n. For the subgroup consisting of the disjoint permutations $i_1,i_2,\ldots,i_m$ and $i_{m+1},i_{m+2},\ldots,i_n$ we write $P_{m,n-m}$. For example, a representative member of $P_{3,2}$ is

$$\begin{pmatrix}1&0&0&0&0\\0&0&1&0&0\\0&1&0&0&0\\0&0&0&0&1\\0&0&0&1&0\end{pmatrix}.$$

Finally, the symmetric group on n distinct objects will always be denoted by S_n.

B. Classical results

As remarked above, the first two-term relation for ${}_3F_2$ was given by Thomae[3] and rediscovered some years later by Ramanujan[4] (p. 104). In Bailey's notation it reads

$${}_3F_2\left(\begin{matrix}a&b&c\\d&e&\end{matrix}\right)=\frac{\Gamma(a')\Gamma(d)\Gamma(e)}{\Gamma(a)\Gamma(d')\Gamma(e')}\,{}_3F_2\left(\begin{matrix}a'&b'&c'\\d'&e'&\end{matrix}\right), \tag{2.1a}$$

where the variables

$$\mathbf{a}=(a,b,c,d,e), \tag{2.1b}$$

$$\mathbf{a}'=(a',b',c',d',e') \tag{2.1c}$$

are related by the linear transformation

$$\mathbf{a}'=t\mathbf{a} \tag{2.2a}$$

with

$$t=\begin{pmatrix}-1&-1&-1&1&1\\-1&0&0&0&1\\-1&0&0&1&0\\-1&-1&0&1&1\\-1&0&-1&1&1\end{pmatrix}. \tag{2.2b}$$

The second two-term relation is due to Bailey[6] (p. 56). In stating it, we use the notation

$$(a)_n=\Gamma(n+a)/\Gamma(a).$$

The relation is then

$${}_4F_3\left(\begin{matrix}a&b&c&-n\\d&e&f&\end{matrix}\right)=\frac{(d')_n(e')_n(f')_n}{(d)_n(e)_n(f)_n}\,{}_4F_3\left(\begin{matrix}a'&b'&c'&-n\\d'&e'&f'&\end{matrix}\right), \tag{2.3}$$

where the variables $\mathbf{a}$ and $\mathbf{a}'$, defined analogously to (2.1b) and (2.1c), are related by the linear transformation

$$\mathbf{a}'=b\mathbf{a} \tag{2.4a}$$

with

$$b=\begin{pmatrix}-1&0&0&0&0&1\\0&-1&0&0&0&1\\0&0&1&0&0&0\\-1&-1&0&0&1&1\\-1&-1&0&1&0&1\\0&0&0&0&0&1\end{pmatrix}. \tag{2.4b}$$

Relation (2.3) is valid only when n is a non-negative integer and the Saalschütz condition is fulfilled:

$$a+b+c-d-e-f-n+1=0. \tag{2.5}$$

In order to give the three-term relation for ${}_3F_2$ [Bailey[6] (p. 15)] in convenient form, we first define a rescaling of ${}_3F_2$ as follows:

$${}_3\hat{F}_2(\mathbf{a})={}_3F_2\left(\begin{matrix}a&b&c\\d&e&\end{matrix}\right)\times[\Gamma(d)\Gamma(e)\Gamma(d+e-a-b-c)]^{-1}, \tag{2.6a}$$

with $\mathbf{a}=(a,b,c,d,e)$. Next, we introduce new parameter column vectors $\mathbf{a}'$ and $\mathbf{a}''$. To define these in terms of $\mathbf{a}$ by linear transformations, we effectively extend the column vectors to six elements by adjoining the element 1 to each and writing

$$\begin{pmatrix}\mathbf{a}'\\1\end{pmatrix}=m_1\begin{pmatrix}\mathbf{a}\\1\end{pmatrix},\quad \begin{pmatrix}\mathbf{a}''\\1\end{pmatrix}=m_2\begin{pmatrix}\mathbf{a}\\1\end{pmatrix}, \tag{2.6b}$$

where m_1, m_2 are the matrices

$$m_1=\begin{pmatrix}1&0&0&0&0&0\\1&0&0&0&-1&1\\1&0&0&-1&0&1\\1&0&-1&0&0&1\\1&-1&0&0&0&1\\0&0&0&0&0&1\end{pmatrix},$$

$$m_2=\begin{pmatrix}0&0&1&-1&0&1\\0&0&1&0&-1&1\\0&0&1&0&0&0\\-1&0&1&0&0&1\\0&-1&1&0&0&1\\0&0&0&0&0&1\end{pmatrix}. \tag{2.6c}$$

The three-term relation then takes the form

$${}_3\hat{F}_2(\mathbf{a})=\alpha'(\mathbf{a}){}_3\hat{F}_2(\mathbf{a}')+\alpha''(\mathbf{a}){}_3\hat{F}_2(\mathbf{a}''). \tag{2.7}$$

The coefficients α', α'' in this expression are given by

$$\alpha'(\mathbf{a})=\frac{\pi\Gamma(1-b)}{\Gamma(d-a)\Gamma(e-a)\Gamma(c)\sin\pi(c-a)}, \tag{2.8a}$$

$$\alpha''(\mathbf{a})=\frac{-\pi\Gamma(1-b)}{\Gamma(d-c)\Gamma(e-c)\Gamma(a)\sin\pi(c-a)}. \tag{2.8b}$$

We note that the device used to write (2.6b)—familiar to geometers working with homogeneous coordinates—allows translations of parameters to be written as linear transformations. We shall return to this point in Sec. IV. Finally, we mention the trivial symmetries of ${}_3F_2$ and ${}_4F_3$ under appropriate permutations of the parameters. Reverting to the original definitions of $\mathbf{a}$ and $\mathbf{a}'$ for these two cases, namely, $\mathbf{a}=(a,b,c,d,e)$ and $\mathbf{a}=(a,b,c,d,e,f)$, respectively, with the $\mathbf{a}'$ similarly defined in terms of the primed parameters, we have

$$ {}_3F_2(\mathbf{a}) = {}_3F_2(\mathbf{a}'), \tag{2.9a} $$

$$ {}_4F_3(\mathbf{a}) = {}_4F_3(\mathbf{a}'). \tag{2.9b} $$

Here $\mathbf{a}'=p\mathbf{a}$ in both cases, with p a permutation matrix. In the first case, p belongs to the (permutation) representation $P_{3,2}$ of $S_3\times S_2$;in the second case, p belongs to the representation $P_{3,3}$ of $S_3\times S_3$.

III. GROUP STRUCTURE OF THE TWO-TERM IDENTITIES

The structure of the two relations (2.1a) and (2.3) becomes trivial when expressed in terms of the proper variables. To find these variables, the first step is to write down two matrices A_1 and A_2 which commute with all elements of $P_{3,2}$ and $P_{3,3}$ respectively, while also satisfying some other restrictions. The matrices A_1 and A_2 are far from unique; we shall write down a suitable pair *ad hoc*, leaving their derivation to the remarks at the end of the section.

Let

$$ A_1 = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{pmatrix}. \tag{3.1} $$

Then

$$ A_1^{-1} p A_1 = p, \quad p\in P_{3,2}, \tag{3.2a} $$

$$ A_1^{-1} t A_1 = [1,5,4,3,2]. \tag{3.2b} $$

Here t is defined by (2.2b). For the permutation matrix notation on the right, see Sec. II A. Equation (3.2a) is obvious in the form $pA_1 = A_1 p$ because each row permutation p and column permutation p has the same action on A_1; (3.2b) is also immediate on verifying the equality $tA_1 = A_1[1,5,4,3,2]$.

The group theoretical result we need to interpret (2.1a) is the following.

Theorem 3.1: The matrices in $P_{3,2}$ together with $[1,5,4,3,2]$ generate P_5.

Proof: Since $[3,1,2,5,4]$ and $[1,2,3,5,4]$ belong to $P_{3,2}$ the matrix $[3,1,2,5,4][1,5,4,3,2][1,2,3,5,4] = [3,4,5,1,2]$ is in the set of matrices generated by $P_{3,2}$ and $[1,5,4,3,2]$. Now $[3,4,5,1,2]$ written in cycle notation is just (13524). Similarly, $[3,2,1,4,5]$, which belongs to $P_{3,2}$, is just the two-cycle (13). As is well known, (13) and (13524) generate P_5 [see, for example, James and Kerber[10] (p. 5)]. Hence, the group generated by $P_{3,2}$ and $[1,5,4,3,2]$ contains P_5. But since all products of 5×5 permutation matrices are themselves 5×5 permutation matrices, we must obtain exactly P_5. ■

To interpret relation (2.1a) in terms of P_5 we define a new function ${}_3E_2$ by

$$ {}_3E_2(\mathbf{x}) = {}_3\hat{F}_2(A_1\mathbf{x}), \tag{3.3a} $$

with

$$ \mathbf{x} = (x,y,z,u,v). \tag{3.3b} $$

Thus the function ${}_3E_2$ is given in terms of ${}_3\hat{F}_2$ [see (2.6a)] by the change of variables

$$ \mathbf{a} = A_1\mathbf{x}. \tag{3.4} $$

Now relation (2.1a) takes the form

$$ {}_3E_2(x,y,z,u,v) = {}_3E_2(x,v,u,z,y) \tag{3.5a} $$

in consequence of (3.3a) and (3.3b). Moreover, the invariance of the original ${}_3F_2$ to separate permutations of the numerator and denominator parameters, and the invariance of the denominator in the definition of ${}_3\hat{F}_2$, is expressed as

$$ {}_3E_2(p\mathbf{x}) = {}_3E_2(\mathbf{x}), \quad p\in P_{3,2}. \tag{3.5b} $$

Theorem 3.1 and the invariance properties of ${}_3E_2$ given by (3.5a) and (3.5b) imply the following.

Theorem 3.2: The group P_5 is an invariance group of the function ${}_3E_2$, i.e.,

$$ {}_3E_2(p\mathbf{x}) = {}_3E_2(\mathbf{x}), \quad p\in P_5. \tag{3.6} $$

This result incorporates relation (2.1a) and the invariance of the original ${}_3F_2$ series under separate permutations (belonging to $P_{3,2}$) of the numerator and denominator parameters into a single relationship. It also extends this result to the group P_5: the function ${}_3E_2$ is invariant under all 5! permutations of the variables (x,y,z,u,v).

We next establish results analogous to Theorems 3.1 and 3.2 for the ${}_4F_3$ series relation given by (2.3). Define the nonsingular matrix A_2 by

$$ A_2 = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}. \tag{3.7} $$

Then

$$ A_2^{-1} p A_2 = p, \quad p\in P_{3,3}, \tag{3.8a} $$

$$ A_2^{-1} b A_2 = [2,1,6,5,4,3]. \tag{3.8b} $$

Here b is the matrix defined by (2.4b).

We can now prove the following result by an argument analogous to that used in Theorem 3.1.

Theorem 3.3: The matrices in $P_{3,3}$ together with the matrix $[2,1,6,5,4,3]$ generate P_6.

To obtain results analogous to Theorem 3.2, we first define the *polynomial* ${}_4Q_3(\mathbf{a})$ by

$$ {}_4Q_3(\mathbf{a}) = (d)_n (e)_n (f)_n \, {}_4F_3\left(\begin{matrix} a & b & c & -n \\ d & e & f & \end{matrix}\right). \tag{3.9} $$

Next we define the new polynomial ${}_4P_3(\mathbf{x})$ by making the change of variables

$$\mathbf{a} = A_2\mathbf{x}, \tag{3.10a}$$

where

$$\mathbf{x} = (x,y,z,u,v,w). \tag{3.10b}$$

Thus

$${}_4P_3(\mathbf{x}) = {}_4Q_3(A_2\mathbf{x}). \tag{3.10c}$$

Relations (2.3) and (2.5) are expressed in terms of the polynomials ${}_4P_3(\mathbf{x})$ and the variables $\mathbf{x}$, respectively, by

$${}_4P_3(x,y,z,u,v,w) = {}_4P_3(y,x,w,v,u,z), \tag{3.11a}$$

$$x+y+z+u+v+w+n-1=0. \tag{3.11b}$$

The invariance of ${}_4F_3(\mathbf{x})$ under separate permutations of numerator and denominator parameters is expressed by

$${}_4P_3(p\mathbf{x}) = {}_4P_3(\mathbf{x}), \quad p\in P_{3,3}. \tag{3.11c}$$

Theorem 3.3 and the invariance properties of ${}_4P_3$ given by (3.11a) and (3.11c) now imply the following.

Theorem 3.4: The group P_6 is an invariance group of the polynomials ${}_4P_3$; i.e.,

$${}_4P_3(p\mathbf{x}) = {}_4P_3(\mathbf{x}), \quad p\in P_6, \tag{3.12}$$

for all x,y,z,u,v,w that satisfy (3.11b).

This theorem extends the special results in (3.11a)–(3.11c) to the full permutation group P_6: for all (x,y,z,u,v,w) that satisfy $x+y+z+u+v+w+n-1=0$, the polynomial ${}_4P_3(x,y,z,u,v,w)$ is invariant under all 6! permutations of these variables. (Wilson[8] points out a lower symmetry of ${}_4F_3$, namely that under S_4.)

The operation of reversing the terminating Saalschützian ${}_4F_3$ series [Bailey[6] (p. 56)] is included in the group P_6. It is expressed in terms of the polynomials ${}_4P_3$ by

$${}_4P_3(x,y,z,u,v,w) = {}_4P_3(w,v,u,z,y,x) \tag{3.13}$$

for all (x,y,z,u,v,w) satisfying the Saalschütz condition.

Remarks: (a) There is considerable freedom in choosing the matrices A_1 and A_2. It is not difficult to prove that the most general matrices commuting with all elements of $P_{3,2}$ and $P_{3,3}$, respectively, are

$$C_1 = \begin{pmatrix} \alpha & \beta & \beta & \gamma' & \gamma' \\ \beta & \alpha & \beta & \gamma' & \gamma' \\ \beta & \beta & \alpha & \gamma' & \gamma' \\ \gamma & \gamma & \gamma & \alpha' & \beta' \\ \gamma & \gamma & \gamma & \beta' & \alpha' \end{pmatrix},$$

$$C_2 = \begin{pmatrix} \alpha & \beta & \beta & \gamma' & \gamma' & \gamma' \\ \beta & \alpha & \beta & \gamma' & \gamma' & \gamma' \\ \beta & \beta & \alpha & \gamma' & \gamma' & \gamma' \\ \gamma & \gamma & \gamma & \alpha' & \beta' & \beta' \\ \gamma & \gamma & \gamma & \beta' & \alpha' & \beta' \\ \gamma & \gamma & \gamma & \beta' & \beta' & \alpha' \end{pmatrix}. \tag{3.14}$$

Consider the determination of A_1. Since the trace of the matrix t is 1, we select any permutation matrix p with the properties $p\in P_5$, $p\notin P_{3,2}$, $\mathrm{tr}(p)=1$, and impose the condition $tC_1 = C_1p$ with C_1 nonsingular. To satisfy this condition, the permutation corresponding to p must belong to the class $(2^2,1)$ of S_5; moreover, each such p admits a solution C_1 when suitable restrictions on $\alpha,\beta,\gamma,\alpha',\beta',\gamma'$ are imposed. In particular, the choice $p = [1,5,4,3,2]$ in (3.2b) requires that $\alpha' = 2\alpha, \beta' = \gamma, \gamma' = \alpha, \gamma = \alpha+\beta$. Similarly, we find that the equation $bC_2 = C_2p$ with $p\in P_6$, $p\notin P_{3,3}$, $\mathrm{tr}(p)=0$, has a nonsingular solution C_2 if and only if p corresponds to a permutation in class (2^3) of S_6, and that each such p admits a solution C_2 for suitable restrictions on the parameters $\alpha,\beta,\gamma,\alpha',\beta',\gamma'$. In particular, the choice $p = [2,1,6,5,4,3]$ in (3.8b) requires that $\alpha' = \alpha+\beta$, $\beta' = 2\alpha$, $\gamma' = \alpha$, $\gamma = \alpha+\beta$. Thus, the most general nonsingular matrices C_1 and C_2, respectively, which can replace A_1 and A_2 in (3.2) and (3.8) are

$$C_1 = \begin{pmatrix} \alpha & \beta & \beta & \alpha & \alpha \\ \beta & \alpha & \beta & \alpha & \alpha \\ \beta & \beta & \alpha & \alpha & \alpha \\ \gamma & \gamma & \gamma & 2\alpha & \gamma \\ \gamma & \gamma & \gamma & \gamma & 2\alpha \end{pmatrix},$$

$$C_2 = \begin{pmatrix} \alpha & \beta & \beta & \alpha & \alpha & \alpha \\ \beta & \alpha & \beta & \alpha & \alpha & \alpha \\ \beta & \beta & \alpha & \alpha & \alpha & \alpha \\ \gamma & \gamma & \gamma & \gamma & 2\alpha & 2\alpha \\ \gamma & \gamma & \gamma & 2\alpha & \gamma & 2\alpha \\ \gamma & \gamma & \gamma & 2\alpha & 2\alpha & \gamma \end{pmatrix}, \tag{3.15}$$

where $\alpha\neq\beta$ and $\gamma = \alpha+\beta$.

(b) The invariance properties of the functions ${}_3E_2(\mathbf{x})$ and ${}_4P_3(\mathbf{x})$ under the action of the transformation groups P_5 and P_6, respectively, can also be realized in terms of the functions ${}_3\hat{F}_2(\mathbf{a})$ and ${}_4Q_3(\mathbf{a})$. Namely, one has

$${}_3\hat{F}_2(p'\mathbf{a}) = {}_3\hat{F}_2(\mathbf{a}), \quad p'\in A_1P_5A_1^{-1}, \tag{3.16a}$$

$${}_4Q_3(p'\mathbf{a}) = {}_4Q_3(\mathbf{a}), \quad p'\in A_2P_6A_2^{-1}. \tag{3.16b}$$

IV. GROUP STRUCTURES UNDERLYING THE THREE-TERM RELATION

In this section we set up the group theoretical apparatus which we use in Sec. V to derive further three-term relations from (2.7). We have shown that $A_1P_5A_1^{-1}$ is an invariance group of ${}_3\hat{F}_2(\mathbf{a})$, while its isomorph P_5 is an invariance group of ${}_3\hat{F}_2(A_1\mathbf{x})$. Clearly, these isomorphic invariance groups will have an important role in our treatment of (2.7).

All this suggests that we look for a simple structure underlying (2.7) which is similar to what we found in Sec. III. The matrices m_1 and m_2 are, however, six dimensional. We therefore extend the matrices $p\in P_{3,2}$, t, and A_1 to 6×6 form as follows:

$$p\mapsto\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}, \quad t\mapsto\begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix}, \quad A_1\mapsto\begin{pmatrix} A_1 & 0 \\ 0 & 1 \end{pmatrix}; \tag{4.1}$$

here 0 is a column or row matrix consisting of five zeros. Under this substitution, (3.2a) and (3.2b) become

$$\begin{pmatrix} A_1 & 0 \\ 0 & 1 \end{pmatrix}^{-1}\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} A_1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}, \quad p\in P_{3,2} \tag{4.2a}$$

and

$$\begin{pmatrix} A_1 & 0 \\ 0 & 1 \end{pmatrix}^{-1}\begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} A_1 & 0 \\ 0 & 1 \end{pmatrix} = [1,5,4,3,2,6]. \tag{4.2b}$$

Theorem 3.1 is still valid if we replace the group $P_{3,2}$ by the group

$$\left\{\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} \middle| p \in P_{3,2}\right\}, \tag{4.3}$$

replace the permutation matrix [1,5,4,3,2] by [1,5,4,3,2,6], and use $P_{5,1}$ in place of P_5.

The natural next step would be to transform the matrices m_1 and m_2 as we transformed $\left(\begin{smallmatrix} p & 0 \\ 0 & 1 \end{smallmatrix}\right)$ above. Unfortunately, this procedure leads to complicated results. To obtain the simple structure we are looking for, we must recognize, first, that the similarity transformation in (4.2a) and (4.2b) is not the only one that leaves those relations invariant, and second, that it is the matrix $m_3 = m_1 m_2$ that we should transform.

The matrix m_3 is given by

$$m_3 = \begin{pmatrix} 0 & 0 & 1 & -1 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 2 \\ 0 & 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4.4}$$

We note that the set of matrices

$$M = \{I, m_1, m_2, m_3\} \tag{4.5}$$

is an Abelian group of involutions.

We find that there exists a nonsingular matrix A such that

$$A^{-1}\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} A = \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}, \quad p \in P_{3,2}, \tag{4.6a}$$

$$A^{-1}\begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix} A = [1,5,4,3,2,6], \tag{4.6b}$$

$$A^{-1} m_3 A = [3,2,1,6,5,4]. \tag{4.6c}$$

The choice of the permutation matrix p' into which m_3 is transformed is narrowed by the requirements $p' \in P_6$, $p' \notin P_{5,1}$, $\mathrm{tr}(p') = 2$, which imply that p' belongs either to the class $(1^2,2^2)$ or to $(1^2,4)$. Taking p' to belong to the second class fails to produce a solution. With $p' \in (1^2,2^2)$, we construct the required A by bordering A_1 with a row and column so that conditions (4.6a)–(4.6c) are satisfied. This determines A up to a multiplicative constant, and we have

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 2 & 1 & 0 \\ 1 & 1 & 1 & 1 & 2 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \tag{4.7}$$

$$A^{-1} = \frac{1}{3}\begin{pmatrix} 1 & -2 & -2 & 1 & 1 & 0 \\ -2 & 1 & -2 & 1 & 1 & 0 \\ -2 & -2 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & -2 & 0 \\ 1 & 1 & 1 & -2 & 1 & 0 \\ 1 & 1 & 1 & -2 & -2 & 1 \end{pmatrix}.$$

We summarize these results in the following theorem.

Theorem 4.1: The set of matrices

$$\left\{\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} \middle| p \in P_{3,2}\right\}, \quad [1,5,4,3,2,6], \tag{4.8a}$$

generates the permutation group $P_{5,1}$. Then $P_{5,1}$ and the matrix [3,2,1,6,5,4] generate P_6. Equivalently, the set of matrices

$$\left\{\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} \middle| p \in P_{3,2}\right\}, \quad \begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix}, \tag{4.8b}$$

generate the group $A\,P_{5,1}A^{-1}$. This group and the matrix m_3 generate $A\,P_6 A^{-1}$.

Proof: The only result not already proven is that $P_{5,1}$ and [3,2,1,6,5,4] generate P_6. But this follows easily from the same argument used to establish Theorem 3.1. ■

The reason for choosing m_3, instead of m_1 or m_2, to determine the matrix A, is found in the relations

$$A^{-1} m_1 A = \lambda[6,5,4,3,2,1], \tag{4.9a}$$

$$A^{-1} m_2 A = \lambda[4,5,6,1,2,3], \tag{4.9b}$$

where λ is an involution defined by

$$\lambda = \frac{1}{3}\begin{pmatrix} -2 & 1 & 1 & 1 & 1 & 1 \\ 1 & -2 & 1 & 1 & 1 & 1 \\ 1 & 1 & -2 & 1 & 1 & 1 \\ 1 & 1 & 1 & -2 & 1 & 1 \\ 1 & 1 & 1 & 1 & -2 & 1 \\ 1 & 1 & 1 & 1 & 1 & -2 \end{pmatrix}. \tag{4.10}$$

It is easy to see that λ commutes with every element of P_6. This implies that no nonsingular matrix B exists such that $B m_1 B^{-1} = p \in P_6$, since then we would have $p(BA) = [6,5,4,3,2,1]\lambda(BA)$, and therefore, $(BA)^{-1} p'(BA) = \lambda$ for $p' \in P_6$. But since $\mathrm{tr}(\lambda) = -4$ and $\mathrm{tr}(p') \geqslant 0$, we have a contradiction; this proves the nonexistence of B. In the same fashion, we can show that m_2 is not similar to any $p \in P_6$.

We are now ready to prove the following.

Theorem 4.2: The set of matrices

$$\left\{\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} \middle| p \in P_{3,2}\right\}, \quad [1,5,4,3,2,6], \ \lambda\,[6,5,4,3,2,1]$$

generates the direct product group of 1440 elements

$$P_6 \times \Lambda = \{P_6, \lambda P_6\}; \tag{4.11}$$

here Λ is the two-element group $\Lambda = \{I, \lambda\}$. The group $A^{-1}MA$ [with M defined by (4.5)] is a subgroup of $P_6 \times \Lambda$. Equivalently, the set of matrices

$$\left\{\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} \middle| p \in P_{3,2}\right\}, \quad \begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix}, \quad m_1,$$

generates the group $A(P_6 \times \Lambda)A^{-1}$, which contains the subgroup M.

Proof: By Theorem 4.1, the set of matrices

$$\left\{\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} \middle| p \in P_{3,2}\right\}, \quad [1,5,4,3,2,6],$$

generates the group $P_{5,1}$. We see that the matrix $A^{-1} m_2 A = \lambda[4,5,6,1,2,3]$ is obtained from the matrix $A^{-1} m_1 A = \lambda[6,5,4,3,2,1]$ by multiplying from the right and left by $[3,2,1,4,5,6] \in P_{5,1}$. Because $m_3 = m_1 m_2$, we have $A^{-1} m_3 A$

$= (A^{-1}m_1A)(A^{-1}m_2A) = [3,2,1,6,5,4]$. This matrix and the group $P_{5,1}$ generate P_6 (Theorem 4.1). The matrix λ is now obtained from $\lambda = (A^{-1}m_1A)[6,5,4,3,2,1]$, and is seen to commute with all elements of P_6. This implies that no further matrices, not in $P_6 \times \Lambda$, can be generated by the matrix set specified in the statement of the theorem. ■

With the application of Theorem 4.2 in mind, we digress here to consider some problems of notation not dealt with in the Introduction. In the rest of the paper we shall frequently be acting on five-tuples such as $\mathbf{a} = (a,b,c,d,e)$ and $\mathbf{x} = (x,y,z,u,v)$ with 6×6 matrices. We recall that $\mathbf{a}$ and $\mathbf{x}$ are the coordinates on which the functions ${}_3E_2$ [(3.5a)] and ${}_3\hat{F}_2$ [(2.7)] are defined. We shall need a special notation for this action. Before introducing an appropriate symbol, we write down some abbreviations for quantities already defined, viz.,

$$G = P_6 \times \Lambda, \quad G_A = AGA^{-1}, \qquad H = P_{5,1}, \quad H_A = AHA^{-1}. \tag{4.12}$$

These definitions will be used from here on.

Let $SL(6,\mathbf{R})$ denote the group of 6×6 nonsingular matrices over $\mathbf{R}$, and let $\mathbf{R}^n$ denote the set of all n-tuples. Let $\mathbf{y} = (y_1,y_2,y_3,y_4,y_5) \in \mathbf{R}^5$ and let $\mathbf{z} = (\mathbf{y},y_6) \in \mathbf{R}^6$ denote any six-tuple with projection π on the first five coordinates given by $\pi\mathbf{z} = \mathbf{y}$ and with y_6 a given function f of $\mathbf{y}$, i.e., $y_6 = f(\mathbf{y})$. We now define the mapping $SL(6,\mathbf{R}): \mathbf{R}^5 \to \mathbf{R}^5$ by the following rule. For $B \in SL(6,\mathbf{R})$ and $\mathbf{y} \in \mathbf{R}^5$, the action of B on $\mathbf{y}$ is denoted by $B\circ\mathbf{y}$, and $B\circ\mathbf{y} \in \mathbf{R}^5$ is defined to be

$$B\circ\mathbf{y} = \pi(B\mathbf{z}) = \mathbf{y}'; \tag{4.13}$$

here $B\mathbf{z}$ is ordinary 6×6 on 6×1 matrix multiplication. This action of $SL(6,\mathbf{R})$ on $\mathbf{R}^5$ satisfies the usual axioms.

Let us take the elements of $\mathbf{R}^5$ to be $\mathbf{a} = (a,b,c,d,e)$ and apply (4.13) to the mappings $G_A: \mathbf{R}^5 \to \mathbf{R}^5$. This gives

$$g_A\circ\mathbf{a} = \mathbf{a}', \quad g_A \in G_A, \quad \mathbf{a} \in \mathbf{R}^5, \tag{4.14}$$

where $\mathbf{a}'$ is obtained from $\mathbf{b} = (\mathbf{a},1) \in \mathbf{R}^6$ by the rule given in (2.6b). In particular, since each g_A has (000001) as its sixth row, the sixth component of $g_A\mathbf{b}$ is also 1.

The three-term relation (2.7), rewritten in terms of the new notation, reads

$${}_3\hat{F}_2(\mathbf{a}) = \alpha'(\mathbf{a})\,{}_3\hat{F}_2(m_1\circ\mathbf{a}) + \alpha''(\mathbf{a})\,{}_3\hat{F}_2(m_2\circ\mathbf{a}). \tag{4.15}$$

In order to take advantage of the simplicity of the matrices in G, in contrast to those in the isomorphic group G_A, we must make the change of coordinates from $\mathbf{a} \in \mathbf{R}^5$ to $\mathbf{x} \in \mathbf{R}^5$ given by

$$\begin{pmatrix}\mathbf{a}\\ 1\end{pmatrix} = A\begin{pmatrix}\mathbf{x}\\ w\end{pmatrix}, \tag{4.16a}$$

that is,

$$\mathbf{a} = \pi(A\mathbf{z}), \tag{4.16b}$$

for $\mathbf{z} = (\mathbf{x},w) \in \mathbf{R}^6$. The sixth coordinate w is therefore not independent, but is given by

$$w = 1 - x - y - z - u - v. \tag{4.16c}$$

Since (4.16a) yields $\mathbf{a} = A_1x$ [cf. (3.4)], we find that it is the function ${}_3E_2$ of the five variables $\mathbf{x} = (x,y,z,u,v)$, defined by (3.3a), that is associated with the new coordinates $\mathbf{x}$. The subscripts 3,2 on this function prove rather unwieldy when further subscripts must be appended. In the sequel, therefore, we shall denote ${}_3E_2(\mathbf{x})$ by $F(\mathbf{x})$.

It is time to take stock of what we have achieved so far. The original three-term relation has been reformulated in new coordinates with the help of a new operation of projection. As we shall see in the next section, this reformulation will enable us to give a complete solution to the problem of finding all the distinct three-term relations for ${}_3F_2$, and in a relatively transparent manner.

New formulation summary: The results given below formulate the original three-term relation in terms of coordinates and functions chosen so that the associated groups G and $H \subset G$ have the simplest possible structures:

$$F(x,y,z,u,v) = {}_3F_2\left(\begin{matrix} x+u+v & y+u+v & z+u+v \\ x+y+z+2u+v & x+y+z+u+2v & \end{matrix}\right) D^{-1},$$

with

$$D = \Gamma(x+y+z+2u+v) \times \Gamma(x+y+z+u+2v)\Gamma(x+y+z). \tag{4.17}$$

The action of the group G on the coordinates $\mathbf{x} \in \mathbf{R}^5$ of $F(\mathbf{x})$ is given by

$$g\circ\mathbf{x} = \pi(g\mathbf{z}), \quad g \in G, \quad \mathbf{x} \in \mathbf{R}^5, \tag{4.18a}$$

where

$$\mathbf{x} = (x,y,z,u,v), \qquad \mathbf{z} = (\mathbf{x}, 1-x-y-z-u-v), \quad \pi\mathbf{z} = \mathbf{x}. \tag{4.18b}$$

The basic relation (4.15) is expressed in terms of the coordinates $\mathbf{x}$ by

$$F(\mathbf{x}) = \alpha(\mathbf{x})F(g_1\circ\mathbf{x}) + \beta(\mathbf{x})F(g_2\circ\mathbf{x}), \tag{4.19}$$

where $g_1, g_2 \in G$ are the 6×6 matrices

$$g_1 = \lambda[6,5,4,3,2,1], \tag{4.20a}$$

$$g_2 = \lambda[4,5,6,1,2,3]. \tag{4.20b}$$

The coefficient functions α and β are defined in terms of the new coordinates by

$$\alpha(\mathbf{x}) = \frac{\pi\Gamma(1-y-u-v)}{\Gamma(y+z+u)\Gamma(y+z+v)\Gamma(z+u+v)\sin\pi(z-x)}, \tag{4.21a}$$

$$\beta(x,y,z,u,v) = \alpha(z,y,x,u,v). \tag{4.21b}$$

We note that the subgroup $H \subset G$ has special significance for (4.19) because it is an invariance group for $F(\mathbf{x})$:

$$F(h\circ\mathbf{x}) = F(\mathbf{x}), \quad h\in H. \tag{4.22}$$

(This is actually Theorem 3.2, adjusted to the present context.) An immediate consequence is the following: let $k\in G$ and let g be an element in the left coset Hk of H in G. Then $g = hk$ for some $h\in H$ and

$$F(g\circ\mathbf{x}) = F(h\circ(k\circ\mathbf{x})) = F(k\circ\mathbf{x}). \tag{4.23}$$

If we define the function F_g by

$$F_g(\mathbf{x}) = F(g\circ\mathbf{x}), \quad g\in G, \tag{4.24}$$

then (4.23) may be written as

$$F_{hk}(\mathbf{x}) = F_k(\mathbf{x}), \quad h\in H. \tag{4.25}$$

Thus, all functions F_g corresponding to the elements of a given left coset of H in G are equal. Since G contains 1440 elements and H contains 120, there are 12 left cosets of H in G. In other words, the action of G on the coordinates $\mathbf{x}$ of F defines exactly 12 new functions of $\mathbf{x}$, one for each coset. It is this fact (as will be explained in greater detail in Sec. V) that accounts for the 12 subtables in Tables 4.2 and 4.3 in Slater[1] (Whipple,[5] Bailey[6]).

The next step is to partition G into its left cosets with respect to H. First we consider the left cosets of $H = P_{5,1}$ in P_6.

Theorem 4.3: The group P_6 contains six left cosets Hk, where the elements $k\in P_6$ may be chosen to be elements of the set

$$K = \{k_r | r = 1,2,...,6\}, \tag{4.26a}$$

where

$$k_1 = I, \quad k_2 = [6,5,4,3,2,1], \quad k_3 = [5,6,4,3,2,1],$$
$$k_4 = [4,5,6,1,2,3], \quad k_5 = [3,2,1,6,5,4], \tag{4.26b}$$
$$k_6 = [3,2,1,5,6,4].$$

Proof: The six sets Hk_r, $r = 1,2,...,6$, are disjoint, since $k'k^{-1}\notin H$ for all pairs $k'\neq k$ and $k',k\in K$. Hence,

$$P_6 = \bigcup_{r=1}^{6} C_r, \quad C_r = Hk_r. \tag{4.27}$$

■

In the next section we shall need the multiplication table for these cosets. This is reproduced as Table I.

The extension of these results on cosets of H in P_6 to cosets of H in G is immediate, and is given by the following.

Theorem 4.4: The left cosets of the subgroup H in G are

$$Hk, \quad k\in\{K,\lambda K\}. \tag{4.28}$$

Denoting the 6×6 array in Table I by C, the multiplication table for the left cosets of H in G is given by the array

TABLE I. Multiplication rules for left cosets of H in G: $C_r = Hk_r$, $r = 1,2,...,6$.

	C_1	C_2	C_3	C_4	C_5	C_6
C_1	C_1	C_2	C_3	C_4	C_5	C_6
C_2	C_2	C_1	C_1	C_5	C_4	C_4
C_3	C_3	C_6	C_6	C_6	C_3	C_3
C_4	C_4	C_5	C_5	C_1	C_2	C_2
C_5	C_5	C_4	C_4	C_2	C_1	C_1
C_6	C_6	C_3	C_2	C_3	C_6	C_5

$$\begin{pmatrix} C & \lambda C \\ \lambda C & C \end{pmatrix}. \tag{4.29}$$

It will be convenient for subsequent applications to denote the 12 matrices in the set $\{K,\lambda K\}$ by

$$k_r, \quad r = 1,2,...,6,1^*,2^*,...,6^*. \tag{4.30a}$$

For $r = 1,2,...,6$, the k_r are the permutation matrices defined by (4.26b), while the corresponding k_{r^*} are just

$$k_{r^*} = \lambda k_r, \quad r = 1,2,...,6. \tag{4.30b}$$

Thus, $k_{1^*} = \lambda$, $k_{2^*} = \lambda k_2 = \lambda\,[6,5,4,3,2,1]$, etc. Since λ is an involution, so that

$$\lambda k_{r^*} = k_r, \tag{4.30c}$$

we must also, for consistency, write $(r^*)^* = r$. We shall refer to r^* as the conjugate of r.

Finally, since only the 12 functions

$$F_{k_r}(x) = F(k_r\circ x), \quad r = 1,2,...,6,1^*,2^*,...,6^*$$

can be generated by the action of G on the coordinates $\mathbf{x}\in\mathbf{R}^5$ [cf. (4.23)–(4.25)], we may simplify the notation still further by writing

$$F_r(x) = F_{k_r}(x), \quad r = 1,2,...,6,1^*,2^*,...,6^*. \tag{4.31}$$

Equation (4.22) exhibits the invariance of the function $F(\mathbf{x})$ under the action of the group H. This property together with the definition (4.31) implies that $F_r(\mathbf{x})$ is invariant under the group $H_r = k_r^{-1}Hk_r$ obtained by the automorphism with $k_r\in K$:

$$F_r(h_r\circ\mathbf{x}) = F_r(\mathbf{x}), \quad h_r\in H_r. \tag{4.32}$$

V. THREE-TERM RELATIONS BETWEEN ${}_3F_2$ SERIES

In the notation defined by (4.30a)–(4.30c) and (4.31), the basic three-term relation (4.19) becomes

$$F_1(\mathbf{x}) = \alpha(\mathbf{x})F_{2^*}(\mathbf{x}) + \beta(\mathbf{x})F_{4^*}(\mathbf{x}), \tag{5.1}$$

since $g_1 = \lambda k_2 = k_{2^*}$ and $g_2 = \lambda k_4 = k_{4^*}$. In this section, we obtain all three-term relations derivable from (5.1) by application of the transformation group G with the action:

$$x\mapsto g\circ x, \quad g\in G, \tag{5.2}$$

and by an elimination procedure to be described below.

The mapping (5.2) can be written in the form

$$\mathbf{x}\mapsto h\circ(k_r\circ\mathbf{x}), \quad h\in H, \quad r = 1,2,...,6,1^*,2^*,...,6^*. \tag{5.3}$$

This suggests carrying out the transformation of relation (5.1) in two steps, by first performing the mapping $\mathbf{x}\mapsto h\circ\mathbf{x}$, $h\in H$, and then the mapping $\mathbf{x}\mapsto k_r\circ\mathbf{x}$.

The first step, applied to (5.1), gives

$$F_1(\mathbf{x}) = \alpha(h\circ\mathbf{x})F_{p^*}(\mathbf{x}) + \beta(h\circ\mathbf{x})F_{q^*}(\mathbf{x}), \tag{5.4a}$$

where the index pair (p^*,q^*) is uniquely determined by finding the left cosets to which $k_{2^*}h$ and $k_{4^*}h$ belong; that is, by solving the inclusion relations

$$k_{2^*}h\in Hk_{p^*}, \quad k_{4^*}h\in Hk_{q^*}. \tag{5.4b}$$

Since the group $H = P_{5,1}$ contains only permutation matrices, it follows that p^* and q^* are both "*-integers" in (5.4a) and (5.4b). Accordingly, the index pair (p,q) may be found by solving

$$k_2 h \in Hk_p, \quad k_4 h \in Hk_q. \tag{5.5}$$

We found the solution of these inclusions for all $h \in H$ by calculating (on MACSYMA) the 12 matrices $k_2 h k_p^{-1}$, $k_4 h k_p^{-1}$, $p = 1,2,...,6$, and identifying the representative pair of matrices in H. The simple solution is as follows: for each pair (p,q) with $p \neq q \in \{1,2,3,4,5\}$ define the subset $H_{p,q}$ of H by

$$H_{p,q} = \{[i_1,i_2,i_3,i_4,i_5,6] \in H \text{ with } i_p = 1,\ i_q = 3\}. \tag{5.6}$$

Each subset $H_{p,q}$ then contains six elements of H. For example, $H_{2,4}$ contains the six permutation matrices $[i_1,1,i_3,3,i_5,6]$ corresponding to the six permutations (i_1,i_3,i_5) of $(2,4,5)$.

Using these notations, the solutions of (5.5) are given in the following.

Theorem 5.1: For each $h \in H_{p,q}$ we have

$$k_2 h \in Hk_{p+1}, \quad k_4 h \in Hk_{q+1} \tag{5.7}$$

for each pair (p,q) with $p \neq q \in \{1,2,3,4,5\}$.

Proof: By MACSYMA, as described above, or by verifying (by hand) that $k_2 h k_{p+1}^{-1} \in H$ and $k_4 h k_{q+1}^{-1} \in H$ for each $h \in H_{p,q}$ (it is only necessary to verify that the sixth component in these products of permutation matrices is 6). ■

Applying Theorem (5.1) to (5.4a) gives

$$F_1(\mathbf{x}) = \alpha(h \circ \mathbf{x}) F_{(p+1)^*}(x) + \beta(h \circ \mathbf{x}) F_{(q+1)^*}(\mathbf{x}), \quad h \in H_{p,q}, \tag{5.8}$$

for each pair (p,q) with $p \neq q \in \{1,2,3,4,5\}$. (We show below that the coefficient functions in this result are invariant for each $h \in H_{p,q}$.)

To obtain the transformation of the basic relation (5.1) by a general element $g \in G$, we must transform (5.8) by $\mathbf{x} \to k_r \circ \mathbf{x}$ [cf.(5.3)]. This results in

$$F_r(\mathbf{x}) = \alpha(g \circ \mathbf{x}) F_s(x) + \beta(g \circ \mathbf{x}) F_t(\mathbf{x}), \tag{5.9a}$$

with

$$g = hk_r, \quad h \in H_{p,q}, \quad r = 1,2,...,6,1^*,2^*,...,6^*. \tag{5.9b}$$

The indices s and t are uniquely determined by the relations

$$k_{(p+1)^*} k_r \in Hk_s, \quad k_{(q+1)^*} k_r \in Hk_t, \tag{5.9c}$$

in which $p \neq q \in \{1,2,3,4,5\}$. [Throughout the remainder of the section, (r,s,t) denote integers with domain $\{1,2,...,6, 1^*,2^*,...,6^*\}$. It will occasionally be convenient to replace the statement

$$r \in \{1,2,...,6,1^*,2^*,...,6^*\}$$

with the phrase "with r in the standard domain."] The triples (r,s,t) of positive integers that can occur in (5.9a) are completely determined from (5.9c) above by the multiplication rules for left cosets of H in G given in Table I. To present the results, it is convenient to represent this table by subscripts alone.

We define the 6×6 array J by

$$J = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 1 & 5 & 4 & 4 \\ 3 & 6 & 6 & 6 & 3 & 3 \\ 4 & 5 & 5 & 1 & 2 & 2 \\ 5 & 4 & 4 & 2 & 1 & 1 \\ 6 & 3 & 2 & 3 & 6 & 5 \end{pmatrix}. \tag{5.10a}$$

The full table for left coset multiplication is represented by the 12×12 index array

$$\mathbf{J} = \begin{Bmatrix} J & J^* \\ J^* & J \end{Bmatrix}, \tag{5.10b}$$

where rows and columns are enumerated by $1,2,...,6,1^*, 2^*,...,6^*$, respectively, which coincide with the entries in column 1 and row 1.

All possible triples (r,s,t) that can occur in (5.9a) are obtained by transcribing the left coset multiplications

$$C_{(p+1)^*} C_r = C_s, \quad C_{(q+1)^*} C_r = C_t \tag{5.11a}$$

into index multiplication (denoted by $\cdot$) in the array $\mathbf{J}$:

$$(p+1)^* \cdot r = s, \quad (q+1)^* \cdot r = t \tag{5.11b}$$

for $p \neq q \in \{1,2,3,4,5\}$. We have tabulated these multiplications in Table II. The values of r in the standard domain are listed in the first row; hence r is a column index. The major row heading is $H_{p,q}$, which includes two rows for each choice of $p < q$. We refer to this pair of rows as the "double-row (p,q)." The pair of entries in column r in the double-row (p,q) is (s,t) with s in the top row. For example, for $h \in H_{3,4}$, the triples (r,s,t) that can occur in (5.9a) are determined by $4^* \cdot r = s$, $5^* \cdot r = t$, and are found from Table II to be

TABLE II. Triples (r,s,t) occurring in three-term relations.

	r	1	2	3	4	5	6	1*	2*	3*	4*	5*	6*
$H_{1,2}$	s	2*	1*	1*	5*	4*	4*	2	1	1	5	4	4
	t	3*	6*	6*	6*	3*	3*	3	6	6	6	3	3
$H_{1,3}$	s	2*	1*	1*	5*	4*	4*	2	1	1	5	4	4
	t	4*	5*	5*	1*	2*	2*	4	5	5	1	2	2
$H_{1,4}$	s	2*	1*	1*	5*	4*	4*	2	1	1	5	4	4
	t	5*	4*	4*	2*	1*	1*	5	4	4	2	1	1
$H_{1,5}$	s	2*	1*	1*	5*	4*	4*	2	1	1	5	4	4
	t	6*	3*	2*	3*	6*	5*	6	3	2	3	6	5
$H_{2,3}$	s	3*	6*	6*	6*	3*	3*	3	6	6	6	3	3
	t	4*	5*	5*	1*	2*	2*	4	5	5	1	2	2
$H_{2,4}$	s	3*	6*	6*	6*	3*	3*	3	6	6	6	3	3
	t	5*	4*	4*	2*	1*	1*	5	4	4	2	1	1
$H_{2,5}$	s	3*	6*	6*	6*	3*	3*	3	6	6	6	3	3
	t	6*	3*	2*	3*	6*	5*	6	3	2	3	6	5
$H_{3,4}$	s	4*	5*	5*	1*	2*	2*	4	5	5	1	2	2
	t	5*	4*	4*	2*	1*	1*	5	4	4	2	1	1
$H_{3,5}$	s	4*	5*	5*	1*	2*	2*	4	5	5	1	2	2
	t	6*	3*	2*	3*	6*	5*	6	3	2	3	6	5
$H_{4,5}$	s	5*	4*	4*	2*	1*	1*	5	4	4	2	1	1
	t	6*	3*	2*	3*	6*	5*	6	3	2	3	6	5

(1,4*,5*), (2,5*,4*), (3,5*,4*),

(4,1*,2*), (5,2*,1*), (6,2*,1*),

(1*,4,5), (2*,5,4), (3*,5,4),

(4*,1,2), (5*,2,1), (6*,2,1).

We have included in Table II only those subsets $H_{p,q} \subset H$ having $p<q$. The table can be extended to include the $H_{p,q}$ having $p>q$ by interchanging the two rows appearing to the right of $H_{q,p}$, *leaving the indices s and t in place.* Since there are six elements of H in each set $H_{p,q}$, where $p \neq q = 1,2,3,4,5$, all 1440 relations (5.9a) are accounted for in the extended table. Some of these relations, however, are identically the same in consequence of properties of the functions $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$. Indeed, we prove in Theorem 5.4 below that the distinct relations (5.9a) are exactly those corresponding to the 120 triples (r,s,t) given in Table II.

We need two theorems on the properties of the coefficient functions $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ that occur in (5.1) and are defined by (4.21a) and (4.21b).

Theorem 5.2: Let $h,h' \in H_{p,q}$ for $p \neq q \in \{1,2,3,4,5\}$. Then

$$\alpha(h \circ \mathbf{x}) = \alpha(h' \circ \mathbf{x}), \quad \beta(h \circ \mathbf{x}) = \beta(h' \circ \mathbf{x}). \tag{5.12}$$

Proof: The functions $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ are invariant under all permutations of (y,u,v), that is, under the group of permutation matrices R defined by

$$R = \{[1,i_2,3,i_4,i_5,6] \mid (i_2,i_4,i_5) \text{ a permutation of } (2,4,5)\}.$$

We can prove the theorem by showing that $h'h^{-1} \in R$ for h', $h \in H_{p,q}$ [see (5.6)]. The product of any two 6×6 permutation matrices, h' and h^{-1}, the first of which has columns $p,q,6$ given by e_1,e_3,e_6, respectively (see the Introduction for the notation), and the second of which has rows $p,q,6$ given by

(100000), (001000), (000001),

always has columns 1,3,6 equal to e_1,e_3,e_6; i.e., $h'h^{-1} \in R$. ■

Theorem 5.3: Let $h \in H_{p,q}$ and $h' \in H_{q,p}$ with $p \neq q \in \{1,2,3,4,5\}$. Then

$$\alpha(h \circ \mathbf{x}) = \beta(h' \circ \mathbf{x}), \quad \beta(h \circ \mathbf{x}) = \alpha(h' \circ \mathbf{x}). \tag{5.13}$$

Proof: Let $h'' = [3,2,1,4,5,6]$. Then, from the definitions (4.21a) and (4.21b) of α and β,

$$\beta(\mathbf{x}) = \alpha(h'' \circ \mathbf{x}), \tag{5.14}$$

and, from the definition (5.6) of $H_{p,q}$,

$$H_{q,p} = h'' H_{p,q}. \tag{5.15}$$

These results together with Theorem 5.2 imply the stated properties of α and β. ■

We can now prove the first of three principal results for three-term relations between ${}_3F_2$ series.

Theorem 5.4: There are 120 distinct relations between the functions $\{F_r \mid r = 1,2,...,6,1^*,2^*,...,6^*\}$ obtainable from the basic relation (5.1) by the group of transformations $\mathbf{x} \mapsto g \circ \mathbf{x}$, $g \in G$. These relations are

$$F_r(\mathbf{x}) = \alpha_{rs^*t^*}(\mathbf{x}) F_{s^*}(\mathbf{x}) + \beta_{rs^*t^*}(\mathbf{x}) F_{t^*}(\mathbf{x}), \tag{5.16a}$$

$$F_{r^*}(\mathbf{x}) = \alpha_{r^*st}(\mathbf{x}) F_s(\mathbf{x}) + \beta_{r^*st}(\mathbf{x}) F_t(\mathbf{x}), \tag{5.16b}$$

where (r,s,t) is any of the 60 triples satisfying

$$r \in \{1,2,...,6\}, \quad s<t \in \{1,2,...,6\} - \{r\}. \tag{5.16c}$$

The coefficient functions in these relations are obtained explicitly in terms of $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ as follows, using Table II. For each triple (r,s,t) determine, in column r, the double-row (p,q) in which s^* and t^* occur (in either order), and let h be any representative element $h \in H_{p,q}$. If s^* occurs in the top row, and t^* in the bottom row, then

$$\alpha_{rs^*t^*}(\mathbf{x}) = \alpha(hk_r \circ \mathbf{x}), \quad \beta_{rs^*t^*}(\mathbf{x}) = \beta(hk_r \circ \mathbf{x}); \tag{5.17a}$$

if s^* occurs in the bottom row and t^* is in the top row, then

$$\alpha_{rs^*t^*}(\mathbf{x}) = \beta(hk_r \circ \mathbf{x}), \quad \beta_{rs^*t^*}(\mathbf{x}) = \alpha(hk_r \circ \mathbf{x}). \tag{5.17b}$$

The coefficients with conjugate indices are given by

$$\alpha_{r^*st}(\mathbf{x}) = \alpha_{rs^*t^*}(\lambda \circ \mathbf{x}), \quad \beta_{r^*st}(\mathbf{x}) = \beta_{rs^*t^*}(\lambda \circ \mathbf{x}). \tag{5.17c}$$

Proof: By Table II, its extension to $p>q$, and Theorems 5.2 and 5.3. Observe in particular that the 120 triples in the set $\{(r,s^*,t^*), (r,t^*,s^*)\}$ obtained by the conditions (5.16c) enumerate exactly the triples in the left half of the extended Table II, with half in Table II. ■

Properties of the coefficient functions $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ are very important for deriving distinct relations between the 12 functions F_r, as shown by the application of Theorems 5.2 and 5.3. These properties can be quite tedious to verify. In what follows we prove an important result that implies the existence of various properties of these coefficients which allows us to ignore, for some purposes, the details. Let us define a relation between the 12 functions F_r to be "F linear" if the relation is invariant under the substitutions $F_r \mapsto \mu F_r$ for $r = 1,2,...,6,1^*,2^*,...,6^*$, $\mu \in \mathbf{R}$. The following theorem then greatly simplifies the proofs of the results obtained in the sequel.

Theorem 5.5: Let (r,s,t) be distinct integers in the standard domain. Then one can derive from the basic relation (5.1) at most one F-linear three-term relation

$$F_r(\mathbf{x}) = \alpha_{rst}(\mathbf{x}) F_s(\mathbf{x}) + \beta_{rst}(\mathbf{x}) F_t(\mathbf{x}) \tag{5.18}$$

between the functions $F_r(\mathbf{x})$, $F_s(\mathbf{x})$, and $F_t(\mathbf{x})$.

Proof: If there are two relations of the form (5.18), then $F_r(\mathbf{x})$, $F_s(\mathbf{x})$, and $F_t(\mathbf{x})$ are pairwise related, i.e., two-term relations exist between these functions. Applying the transformation $\mathbf{x} \mapsto k_p \circ \mathbf{x}$ to these three (hypothetical) two-term relations, we find two-term relations between all functions with index pairs given by $(r \cdot p, s \cdot p)$, $(r \cdot p, t \cdot p)$, and $(s \cdot p, t \cdot p)$, with $p \in \{1,2,...,6,1^*,2^*,...,6^*\}$.

Let us write $r \sim s$ if $F_r(\mathbf{x}) = \gamma_{rs}(\mathbf{x}) F_s(\mathbf{x})$, where γ_{rs} is a quotient of products of gamma functions. The relation $\sim$ is then an equivalence relation between the integers r and s. The existence of two relations of the form (5.18) then implies that

$$r \cdot p \sim s \cdot p \sim t \cdot p, \quad p \in \{1,2,...,6,1^*,2^*,...,6^*\}. \tag{5.19a}$$

Our strategy is to use this last result, which is implied by the existence of two relations of the form (5.18), to show that $2^* \sim 4^*$, i.e., that $F_{2^*}(\mathbf{x})$ and $F_{4^*}(\mathbf{x})$ are related by a two-term relation. But this is false, since, by assumption, the basic relation (5.1) is a three-term relation. This contradiction will then prove the theorem.

We need to show that the set of equivalence relations (5.19a) implies $2^* \sim 4^*$ for every possible choice of (r,s,t)

with $r<s<t$. In many instances, it is possible to show that

$$r\cdot p\sim s\cdot p,\quad p\in\{1,2,...,6,1^*,2^*,...,6^*\} \tag{5.19b}$$

implies $2^*\sim 4^*$. We consider these cases first.

Recall that $r\sim s$ implies all the equivalences (5.19b), i.e., each pair of integers occurring in the same column and row r and row s of the index array **J** is equivalent. For example, for $(r,s)=(1,3)$ we find from columns 2^* and 4^* that $1\sim 3$ implies $2^*\sim 6^*$ and $6^*\sim 4^*$, i.e., $1\sim 3$ implies $2^*\sim 4^*$. The conjugate result, $1^*\sim 3^*$ implies $2\sim 4$ is obtained similarly. In this way, we verify that

$$r\sim s \text{ implies } 2^*\sim 4^*,\quad r^*\sim s^* \text{ implies } 2\sim 4 \tag{5.20a}$$

for all pairs

$$(r,s)=(1,3),(1,5),(1,6),(2,3)(2,4),$$
$$(2,6),(3,4),(3,5),(4,6),(5,6). \tag{5.20b}$$

For the "missing" pairs $(r,s)=(1,2),(1,4),(2,5),(3,6),(4,5)$, we find $1\sim 2$ implies $1\sim 3$, $1\sim 4$ implies $2\sim 6$, $2\sim 5$ implies $1\sim 4$, $3\sim 6$ implies $2\sim 6$, and $4\sim 5$ implies $1\sim 2$. Combining these equations first among themselves, as necessary, and then with (5.20a) and (5.20b), we find that relation (5.20a) is also valid for

$$(r,s)=(1,2),(1,4),(2,5),(3,6),(4,5). \tag{5.20c}$$

Next, we use the fact that the existence of two three-term relations between the same three functions implies the equivalence, not of pairs of indices, but of triples, as given by (5.19a). Since each triple always includes a pair of indices in the sets (5.20b) and (5.20c) or the conjugates of these pairs, we conclude that $r\sim s\sim t$ implies $2^*\sim 4^*$ for all possible triples. ■

Remark: Theorem 5.5 implies that the relations between the α and β functions given in (5.17a) and (5.17b) must be true, for otherwise the theorem would be false (by contradiction). Theorem 5.5 simplifies considerably the task of enumerating all additional F-linear three-term relations that can be derived from the results given in Table II.

Theorem 5.4 gives 120 relations derivable from the basic relation (5.1) by direct application of the group of transformations $\mathbf{x}\mapsto g\circ\mathbf{x}$, $g\in G$. Further F-linear three-term relations can be found from these 120 by the process of *elimination*. We show next how this is done. There are two kinds of relations (given in Theorems 5.6 and 5.7 below).

Consider three relations of the form given by (5.16a) in Theorem 5.4:

$$F_r(\mathbf{x})=\alpha_{rp^*q^*}(\mathbf{x})F_{p^*}(\mathbf{x})+\beta_{rp^*q^*}(\mathbf{x})F_{q^*}(\mathbf{x}), \tag{5.21a}$$
$$F_s(\mathbf{x})=\alpha_{sp^*q^*}(\mathbf{x})F_{p^*}(\mathbf{x})+\beta_{sp^*q^*}(\mathbf{x})F_{q^*}(\mathbf{x}), \tag{5.21b}$$
$$F_t(\mathbf{x})=\alpha_{tp^*q^*}(\mathbf{x})F_{p^*}(\mathbf{x})+\beta_{tp^*q^*}(\mathbf{x})F_{q^*}(\mathbf{x}), \tag{5.21c}$$

where each triple (r,p,q), (s,p,q), (t,p,q) is chosen by the rule (5.16c) and such that $r<s<t$. We can now prove the second principal result for three-term relations.

Theorem 5.6: There are 40 distinct F-linear relations between the functions F_r, with r in the standard domain, obtainable by elimination between triples of relations of type (5.21a)–(5.21c). Twenty are given by

$$\gamma_{rst}(\mathbf{x})F_r(\mathbf{x})+\delta_{rst}(\mathbf{x})F_s(\mathbf{x})+\epsilon_{rst}(\mathbf{x})F_t(\mathbf{x})=0, \tag{5.22a}$$

and 20 by the conjugates to these:

$$\gamma_{r^*s^*t^*}(\mathbf{x})F_{r^*}(\mathbf{x})+\delta_{r^*s^*t^*}(\mathbf{x})F_{s^*}(\mathbf{x})$$
$$+\epsilon_{r^*s^*t^*}(\mathbf{x})F_{t^*}(\mathbf{x})=0, \tag{5.22b}$$

there being a relation of each type for each triple (r,s,t) such that

$$r<s<t\in\{1,2,...,6\}. \tag{5.22c}$$

The coefficient functions are given explicitly in terms of $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ by Eqs. (5.17a)–(5.17c) and the following relations and their conjugates:

$$\gamma_{rst}(\mathbf{x})=\alpha_{sp^*q^*}(\mathbf{x})\beta_{tp^*q^*}(\mathbf{x})-\alpha_{tp^*q^*}(\mathbf{x})\beta_{sp^*q^*}(\mathbf{x}), \tag{5.23a}$$
$$\delta_{rst}(\mathbf{x})=\alpha_{tp^*q^*}(\mathbf{x})\beta_{rp^*q^*}(\mathbf{x})-\alpha_{rp^*q^*}(\mathbf{x})\beta_{tp^*q^*}(\mathbf{x}), \tag{5.23b}$$
$$\epsilon_{rst}(\mathbf{x})=\alpha_{rp^*q^*}(\mathbf{x})\beta_{sp^*q^*}(\mathbf{x})-\alpha_{sp^*q^*}(\mathbf{x})\beta_{rp^*q^*}(\mathbf{x}). \tag{5.23c}$$

Proof: Equation (5.22a), with coefficient functions given by (5.23a)–(5.23c) follows directly from the three relations (5.21a)–(5.21c). The conjugate relation (5.22b) is obtained similarly. Thus, to prove the theorem, we must show that there are 20 distinct sets of three relations of the form (5.21a)–(5.21c); i.e., a set for each triple (r,s,t) satisfying conditions (5.22c).

Consider the left half of Table II (columns 1 through 6). The set of 60 triples $\{(r,p^*,q^*)\}$ associated with this half of the table can be described as follows. Select any pair $p\neq q$ from the set $\{1,2,...,6\}$, and then r from the set $\{1,2,...,6\}-\{p,q\}=\{r_1,r_2,r_3,r_4\}$. Then one of the two triples (r_i,p^*,q^*) or (r_i,q^*,p^*), but not both, occurs in the table for each $i=1,2,3,4$. This is true for all 30 choices of the pair (p,q), thus giving all $(30/2)\times 4=60$ triples. For each pair (p,q), then, there are four sets of three relations of the form (5.21a)–(5.21c), a set for each triple (r,s,t) with $r<s<t$ selected from $\{r_1,r_2,r_3,r_4\}$.

We illustrate the preceding results with some examples. For $p=1$, $q=5$, we find $(2,1^*,5^*)$, $(3,1^*,5^*)$, $(4,5^*,1^*)$, and $(6,1^*,5^*)$ from the table, so that $\{r_1,r_2,r_3,r_4\}=\{2,3,4,6\}$; the four sets of three relations (5.21a)–(5.21c) all have $(p,q)=(1,5)$ and correspond to $(r,s,t)=(2,3,4)$, $(2,3,6)$, $(2,4,6)$, or $(3,4,6)$. A second example $p=1$, $q=2$ leads to $(r,s,t)=(3,4,5)$, $(3,4,6)$, or $(4,5,6)$, which illustrates that a given triple—in this case $(r,s,t)=(3,4,6)$—may be repeated, even though the index pairs $(p,q)=(1,5)$ and $(1,2)$ are not equal. For this reason, the same three functions F_3, F_4, and F_6 can be expressed in terms of F_{1^*} and F_{5^*}, or in terms of F_{1^*} and F_{2^*}. Either of these sets of three relations must lead, however, by Theorem 5.5, to the same relation (5.22a), up to a common multiple of the coefficient functions.

Considering all 15 sets of four triples that can be obtained from the left half of Table II, we find all the possible $\binom{6}{3}=20$ triples (r,s,t) with $r<s<t\in\{1,2,...,6\}$. Some triples are repeated, as noted above, but by Theorem 5.5, only one new relation per triple can be produced. The 20 relations corresponding to these index choices are relations between the F_r with $r=1,2,...,6$; hence, all are new (i.e., not obtainable directly from Table II). Since the right half of the table is gotten by applying the *-operation to the left half, we also obtain relations (5.22b) for all r,s,t satisfying (5.22c). ■

There is a second process of elimination that leads to three-term relations not obtainable from Theorems 5.4 and 5.6. We select any triple (r,s^*,t^*) with $r\neq s\neq t\in\{1,2,...,6\}$ from the left half of the table, where we always find either (s^*,r,t) or (s^*,t,r). The corresponding three-term identities are

$$F_r(\mathbf{x}) = \alpha_{rs^*t^*}(\mathbf{x})F_{s^*}(\mathbf{x}) + \beta_{rs^*t^*}(\mathbf{x})F_{t^*}(\mathbf{x}), \quad (5.24a)$$

$$F_{s^*}(\mathbf{x}) = \alpha_{s^*rt}(\mathbf{x})F_r(\mathbf{x}) + \beta_{s^*rt}(\mathbf{x})F_t(\mathbf{x}), \quad (5.24b)$$

where the coefficient functions are given explicitly by (5.17a)–(5.17c). We now eliminate $F_{s^*}(x)$ between (5.24a) and (5.24b), which leads to the third principal result for three-term relations.

Theorem 5.7: There are 60 distinct F-linear relations between the F_r, with r in the standard domain, obtainable by elimination between pairs of relations of type (5.24a) and (5.24b). Thirty are given by

$$\xi_{rtt^*}(\mathbf{x})F_r(\mathbf{x}) + \eta_{rtt^*}(\mathbf{x})F_t(\mathbf{x}) + \zeta_{rtt^*}(\mathbf{x})F_{t^*}(\mathbf{x}) = 0, \quad (5.25a)$$

and 30 by the conjugates to these:

$$\xi_{r^*t^*t}(\mathbf{x})F_{r^*}(\mathbf{x}) + \eta_{r^*t^*t}(\mathbf{x})F_{t^*}(\mathbf{x}) + \zeta_{r^*t^*t}(\mathbf{x})F_t(\mathbf{x}) = 0, \quad (5.25b)$$

there being a relation of each type for each

$$r\neq t\in\{1,2,...,6\}. \quad (5.25c)$$

The coefficient functions are given explicitly in terms of $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ by Eqs. (5.17a)–(5.17c) and the following relations and their conjugates:

$$\xi_{rtt^*}(\mathbf{x}) = \alpha_{rs^*t^*}(\mathbf{x})\alpha_{s^*rt}(\mathbf{x}) - 1, \quad (5.26a)$$

$$\eta_{rtt^*}(\mathbf{x}) = \alpha_{rs^*t^*}(\mathbf{x})\beta_{s^*rt}(\mathbf{x}), \quad (5.26b)$$

$$\zeta_{rtt^*}(\mathbf{x}) = \beta_{rs^*t^*}(\mathbf{x}). \quad (5.26c)$$

All indices in these relations should be conjugated in obtaining the coefficient functions in Eqs. (5.25b).

Proof: Elimination of $F_{s^*}(\mathbf{x})$ between the expressions (5.24a) and (5.24b) gives (5.25a) with the coefficients defined by (5.26a)–(5.26c). We find from Table II that all choices of indices $r\neq t\in\{1,2,...,6\}$ occur, some pairs (r,t) more than once for the same s^*. Whenever (5.24) occur, so do their conjugates. By Theorem 5.5, we can obtain one relation, up to a common factor between the coefficients, for each triple (r,t,t*). ■

Our final theorem is the following.

Theorem 5.8: All F-linear three-term relations between the 12 functions F_r, with r in the standard range, which are obtainable from (5.1) by transformations from the group G and by elimination, are given in Theorems 5.4, 5.6, and 5.7.

Proof: There are $\binom{12}{3} = 220$ distinct choices of the indices $r<s<t$, with the indices in the standard range, and we have given exactly this number of three-relations in Theorems 5.4, 5.6, and 5.7. Moreover, none of these relations can degenerate to a two-term relation (for general values of $\mathbf{x}$), since Theorem 5.5 would then be violated. ■

We note that each of the possible 220 choices of the triple (r,s,t) corresponds to an F-linear three-term relation. If there were a direct way of seeing this, our paper could be considerably shortened.

Remark: As noted earlier, the coefficient functions $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ must possess a number of properties beyond those already given in Theorems 5.2 and 5.3 in order that Theorem 5.5 be valid. Let us give an illustrative example. Consider the transformation of the basic relation (5.1) corresponding to the elements of the Abelian group of involutions $\{I,k_{2^*},k_{4^*},k_5\}$, which is the transformation $A^{-1}MA$ of the group M defined by (4.5). This leads to three additional relations:

$$F_{2^*}(\mathbf{x}) = \alpha(k_{2^*}\circ\mathbf{x})F_1(\mathbf{x}) + \beta(k_{2^*}\circ\mathbf{x})F_5(\mathbf{x}), \quad (5.27a)$$

$$F_{4^*}(\mathbf{x}) = \alpha(k_{4^*}\circ\mathbf{x})F_5(\mathbf{x}) + \beta(k_{4^*}\circ\mathbf{x})F_1(\mathbf{x}), \quad (5.27b)$$

$$F_5(\mathbf{x}) = \alpha(k_5\circ\mathbf{x})F_{4^*}(\mathbf{x}) + \beta(k_5\circ\mathbf{x})F_{2^*}(\mathbf{x}). \quad (5.27c)$$

Eliminating $F_{2^*}(\mathbf{x})$ between (5.1) and (5.27a) leads to a relation between $F_1(\mathbf{x})$, $F_{4^*}(\mathbf{x})$, and $F_5(\mathbf{x})$, which is exactly the relation (5.27b) because of the identities,

$$\begin{aligned}\alpha(\mathbf{x})\alpha(k_{2^*}\circ\mathbf{x}) + \beta(\mathbf{x})\beta(k_{4^*}\circ\mathbf{x}) &= 1,\\ \alpha(\mathbf{x})\beta(k_{2^*}\circ\mathbf{x}) + \beta(\mathbf{x})\alpha(k_{4^*}\circ\mathbf{x}) &= 0.\end{aligned} \quad (5.28)$$

The correctness of these relations may be verified directly from the definitions (4.21a) and (4.21b). The compatibility of the set of four relations (5.1) and (5.27a)–(5.27c) is implied by (5.28).

We conclude this section by noting the relationship between the results obtained here and those of Whipple[5] (see also Bailey[6] and Slater[1]). We refer to Slater's tabulation of these results, and use the notation introduced there. The set $F_p(\rho)$ [resp. $F_n(\rho)$] for each $\rho\in\{0,1,...,5\}$ is defined by

$$F_p(\rho) = \{F_p(\rho;\sigma,\tau)\,|\,\sigma<\tau\in\{0,1,...,5\}-\{\rho\}\}, \quad (5.29a)$$

[resp.

$$F_n(\rho) = \{F_n(\rho;\sigma,\tau)\,|\,\sigma<\tau\in\{0,1,...,5\}-\{\rho\}\}.] \quad (5.29b)$$

Each of the ten symbols in a given one of these sets denotes a ${}_3F_2$ series having distinct sets of numerator and denominator parameters. Thus, we have defined in Eqs. (5.29) 12 sets, each containing ten ${}_3F_2$ series with distinct parameter sets. (The letters p and n serve to denote two types of sets.) In all, we have 120 ${}_3F_2$ series, each with its distinct parameter set, distributed into 12 sets of ten each. The 12 sets are given in Tables 4.2 and 4.3 of Slater, where typical parameters of the ten functions in each set are listed.

It is convenient to extend the sets $F_p(\rho)$ [resp. $F_n(\rho)$] to 120 functions, obtained by including all 12 "place" permutations of the numerator and denominator parameters for each of the ten ${}_3F_2$ series in the set. We denote these extended sets by the notation $\mathcal{F}_p(\rho)$ [resp. $\mathcal{F}_n(\rho)$]. We now have 1440 distinct ${}_3F_2$ series (counting the place permutations of numerator and denominator parameters as distinct) distributed into 12 sets, each containing 120 functions. Finally, we also introduce the 12 sets of functions $\mathcal{F}_r$ defined by

$$\mathcal{F}_r = \{F_r(h\circ\mathbf{x})\,|\,h\in H;\ \mathbf{x}=A^{-1}\circ\mathbf{a}\}, \quad (5.30a)$$

each r in the standard domain. The functions in this set may also be written [cf. (2.6a), (3.4), (4.17), (4.18a)] as

$$F_r(h\circ\mathbf{x})|_{\mathbf{x}=A^{-1}\circ\mathbf{a}} = {}_3\hat{F}_2(k_r';h'\circ\mathbf{a}), \quad (5.30b)$$

where

$$k'_r = Ak_rA^{-1}, \quad h' = AhA^{-1}. \tag{5.30c}$$

With these preliminaries, the results obtained by Whipple (Tables 4.2 and 4.3 in Slater) have the following comprehensive explanation in terms of the group G and the subgroup H.

(i) The sets of functions $\mathcal{F}_p(\rho)$, $\mathcal{F}_n(\rho)$, and $\mathcal{F}_r$ are related by

$$\mathcal{F}_p(\rho) = \mathcal{F}_{\rho+1}, \quad \mathcal{F}_n(\rho) = \mathcal{F}_{(\rho+1)^*}, \quad \rho = 0,1,...,5 \tag{5.31}$$

[see (4.26a) and (4.26b), (4.30b) and (4.30c), and (4.31)]. Thus, the 12 sets of functions given in Slater's two tables are one-to-one with the left cosets of H in G (equivalently, with the left cosets of H_A in G_A). The functions in a given set, e.g., $F_p(\rho)$ [resp. $F_n(\rho)$], which is now trivially extended to $\mathcal{F}_p(\rho)$ [resp. $\mathcal{F}_n(\rho)$], and contains 120 functions as described above, are those with parameters

$$\mathbf{a}' = k'_{\rho+1}h'\circ\mathbf{a} \quad (\text{resp. } \mathbf{a}' = k'_{(\rho+1)^*}h'\circ\mathbf{a}), \quad h' \in H_A.$$

The functions within a given set are then all equal because the subgroup H is an invariance group of the function $F(\mathbf{x})$, or, equivalently, because $H_r = k_r^{-1}Hk_r$ is an invariance group of the function $F_r \in \mathcal{F}_r$ [see (4.32)]. It is important to observe that this implies there are no new results for two-term relations beyond $F(h\circ\mathbf{x}) = F(\mathbf{x})$, $h \in H$, which are contained in Slater's tables: the invariance of the functions in the set $\mathcal{F}_r$ under the group H_r is equivalent to the invariance of the functions in the set $\mathcal{F}_1 = \mathcal{F}_p(0)$ under the group H.

(ii) Theorems 5.4, 5.6, and 5.7 together give 220 relations between the 12 functions $F_r(\mathbf{x})$, with r in the standard range. As noted, these relations split into a set of 110 relations and a set of 110 conjugate relations, where the relation conjugate to a given one is obtained by the transformation $\mathbf{x} \mapsto \lambda\circ\mathbf{x}$. [This transformation interchanges the letters n and p in (5.29) and (5.31).] The results given in Slater's Sec. 4.3.2 (or Sec. 3.7 in Bailey) are related to those given in Theorems 5.4, 5.6, and 5.7 as follows. Slater's (4.3.2.1) and its conjugate (4.3.2.2) are two of the 120 relations given in Theorem 5.4; (4.3.2.3) and its conjugate (4.3.2.4) are two of the 40 relations given in Theorem 5.6; and (4.3.2.5) and its conjugate (4.3.2.6) are two of the 60 relations given in Theorem 5.7.

[1] L. J. Slater, *Generalized Hypergeometric Functions* (Cambridge U. P., Cambridge, 1966).

[2] L. C. Biedenharn and J. D. Louck, *The Racah–Wigner Algebra in Quantum Theory, Encyclopedia of Mathematics and its Applications*, Vol. 9 (Addison–Wesley, Reading, MA, 1981).

[3] J. Thomae, "Ueber die functionen welche durch Reihen...," J. Reine Angew. Math. **87**, 26 (1879).

[4] G. H. Hardy, *Ramanujan* (Chelsea, New York, 1978), 3rd ed.

[5] F. J. Whipple, "A group of generalized hypergeometric series: Relations between 120 allied series," Proc. London Math. Soc. **23**, 104 (1925).

[6] W. N. Bailey, *Generalized Hypergeometric Series* (Cambridge U. P., Cambridge, 1935).

[7] L. C. Biedenharn and M. Lohe (private communication).

[8] J. Wilson, "Some hypergeometric orthogonal polynomials," SIAM J. Math. Anal. **11**, 670 (1980).

[9] L. C. Biedenharn, "Generalization of an identity of Bailey," invited talk presented at N.S.F.–C.B.M.S. regional conference on special functions, physics and computer algebra, Arizona State University, Tempe, Arizona 20–24 May 1985.

[10] G. James and A. Kerber, *The Representation Theory of the Symmetric Group, Encyclopedia of Mathematics and its Applications*, Vol. 16 (Addison–Wesley, Reading, MA, 1981).

Continuous Hahn polynomials and the Heisenberg algebra

Carl M. Bender and Lawrence R. Mead[a)]
Department of Physics, Washington University, St. Louis, Missouri 63130

Stephen S. Pinsky
Department of Physics, Ohio State University, Columbus, Ohio 43210

(Received 2 September 1986; accepted for publication 5 November 1986)

Continuous Hahn polynomials have surfaced in a number of somewhat obscure physical applications. For example, they have emerged in the description of two-photon processes in hydrogen, hard-hexagon statistical mechanical models, and Clebsch–Gordan expansions for unitary representations of the Lorentz group SO(3,1). In this paper it is shown that there is a simple and elegant way to construct these polynomials using the Heisenberg algebra.

I. INTRODUCTION

Most of the familiar orthogonal polynomials of mathematical physics satisfy second-order linear eigenvalue differential equations. The Hahn polynomials are unusual in that they satisfy a second-order difference rather than differential equation.

Originally, the Hahn polynomials[1] were constructed as totally discrete analogs of the more conventional polynomials of mathematical physics. These polynomials were originally defined by two three-term recursion relations, one in the index and one in the argument, and were shown to satisfy a discrete orthogonality relation. It was only very recently in 1985, that Atakishiyev and Suslov[2] and Askey[3] generalized these discrete polynomials to continuous polynomials in the following sense: (1) the argument can be extended from a discrete variable to a continuous variable; (2) the orthogonality relation can be written as an integral rather than as a sum over a weight function; and (3) the index can be analytically continued to complex numbers. This continuous generalization establishes a complete analog between the Hahn polynomials[4] and the more conventional polynomials of mathematical physics except for the fact that the continuous Hahn polynomials satisfy functional difference rather than differential equations.

In this paper we consider a special class of continuous Hahn polynomials which we designate $S_n(x)$. We have organized our presentation as follows. Section II discusses the elementary properties of $S_n(x)$, some of which are new results. In Sec. III we give the main result of this paper; namely, the connection between the Heisenberg algebra and $S_n(x)$. More detailed mathematical properties of $S_n(x)$, such as its asymptotic behavior, its zeros, its representation as a generalized hypergeometric function, and the expansion of functions as a series of $S_n(x)$, are discussed in Sec. IV.

II. ELEMENTARY PROPERTIES OF $S_n(x)$

The first few polynomials $S_n(x)$ are

$$\begin{aligned}
S_0(x) &= (1),\\
S_1(x) &= x,\\
S_2(x) &= \tfrac{1}{2}(x^2-1),\\
S_3(x) &= \tfrac{1}{6}(x^3-5x),\\
S_4(x) &= \tfrac{1}{24}(x^4-14x^2+9),\\
S_5(x) &= \tfrac{1}{120}(x^5-30x^3+89x),\\
S_6(x) &= \tfrac{1}{720}(x^6-55x^4+439x^2-225),\\
S_7(x) &= (1/7!)(x^7-91x^5+1519x^3-3429x),\\
S_8(x) &= (1/8!)(x^8-140x^6+4214x^4\\
&\quad -24\,940x^2+11\,025),\\
S_9(x) &= (1/9!)(x^9-204x^7+10\,038x^5\\
&\quad -122\,156x^3+230\,481x),\\
S_{10}(x) &= (1/10!)(x^{10}-285x^8+21\,378x^6\\
&\quad -463\,490x^4+2250\,621x^2-893\,025).
\end{aligned} \tag{2.1}$$

To compute these polynomials we can use the two-term recursion relation

$$nS_n(x) = xS_{n-1}(x) - (n-1)S_{n-2}(x). \tag{2.2}$$

If we define the generating function $G(t)$ by

$$G(t) = \sum_{n=0}^{\infty} t^n S_n(x), \tag{2.3}$$

then the recursion relation (2.2) gives a simple differential equation satisfied by $G(t)$:

$$(1+t^2)G'(t) = (x-t)G(t). \tag{2.4}$$

The solution to (2.4) satisfying the normalization condition $G(0)=1$ is

$$G(t) = e^{x\arctan t}/(1+t^2)^{1/2}. \tag{2.5}$$

Applying the Cauchy integral formula to (2.5) gives a simple integral representation for $S_n(x)$:

$$S_n(x) = \frac{1}{2\pi i}\oint_C \frac{dz}{z^{n+1}}\,\frac{e^{x\arctan z}}{(1+z^2)^{1/2}}, \tag{2.6}$$

where C is a contour encircling the origin in the z plane.

The functional difference equation that the polynomials $S_n(x)$ satisfy is

$$\begin{aligned}
&(1-ix)S_n(x+2i) + (1+ix)S_n(x-2i)\\
&\quad = (4n+2)S_n(x).
\end{aligned} \tag{2.7}$$

It is not easy to find equations that relate $S_n(x)$ and its derivatives. The simplest such relation we have found is

$$S_n'(x) = \sum_{j=0}^{[(n-1)/2]} \frac{S_{n-1-2j}(x)}{2j+1}(-1)^j. \tag{2.8}$$

[a)] Permanent address: Physics Department, University of Southern Mississippi, Hattiesburg, Mississippi 39401.

 0022-2488/87/030509-05$02.50

It is easy to discover the weight function $W(x)$ for the orthogonality relation satisfied by $S_n(x)$ using experimental methods. If we require that $S_n(x)$ satisfy an orthonormality relation of the general form

$$\int_{-a}^{a} dx\, W(x) S_n(x) S_m(x) = \delta_{mn}, \tag{2.9}$$

we can assume that $W(x)$ is an even function because the polynomials exhibit parity. Then we can compute the first few even moments μ_{2n} of $W(x)$:

$$\mu_{2n} = \int_{-a}^{a} dx\, W(x) x^{2n}. \tag{2.10}$$

The results are $\mu_0 = 1$, $\mu_2 = 1$, $\mu_4 = 5$, $\mu_6 = 61$, $\mu_8 = 1385$,..., which we recognize are just the absolute values of the Euler numbers

$$\mu_{2n} = |E_{2n}|. \tag{2.11}$$

From the integral formula[5]

$$\frac{1}{2}\int_{-\infty}^{\infty} \frac{dx\, x^{2n}}{\cosh(\pi x/2)} = |E_{2n}|, \tag{2.12}$$

we immediately identify $a = \infty$ in (2.10) and the weight function

$$W(x) \equiv 1/2\cosh(\pi x/2). \tag{2.13}$$

It is easy to verify the orthogonality condition

$$\int_{-\infty}^{\infty} \frac{dx\, S_n(x) S_m(x)}{2\cosh(\pi x/2)} = \delta_{mn}. \tag{2.14}$$

We insert the complex integral representation for $S_n(x)$ in (2.6) and interchange orders of integration. Then we use the integral identity[6]

$$\frac{1}{2}\int_{-\infty}^{\infty} dx \frac{\cosh(xz)}{\cosh(\pi x/2)} = \frac{1}{\cos z} \quad (|z| < \pi/2), \tag{2.15}$$

and the trigonometric identity

$$(z^2+1)^{1/2}(z'^2+1)^{1/2}\cos(\arctan z + \arctan z') = 1 - zz'. \tag{2.16}$$

The result is

$$\int_{-\infty}^{\infty} \frac{dx\, S_n(x) S_m(x)}{2\cosh(\pi x/2)} = \frac{1}{(2\pi i)^2}\oint_C \oint_C \frac{dz\, dz'}{z^{m+1} z'^{n+1}(1-zz')} = \delta_{mn}.$$

Further mathematical properties of the polynomials $S_n(x)$ are discussed in Sec. IV.

III. CONNECTION WITH QUANTUM MECHANICS

We now proceed with a discussion of the connection between $S_n(x)$ and the Heisenberg algebra.

The Heisenberg algebra consists of two Hermitian operators p and q which satisfy the commutation relation

$$[q,p] = i. \tag{3.1}$$

In terms of q and p we construct a set of homogeneous polynomial operators $T_{m,n}$. Here $T_{m,n}$ is defined as the sum of all possible terms containing m factors of p and n factors of q and is thus a totally symmetric Hermitian object containing $(m+n)!/(m!n!)$ individual terms. For example,

$$T_{0,1} = q,$$
$$T_{1,1} = pq + qp,$$
$$T_{2,1} = p^2 q + pqp + qp^2,$$
$$T_{2,2} = qpqp + pqpq + pq^2 p + pq^2 q + p^2 q^2 + q^2 p^2.$$

The operators $T_{m,n}$ exhibit some very elementary properties:

$$[q, T_{m,n}] = i(m+n) T_{m-1,n}, \tag{3.2}$$

$$[T_{m,n}, p] = i(m+n) T_{m,n-1}, \tag{3.3}$$

$$T_{m,m+k} = \frac{(2m+k)!m!}{(2m)!(m+k)!}\frac{1}{2}\{T_{m,m}, q^k\}_+, \tag{3.4}$$

$$T_{m+k,m} = \frac{(2m+k)!m!}{(2m)!(m+k)!}\frac{1}{2}\{T_{m,m}, p^k\}_+. \tag{3.5}$$

The connection between the operators $T_{m,n}$ and the polynomials $S_n(x)$ is extremely simple. First consider $T_{n,n}$. Using the commutation relation (3.1) one can always restructure $T_{n,n}$ as a polynomial in $T_{1,1}$; this polynomial is proportional to $S_n(T_{1,1})$ (see Ref. 7):

$$T_{n,n} = [1/(2n-1)!!] S_n(T_{1,1}). \tag{3.6}$$

Using (3.4) we can generalize (3.6) slightly to read

$$T_{m,m+k} = \frac{(2m+k)!}{(m+k)!2^{m+1}}\{q^k, S_m(T_{1,1})\}_+. \tag{3.7}$$

In fact, we could regard (3.6) as the defining equation for $S_n(x)$. The formula in (3.4) would then be in exact analogy with the defining equation for Chebyshev polynomial; the fact that $\cos(n\theta)$ is a polynomial in $\cos\theta$ allows one to define the nth Chebyshev polynomial $T_n(x)$ by

$$\cos(n\theta) \equiv T_n(\cos\theta).$$

We conclude this section with a heuristic discussion of the connection between the algebra of the polynomials $S_n(x)$ and quantum mechanics. We argue that the polynomials $S_n(x)$ are, in fact, the discrete analogs of the Hermite polynomials $\mathrm{He}_n(x)$. The first few Hermite polynomials are

$$\begin{aligned} \mathrm{He}_0(x) &= 1 \\ \mathrm{He}_1(x) &= x, \\ \mathrm{He}_2(x) &= x^2 - 1, \\ \mathrm{He}_3(x) &= x^3 - 3x, \\ \mathrm{He}_4(x) &= x^4 - 6x^2 + 3, \\ \mathrm{He}_5(x) &= x^5 - 10x^3 + 15x. \end{aligned} \tag{3.8}$$

The Hermite polynomials satisfy an eigenvalue differential equation[8]

$$\mathrm{He}_n''(x) - x\,\mathrm{He}_n'(x) + n\,\mathrm{He}_n(x) = 0. \tag{3.9}$$

The eigenvalue difference equation in (2.7) may be recast in a form in which the discrete differences are explicit:

$$\frac{S_n(x+2i) - 2S_n(x) + S_n(x-2i)}{(2i)^2} - x\frac{S_n(x+2i) - S_n(x-2i)}{4i} + nS_n(x) = 0. \tag{3.10}$$

The first term in (3.10) is the central second difference,

$$D^2 S_n^{(x)} = [S_n(x+h) - 2S_n + S_n(x-h)]/h^2,$$

and the second term is the central first difference,

$$DS_n(x) = [S_n(x+h) - S_n(x-h)]/(2h),$$

where $h = 2i$. Thus (3.9) takes the form

$$D^2 S_n(x) - xDS_n(x) + nS_n(x) = 0, \tag{3.11}$$

which is the lattice analog of 3.9. This similarity between difference and differential formulations of quantum mechanics has arisen in a very natural way. The polynomials $S_n(x)$ emerge from a formulation of quantum mechanics on a discrete-time lattice using the method of finite elements (see Ref. 7). Apparently, the Hahn polynomials $S_n(x)$ are the natural basis for the states in discrete-time quantum mechanics just as the Hermite polynomials $\mathrm{He}_n(x)$ are the natural coordinate space basis in the continuum. Both sets of polynomials arise from the same underlying Heisenberg algebra $[q,p] = i$.

IV. FURTHER MATHEMATICAL PROPERTIES OF $S_n(x)$

A. Raising and lowering operators

It is well known that the Hermite polynomials $\mathrm{He}_n(x)$ listed in (3.8) can be constructed by means of raising and lowering operators $a^\dagger, a$. If we define $a = d/dx$ and $a^\dagger = x - d/dx$, we find that

$$a\,\mathrm{He}_n(x) = n\,\mathrm{He}_{n-1}(x),$$

$$a^\dagger\,\mathrm{He}_n(x) = \mathrm{He}_{n+1}(x),$$

where $[a,a^\dagger] = 1$.

In similar fashion we can find a lowering operator A for the Hahn polynomials $S_n(x)$:

$$A = \tan\left(\frac{d}{dx}\right). \tag{4.1}$$

The operator A has the property that

$$AS_n(x) = \begin{cases} S_{n-1}(x), & n>0, \\ 0, & n=0, \end{cases} \tag{4.2}$$

which follows easily from the integral representation for $S_n(x)$ in (2.6). The operator A is well defined as a convergent Taylor series in powers of d/dx.

By the same logic the raising operator $A^\dagger$ is

$$A^\dagger = \cot\left(\frac{d}{dx}\right). \tag{4.3}$$

Evidently this operator exists only in a formal sense; $A^\dagger$ is actually a nonlocal integral operator.

Note that since both A and $A^\dagger$ are functions of d/dx only, we seem to conclude formally that

$$[A,A^\dagger] = 0, \tag{4.4}$$

in contrast with $[a,a^\dagger] = 1$. Indeed, the identity (4.4) holds when $[A,A^\dagger]$ operates on $S_n(x)$ $(n>0)$. However, (4.4) is false when it operates on $S_0(x) = 1$ apparently because of ambiguities associated with the definition of $[d/dx]^{-1}$.

B. Representation of $S_n(x)$ as a generalized hypergeometric function

In Ref. 3 a general four-parameter class of continuous Hahn polynomials $P_n(x)$ is described. In this paper $P_n(x)$ satisfies an orthogonality relation

$$\int_{-\infty}^{\infty} P_n(x)P_m(x)W(x)dx = \frac{\Gamma(n+a+c)\Gamma(n+a+d)\Gamma(n+b+c)\Gamma(n+b+d)}{(2n+a+b+c+d-1)\Gamma(n+a+b+c+d-1)}\delta_{n,m}, \tag{4.5}$$

where

$$P_n(x) = i^n \frac{\Gamma(a+c+n)\Gamma(a+d+n)}{\Gamma(a+c)\Gamma(a+d)n!}\,{}_3F_2(-n,n+a+b+c+d-1,a-ix;a+c,a+d;1) \tag{4.6}$$

and

$$W(x) = \frac{\Gamma(a+ix)\Gamma(b+ix)\Gamma(c-ix)\Gamma(d-ix)}{2\pi}. \tag{4.7}$$

We obtain the connection between $S_n(x)$ and $P_n(x)$ if we expand our weight function in (2.13) as a product of four gamma functions. By comparing the result with (4.7) we can identify the parameters a,b,c,d:

$$a = c = \tfrac{1}{4}, \quad b = d = \tfrac{3}{4}. \tag{4.8}$$

From this result and a comparison of the orthogonality relations in (4.5) and (2.14), we can identify our polynomials $S_n(x)$ as generalized hypergeometric functions of argument 1:

$$S_n(x) = \frac{i^n}{\sqrt{n!}}\,{}_3F_2\left(-n,n+1,\frac{1}{4}-\frac{ix}{4};\frac{1}{2},1;1\right). \tag{4.9}$$

The connection between our continuous Hahn polynomials and $3j$ symbols is given in Eq. (1) of Ref. 3.

C. Asymptotic behavior of $S_n(x)$ and distribution of zeros

One of the most prominent features of polynomials are their zeros. Using MACSYMA we computed the first one hundred polynomials $S_n(x)$ and their zeros. We found that the zeros of $S_n(x)$ are real and occur in symmetrical pairs centered about $x = 0$ in a band that ranges just above $x = -2n$ to just below $x = +2n$. Our numeral results also suggest that as $n \to \infty$ the separation between adjacent zeros near a fixed value of x slowly vanishes. In addition, the zeros of $S_n(x)$ and $S_{n+1}(x)$ interlace.

We now present an explicit asymptotic analysis which demonstrates the correctness of the above numerical observations for large n. We begin by rewriting the integral representation for $S_n(x)$ in (2.6) in Laplace form:

$$S_n(x) = \oint \frac{dz}{2\pi i}\psi(z)e^{\phi(z)}, \tag{4.10}$$

where

$$\psi(z) = 1/z(1+z^2)^{1/2}, \tag{4.11}$$

$$\phi(z) = -n \ln z + x \arctan z \tag{4.12}$$

It is simplest to use the method of steepest descents[9] to evaluate $S_n(x)$ in (4.10) as $n \to \infty$. We find the saddle points by solving $\phi'(z) = 0$. There are two saddle points located at

$$z_{\pm} = [x \pm (x^2 - 4n^2)^{1/2}]/2n. \tag{4.13}$$

Note that the value of $z_{\pm}$ is a function of n. If $|x| < 2n$ then $z_{\pm} \sim x/2n \pm i$ $(n \to \infty)$.

The controlling factor[9] (the most rapidly varying component) of the leading asymptotic behavior of $S_n(x)$ as $n \to \infty$ is given by

$$e^{\phi(z_+)} + e^{\phi(z_-)}. \tag{4.14}$$

When $\phi(z_{\pm})$ is imaginary $S_n(x)$ is oscillatory and when $\phi(z_{\pm})$ is real $S_n(x)$ is growing and not oscillatory. From (4.13) we can see that the transition between these two distinct behaviors occurs when $z_{\pm}$ changes from complex to real. This happens at $|x| = 2n$. Thus, asymptotically, the zeros ζ_k $(k = 1,2,3,...,n)$ of $S_n(x)$ must lie in the range $-2n \leqslant \zeta_k \leqslant 2n$.

It is easy to illustrate this result numerically using ζ_n, the largest positive zero of $S_n(x)$. We have determined that

$$\zeta_{10}/10 = 1.3428, \quad \zeta_{20}/20 = 1.5638,$$
$$\zeta_{25}/25 = 1.6191, \quad \zeta_{30}/30 = 1.6594,$$
$$\zeta_{50}/50 = 1.7520, \quad \zeta_{100}/100 = 1.8399.$$

If we extrapolate these values of ζ_n/n we find that $\zeta_n/n \to 2$ as $n \to \infty$.

Next, we substitute (4.13) into (4.14) to obtain an explicit form for the controlling factor of the asymptotic behavior of $S_n(x)$ for large n. When $-2n \leqslant x \leqslant 2n$ we obtain the oscillatory controlling factor

$$\cos(x \ln(4ne/x)^{1/2}).$$

Thus the k th zero ζ_k of $S_n(x)$ satisfies the equation

$$\zeta_k \ln(4ne/\zeta_k)^{1/2} \sim (k - (n+1)/2)\pi \qquad (n \to \infty, \quad k = 1,2,...,n). \tag{4.15}$$

Therefore the separation $\Delta = \zeta_{k+1} - \zeta_k$ between two large consecutive zeros of $S_n(x)$ satisfies the asymptotic formula

$$\Delta \sim 2\pi/\ln(4n/\bar{\zeta}) \quad (n \to \infty, \quad 1 \ll \bar{\zeta} \ll 2n), \tag{4.16}$$

where $\bar{\zeta} = (\zeta_k + \zeta_{k+1})/2$. Note that for fixed $\bar{\zeta}$, Δ decays logarithmically as n increases.

The asymptotic result in (4.16) is extremely accurate. In Table I we compare the exact value of Δ with the asymptotic prediction in (4.16) for the positive zeros of $S_{50}(x)$. Observe that except near $k = 25$ and $k = 50$ the relative error is much less than 1%.

From (4.15) it is also evident that the zeros of consecutive polynomials $S_n(x)$ and $S_{n+1}(x)$ must interlace.

D. Expansions of functions in series of $S_n(x)$

In this subsection we construct a general procedure for expanding an arbitrary function $f(x)$ as a series in $S_n(x)$:

$$f(x) = \sum_{n=0}^{\infty} a_n S_n(x). \tag{4.17}$$

Let us assume that $f(t)$ has a Fourier transform representation:

TABLE I. A comparison between the exact values of and the asymptotic predictions for the differences between pairs of consecutive zeros of $S_{50}(x)$. The zeros of $S_{50}(x)$ are labeled ζ_k $(k = 1,2,...,50)$. Only the positive zeros $(k = 26,27,...,50)$ are listed. The exact separation between consecutive zeros is denoted $\Delta_k^{(\text{exact})} = \zeta_{k+1} - \zeta_k$. The asymptotic prediction $\Delta_k^{(\text{asymptotic})}$ is given in (43). Except for the smallest and largest values of ζ the agreement between $\Delta_k^{(\text{exact})}$ and $\Delta_k^{(\text{asymptotic})}$ is very strong.

k	ζ_k	$\Delta_k^{(\text{exact})}$	$\Delta_k^{(\text{asymptotic})}$
26	0.4885	1.1522	1.2001
27	1.6407	1.3915	1.4121
28	3.0323	1.5757	1.5874
29	4.6079	1.7370	1.7464
30	6.3449	1.8887	1.8971
31	8.2337	2.0364	2.0443
32	10.2701	2.1833	2.1907
33	12.4534	2.3319	2.3385
34	14.7853	2.4841	2.4894
35	17.2694	2.6417	2.6448
36	19.9111	2.8066	2.8063
37	22.7177	2.9808	2.9755
38	25.6985	3.1667	3.1541
39	28.8651	3.3671	3.3439
40	32.2322	3.5858	3.5474
41	35.8181	3.8276	3.7673
42	39.6457	4.0989	4.0073
43	43.7446	4.4091	4.2720
44	48.1536	4.7720	4.5677
45	52.9257	5.2098	4.0034
46	58.1355	5.7605	5.2925
47	63.8960	6.4976	5.7566
48	70.3936	7.5927	6.3358
49	77.9863	9.6115	7.1239
50	87.5978	...	...

$$f(t) = \int_{-\infty}^{\infty} ds\, e^{ist} F(s). \tag{4.18}$$

From the orthogonality relation in (2.14), the general formula for the coefficients a_n is

$$a_n = \int_{-\infty}^{\infty} dt \frac{S_n(t) f(t)}{2\cosh(\pi t/2)}. \tag{4.19}$$

Substituting the expression for $f(t)$ in (4.18) and the integral representation for S_n in (2.6) gives

$$a_n = \frac{1}{2\pi i} \int_{-\infty}^{\infty} ds\, F(s) \oint \frac{dz}{z^{n+1}(1+z^2)^{1/2}} \times \int_{-\infty}^{\infty} dt \frac{e^{t(is + \arctan z)}}{2\cosh(\pi t/2)}, \tag{4.20}$$

where we have interchanged orders of integration.

We can evaluate the t integral using (2.15):

$$a_n = \frac{1}{2\pi i} \int_{-\infty}^{\infty} ds\, F(s) \oint dz \frac{\sec(is + \arctan z)}{z^{n+1}(1+z^2)^{1/2}}.$$
$$= \int_{-\infty}^{\infty} ds \frac{F(s)}{\cosh s} \frac{1}{2\pi i} \oint \frac{dz}{z^{n+1}} \frac{1}{1 - iz \tanh s}.$$

Next, we evaluate the z integral by expanding the denominator and integrating term by term:

$$a_n = i^n \int_{-\infty}^{\infty} \frac{ds\, F(s)}{\cosh s} (\tanh s)^n. \tag{4.21}$$

We now consider some special cases.

Example 1: $f(x)=\delta(x)$, $F(s)=(2\pi)^{-1}$. Using the general formula[10]

$$\int_0^\infty \frac{\sinh^\mu x}{\cosh^\nu x}=\frac{1}{2}\frac{\Gamma((\mu+1)/2)\Gamma((\nu-\mu)/2)}{\Gamma((\nu+1)/2)} \tag{4.22}$$

we obtain

$$a_{2p}=\frac{(-1)^p\Gamma(p+\frac{1}{2})}{p!2\sqrt{\pi}},\quad a_{2p+1}=0.$$

Thus

$$\delta(x)=\sum_{p=0}^{\infty} S_{2p}(x)\frac{(-1)^p\Gamma(p+\frac{1}{2})}{p!2\sqrt{\pi}}. \tag{4.23}$$

Note that this formula is a special case of the general statement of completeness of the polynomials $S_n(x)$:

$$\delta(x-a)=\sum_{n=0}^{\infty}\frac{S_n(x)S_n(a)}{2[\cos h(\pi x/2)\cosh(\pi a/2)]^{1/2}}. \tag{4.24}$$

If we set $a=0$ in (4.24) and use the property that

$$\begin{aligned} &S_{2n}(0)=\Gamma(n+\tfrac{1}{2}(-1)^n/n!\sqrt{\pi},\\ &S_{2n+1}(0)=0, \end{aligned} \tag{4.25}$$

we recover (4.23).

Example 2: $f(x)=\delta(x-a)$, $F(s)=e^{-isa}/2\pi$. Here we are rederiving the expansion in (4.24). For this choice of $F(s)$ the integral (4.21) can be done by considering a rectangular complex contour whose vertices are located at $(-\infty,+\infty,+\infty+i\pi,-\infty+i\pi)$. This contour encloses the $(n+1)$-order pole at $s=i\pi/2$. Translating variables $s=iz+i\pi/2$ and comparing with the expansion in (4.24) gives an interesting Rodrigues-like formula for $S_n(x)$:

$$S_n(x)=\frac{1}{n!}\lim_{z\to 0}\left(\frac{d}{dz}\right)^n\frac{z^{n+1}e^{xz}(\cos z)^n}{(\sin z)^{n+1}}. \tag{4.26}$$

Example 3: $f(x)=[2\cosh(\pi t/2)]^{-1}$, $F(s)=(2\pi\cosh s)^{-1}$. For this case we evaluate (4.21) using (4.22) and obtain a formula for the expansion of the weight function $W(x)$:

$$\frac{1}{2\cosh(\pi x/2)}=\sum_{p=0}^{\infty}\frac{S_{2p}(x)(-1)^p}{\pi(2p+1)}. \tag{4.27}$$

Example 4: $f(x)=\sin(ax)$, $F(s)=(1/2i)[\delta(s-a)-\delta(s+a)]$. In this case we obtain from (4.21)

$$\sin(ax)=\sum_{p=0}^{\infty}\frac{(-1)^p(\tanh a)^{2p+1}}{\cosh a}S_{2p+1}(x). \tag{4.28}$$

Similarly, we have

$$\cos(ax)=\sum_{p=0}^{\infty}\frac{(-1)^p(\tanh a)^{2p}}{\cosh a}S_{2p}(x). \tag{4.29}$$

Finally, combining (4.28) and (4.29) and replacing a by $-ia$, we have

$$e^{ax}=\sum_{n=0}^{\infty}\frac{(\tan a)^n}{\cos a}S_n(x), \tag{4.30}$$

which is valid as long as $|a|<\pi/2$. This formula was used in Ref. 7.

ACKNOWLEDGMENTS

We thank R. Askey for an informative discussion. One of us, L. R. M., thanks Washington University for its hospitality.

We are grateful to the U. S. Department of Energy for financial support.

[1]W. Hahn, Math. Nachr. **2**, 4 (1949).
[2]N. M. Atakishiyev and S. K. Suslov, J. Phys. A **18**, 1583 (1985).
[3]R. Askey, J. Phys. A **18**, L 1017 (1985).
[4]For a complete list of references on discrete Hahn polynomial and associated functions see Refs. 2 and 3.
[5]I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products* (Academic, New York, 1965), 3.523(4).
[6]See Ref. 5, 3.511(4).
[7]This identity was first observed in C. M. Bender, L. R. Mead, and S. Pinsky, Phys. Rev. Lett. **56**, 2445 (1986).
[8]See Ref. 5, p. xxxv.
[9]C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (McGraw-Hill, New York, 1978).
[10]See Ref. 5, p. 344, 3.572. Note that this formula is given incorrectly in the first edition.

The Gel'fand realization and the generating function of the Clebsch–Gordan coefficients of SL(2,*R*) in the hyperbolic basis

Debabrata Basu
Department of Physics, Indian Institute of Technology, Kharagpur-721302, West Bengal, India

(Received 19 February 1986; accepted for publication 24 September 1986)

It is shown that the canonical realization of the representations of SL(2,*R*) proposed by Gel'fand and co-workers yields a generating function of the Clebsch–Gordan coefficients of the group in the hyperbolic basis. This function is the coupled state and appears as the solution of an ordinary differential equation reducible to the hypergeometric equation. The desired expansion of the generating function that yields the Clebsch–Gordan coefficients is essentially a generalization of Barnes' theory of analytic continuation of the hypergeometric function. In this paper the normalized Clebsch–Gordan coefficients for the coupling of two representations of the positive discrete class are calculated. The final result is an analytic continuation of the corresponding expression in the SO(2) basis. The possible application of the generating function to the reduction of the Kronecker product of three irreducible representations is discussed.

I. INTRODUCTION

The Clebsch–Gordan problem of the group SL(2,*R*) was investigated by Pukanszky,[1] Holman and Biedenharn,[2] Ferretti and Verde,[3] Wang,[4] Verdiev, Kerimov, and Smorodinskii,[5] Barut and Wilson,[6] and us[7] among others.[8] Pukanszky[1] confined his attention to the structure of the Clebsch–Gordan (CG) series for the coupling of two representations of the continuous class. However, he did not attempt the remaining couplings or the problem of explicit evaluation of the Clebsch–Gordan coefficients (CGC's). These aspects of the problem were considered by Holman and Biedenharn[2] (HB), Wang,[4] and us.[7] HB based their investigations on the fundamental recurrence relation satisfied by the CGC's. Their first paper was mainly concerned with the coupling of two representations of the discrete class, and other cases of coupling were considered in the second paper. The CG series was obtained by them by examining the resolvent of the Laplace–Beltrami operator in the space of Bargmann's representation functions. The non-normalized CGC's determined by Ferretti and Verde[3] formed the starting point of the investigations of Wang[4] who attempted to normalize them by adopting a summation prescription originally due to HB. All these authors used the compact SO(2) basis for the unitary irreducible representations (UIR's) of SL(2,*R*). More recently we made a departure from the previous practice by evaluating the CGC's in the noncompact $E(1)$ basis.[9] The problem of evaluation of the CGC's in the hyperbolic SO(1,1) basis was attempted some time ago by Mukunda and Radhakrishnan.[10] However, some of their results given in terms of the generalized hypergeometric functions of the ${}_3F_2(1)$ type turn out to be divergent. This may be attributed to their use of the oscillator realization, which does not seem to be particularly suitable for this problem.

In this paper we make a fresh attack on this problem along entirely different lines. We show that the realization of the representations of SL(2,*R*) proposed by Gel'fand and co-workers[11] constitutes a convenient starting point for the CG problem of the group in the SO(1,1) basis. In a previous paper[12] (I) we analyzed this representation space in some detail and obtained the unitary transformations connecting the three subgroup reductions. We now show that the use of the Gel'fand realization leads to a generating function of the CGC's in the continuous SO(1,1) basis. A similar generating function made its appearance some time ago in connection with the CG problem of SL(2,*R*) in the compact SO(2) basis.[7] The generators of the group were constructed in the space of homogeneous functions of two complex variables ξ_1, ξ_2 transforming according to the fundamental representation of SU(1,1) [isomorphic to SL(2,*R*)], which is essentially the Bargmann realization.[13] The bases of the coupled representation f_{jm} were then shown to satisfy an ordinary differential equation reducible to the hypergeometric equation. The CGC's then become identical with the coefficients of Fourier or Taylor expansion of an appropriate solution of this equation. The connection of this approach with that of HB can be established by writing the series solution of this equation in the form $\Sigma\, a_{m_2} x^{m_2}$. Substitution of this solution in the differential equation yields

$$(j_2+m_2+1)(j_1-m_1+1)a_{m_2+1} + [j_1(j_1+1)+j_2(j_2+1)-j(j+1)+2m_1m_2] \times a_{m_2} + (j_2-m_2+1)(j_1+m_1+1)a_{m_2-1}=0.$$

This recurrence relation is completely equivalent to the recurrence relation of the CGC's derived by HB. However, since the SO(1,1) basis is continuous no such discrete recurrence relation exists in this basis in the usual sense.

On the other hand, the bases of the coupled representation in the Gel'fand realization still satisfy an ordinary differential equation of second order. This equation turns out to be formally the same as the one in the SO(2) basis, but with m replaced by $-i\lambda$. This simplification may be attributed to the close similarity between the monomial eigenbases of the continuous SO(1,1) basis in the Gel'fand realization and those of the discrete SO(2) basis in the Bargmann realiza-

 0022-2488/87/030514-06$02.50

tion. The basic difference in the use of the generating function in the two problems lies in the intrinsic difference in the structure of the representation spaces of Gel'fand and Bargmann. This has the consequence of not only restricting the values of j_1, j_2, but also the domain of the variable x of the differential equation, which is determined by the exponentiability of the generators to the UIR's of the group. For example, for the Kronecker product of two positive representations, the variable x in both the problems is the ratio of two complex numbers x_1 and x_2, but the domain of the two variables as well as the scalar product in the Hilbert space (see Sec. II) are entirely different. In the Gel'fand realization x_1 and x_2 represent two complex numbers each spanning the upper half-plane $\mathrm{Im}\, x_1 > 0$, $\mathrm{Im}\, x_2 > 0$ whereas in the Bargmann realization x_1 and x_2 are complex numbers varying over the open unit disk $0 < |x_1| < 1, 0 < |x_2| < 1$. The solution of the above differential equation, which is once again reducible to the hypergeometric (HG) equation by a simple substitution, constitutes the generating function of the CGC's in the SO(1,1) basis.

We start with the coupling of two UIR's belonging to the positive discrete class. Since the SO(1,1) basis spans a continuum we look for an expansion of the generating function as an integral over the continuous SO(1,1) state label. In a sense this expansion is a generalization of Barnes' theory of analytic continuation[14] for the product of a binomial and a hypergeometric function (HGF). Although the generating function has two different Taylor expansions inside and outside the unit circle, it represents a single analytic function. It then follows from Barnes' theory that the desired integral representation must be the same in all regions of the complex plane. The coefficient of the product state in this integral is the unnormalized CGC. To get the normalized CGC we compare this integral with the inverse expansion which is essentially the CG series.

II. THE FUNDAMENTAL EQUATION AND THE DISCRETE PART OF THE SPECTRUM

The group SL(2,R) [isomorphic to SU(1,1)] consists of all 2×2 real matrices with determinant 1. In the realization of Gel'fand and co-workers,[11] the representations of SL(2,R) are constructed in the space $D_{(j,\epsilon)}$ of functions $f(x)$ of a single real or complex variable x. As shown in Paper I the generators J_1, J_2, J_3 can be represented as differential operators of the form

$$\begin{aligned} J_1 &= -i\left[\frac{(1-x^2)}{2}\frac{d}{dx} + jx\right], \\ J_2 &= i\left[x\frac{d}{dx} - j\right], \\ J_3 &= i\left[\frac{(1+x^2)}{2}\frac{d}{dx} - jx\right]. \end{aligned} \tag{2.1}$$

The generators (2.1) can be exponentiated to the representations of the positive discrete class when x is a complex variable spanning the half-plane $\mathrm{Im}\, x > 0$. The representation space then consists of functions analytic in the upper half-plane and the generators (2.1) are Hermitian under the scalar product

$$(f_1, f_2) = \frac{i}{2\Gamma(-2j-1)} \iint_{\mathrm{Im}\, x>0} f_1(x)\bar{f}_2(x) \times (\mathrm{Im}\, x)^{-2j-2}\, dx\, d\bar{x}\,. \tag{2.2}$$

For the negative discrete class the representation space consists of functions analytic in the lower half-plane and the scalar product is given by

$$(f_1, f_2) = \frac{i}{2\Gamma(-2j-1)} \iint_{\mathrm{Im}\, x<0} f_1(x)\bar{f}_2(x) \times |\mathrm{Im}\, x|^{-2j-2}\, dx\, d\bar{x}\,. \tag{2.3}$$

The principal and the exceptional series of representations are realized, on the other hand, in the Hilbert space of functions defined on the real line. The generators (2.1) are Hermitian for the principal series, under a local scalar product, and, for the exceptional series, under a nonlocal scalar product. Although the intrinsic structure of the representation space is different for each class of representation the formal differential operators (2.1) are the same for all UIR's.

We shall now consider the Kronecker product

$$T_g^{j_1} \times T_g^{j_2}\,.$$

The variables for the carrier space of the representations are, respectively, x_1 and x_2, which can be complex or real depending on the nature of the representations coupled. In the hyperbolic SO(1,1) basis, the product states are "monomials" of the form

$$x_1^{j_1 - i\lambda_1} x_2^{j_2 - i\lambda_2}, \tag{2.4}$$

when j_1 and j_2 belong to the UIR's of the discrete class. For the representations of the principal or exceptional series these are distributions of the form (see Paper I)

$$(x_1 \pm i0)^{j_1 - i\lambda_1} (x_2 \pm i0)^{j_2 - i\lambda_2}\,. \tag{2.5}$$

Although the product states (2.5) are fundamentally different from (2.4) the formal operations presented below can be justified for both.

By definition, the coupled states $g_{j\lambda}$ are the simultaneous eigenstates

$$(J_3^2 - J_1^2 - J_2^2) g_{j\lambda} = j(j+1) g_{j\lambda}\,, \tag{2.6a}$$

$$J_2 g_{j\lambda} = i\left(x_1\frac{\partial}{\partial x_1} + x_2\frac{\partial}{\partial x_2} - j_1 - j_2\right) g_{j\lambda} = \lambda g_{j\lambda}\,. \tag{2.6b}$$

Equation (2.6b) implies that $g_{j\lambda}$ is a homogeneous function of degree $(j_1 + j_2 - i\lambda)$ in x_1 and x_2. This suggests the following transformations:

$$\begin{aligned} &x_1 \to x_1, \quad x_2 \to x_1 x\,, \quad x = x_2/x_1\,, \\ &g_{j\lambda}(x_1, x_2) = x_1^{j_1 + j_2 - i\lambda} e_{j\lambda}(x)\,. \end{aligned} \tag{2.7}$$

Now we have to convert the partial derivatives acting on functions of x_1 and x_2 into those acting on functions of x_1 and $x = x_2/x_1$. This can be done by noting

$$\frac{\partial}{\partial x_1} \to \frac{\partial}{\partial x_1} - \frac{x}{x_1}\frac{\partial}{\partial x}, \quad \frac{\partial}{\partial x_2} \to \frac{1}{x_1}\frac{\partial}{\partial x}. \tag{2.8}$$

Using Eqs. (2.7) and (2.8) and eliminating the variable x_1 in Eq. (2.6a) we obtain, after some calculations the ordinary differential equation satisfied by the function $e_{j\lambda}(x)$,

$$x(1-x)^2\frac{d^2e_{j\lambda}}{dx^2}+\left[4j_2-2i\lambda-2j_2x-2\right.$$
$$\left.+\frac{j_1-j_2+i\lambda+1}{x}-x(j_1+j_2-i\lambda-1)\right]x\frac{de_{j\lambda}}{dx}$$
$$-[j_2(j_2-1)-j_1(j_1+1)+j(j+1)$$
$$-2i\lambda j_2-2j_2x(j_1+j_2-i\lambda)]e_{j\lambda}=0\,. \tag{2.9}$$

It is interesting to note that the differential equation (2.9) is formally identical to the corresponding equation in Ref. 7 with m replaced by $-i\lambda$.

For the determination of the spectrum of j-values appearing in the reduction it seems necessary to express the solution of Eq. (2.9) in terms of known functions of analysis. This is done by the substitution

$$e_{j\lambda}(x)=(1-x)^{\sigma-j-1}F(x)\,, \tag{2.10a}$$
$$\sigma=j_1+j_2+1\,, \tag{2.10b}$$

which reduces Eq. (2.9) to the standard differential equation satisfied by the HGF:

$$x(1-x)\frac{d^2F}{dx^2}+[2j+(1-x)(j_0+i\lambda+1-2j)]$$
$$\times\frac{dF}{dx}+(j-i\lambda)(j_0-j)F=0\,, \tag{2.11a}$$
$$j_0=j_1-j_2\,. \tag{2.11b}$$

When j belongs to the discrete spectrum, the appropriate solution is given by

$$F=F(-j+i\lambda,j_0-j;\,-2j;\,1-x)\,. \tag{2.12}$$

When j lies in the continuous spectrum we have to take suitable linear combinations of the first and second solution of the HG equation (2.11a).

The discrete part of the spectrum is obtained by applying the operator

$$K_2=J_2^{(1)}-J_2^{(2)}=i\left(x_1\frac{\partial}{\partial x_1}-2x\frac{\partial}{\partial x}-j_0\right) \tag{2.13}$$

to the coupled eigenfunction

$$g_{j\lambda}=x_1^{j_1+j_2-i\lambda}(1-x)^{\sigma-j-1}$$
$$\times F(-j+i\lambda,j_0-j;\,-2j;\,1-x)\,, \tag{2.14}$$

and using the Hermiticity condition. Operating $g_{j\lambda}$ by the operator (2.13) we have

$$K_2\,g_{j\lambda}(x)=\frac{-i(j^2+\lambda^2)(j_0^2-j^2)(\sigma+j)}{2j^2(4j^2-1)}g_{j-1\lambda}$$
$$+\frac{\lambda\sigma j_0}{j(j+1)}g_{j\lambda}+2i(\sigma-j-1)g_{j+1\lambda}\,. \tag{2.15}$$

This result is obtained by using the recurrence relations,

$$(1-x)\frac{dF}{dx}=\frac{ab}{c}F-\frac{ab(c-b)(c-a)}{c^2(c+1)}$$
$$\times xF(a+1,b+1;c+2)\,, \tag{2.16a}$$
$$(1-x)F=[ab(c-a)(c-b)/c^2(c^2-1)]$$
$$\times x^2F(a+1,b+1;c+2)$$
$$-[c(c-a-b-1)+2ab/c(c-2)]xF$$
$$+F(a-1,b-1;c-2)\,, \tag{2.16b}$$

with

$$F=F(a,b;c;x)\,,\ \text{etc.}$$

We now introduce the normalized coupled eigenbases

$$f_{j\lambda}=N_{j\lambda}g_{j\lambda}\,, \tag{2.17}$$

where $N_{j\lambda}$ is the normalization constant. The Hermiticity of K_2, i.e.,

$$(K_2f_{j\lambda},f_{j-1\lambda'})=(f_{j\lambda},K_2f_{j-1\lambda'})\,, \tag{2.18}$$

now yields

$$\left|\frac{N_{j\lambda}}{N_{j-1\lambda}}\right|^2=\frac{(\bar\sigma^2-j^2)4j^2(4j^2-1)}{(j^2+\lambda^2)(j_0^2-j^2)|\sigma+j|^2}\,. \tag{2.19}$$

This equation determines the normalization factor and the range of j-values but with a degree of uncertainty. First, Eq. (2.19) asserts that $N_{0\lambda}=N_{-1/2\,\lambda}=0$ so that the identity and $D_{-1/2}$ representations do not appear in the reduction. Second, since the remaining factor on the rhs of (2.19) is positive, the ratio $|N_{j\lambda}/N_{j-1\lambda}|^2$ will be positive if

$$(\bar\sigma^2-j^2)/(j_0^2-j^2)>0.$$

We shall analyze this condition case by case. Let us first consider the coupling of two discrete representations, i.e., $D_{j_1}^+\times D_{j_2}^+$ or $D_{j_1}^-\times D_{j_2}^+$. For this case $(\sigma+j)/(|j_0|-j)<0$. Therefore we must have $(\sigma-j)/(|j_0|+j)<0$. This is possible if

$$j=\sigma-1,\sigma-2,\ldots,-\infty\,,$$

or if

$$j=-|j_0|,\,-|j_0|+1,\ldots,-1\ (\text{or}\ -\tfrac{1}{2})\,.$$

The first region has an upper boundary at $j=j_1+j_2$, which j cannot cross, and the second region has a lower boundary at $j=-|j_0|$. The first case corresponds to the coupling $D_{j_1}^+\times D_{j_2}^+$ and the second case to $D_{j_1}^-\times D_{j_2}^+$. For other cases of coupling involving the continuous representations, Eq. (2.19) permits all values of $j\leqslant -1,\,-\frac{1}{2}$ to appear in the reduction.

III. THE COUPLING $D_{j_1}^+\times D_{j_2}^+$

A. The CG series

To start with, we consider the coupling of two UIR's of the positive discrete class. For the determination of the normalized CGC's and the complete spectrum of j-values it is necessary to expand the monomials $x^{j_2-i\lambda_2}$ in terms of the coupled eigenfunctions $e_{j\tau}$. The expansion coefficients will be the complex conjugates of the CGC's. We start with the identity[15]

$$x^\mu=\sum_{r=0}^{\infty}\frac{(-1)^r(a)_r(b)_r}{(c+r-1)_r r!}$$
$$\times{}_3F_2\left[\begin{matrix}-\mu,&c+r-1,&-r\\ a,&b&\end{matrix}\right](1-x)^r$$
$$\times F(a+r,b+r;c+2r;1-x)\,, \tag{3.1}$$

with

$$\mu=j_2-i\lambda_2,\quad r=\sigma-j-1,\quad a=1-\sigma+i\tau\,,$$
$$b=-2j_2,\quad c=2-2\sigma,\quad \tau=\lambda_1+\lambda_2\,. \tag{3.2}$$

A little simplification yields the desired expansion:

$$x^{j_2-i\lambda_2}=\sum_{-\infty}^{j_1+j_2}(-)^{\sigma-j-1}\times\Gamma\begin{bmatrix}-j+i\lambda, & j_0-j, & -\sigma-j\\ 1-\sigma+i\tau, & -2j_2, & -2j-1,\quad \sigma-j\end{bmatrix}\times{}_3F_2\begin{bmatrix}-j_2+i\lambda_2, & -\sigma-j, & j+1-\sigma\\ 1-\sigma+i\tau, & -2j_2\end{bmatrix}e_{j\tau}\,. \tag{3.3}$$

B. Expansion of the generating function and Barnes' theory of analytic continuation

We shall now consider the problem of expansion of the coupled state $e_{j\lambda}(x)$ in terms of the product states $x^{j_2-i\lambda_2}$. Since the spectrum of λ_2 is the entire real line, the desired expansion of $e_{j\lambda}(x)$ is of the form

$$e_{j\lambda}(x)=\int_{-i\infty}^{i\infty}a(z)x^{j_2-z}\,dz\,, \tag{3.4}$$

where the path of integration is the entire imaginary axis. The coefficient $a(z)$ is then the unnormalized CGC of SL(2,R) in the hyperbolic basis.

We note that the coupled state $e_{j\lambda}(x)$, being the product of a binomial and a HGF, defines a single analytic function of the complex variable x. The Taylor expansions of $e_{j\lambda}$ inside and outside the unit circle $|x|=1$ are, however, different. When the contour of (3.4) is closed on the left we get the Taylor expansion of $e_{j\lambda}$ for $|x|<1$, and when it is closed on the right we get the expansion for $|x|>1$. But either of the expansions represents the same analytic function. It therefore follows that the expansions of $e_{j\lambda}(x)$ in the two regions $|x|<1$ and $|x|>1$ must yield the same coefficient $a(z)$. In short, (3.4) is essentially the same as the expansion of the HGF in Barnes' theory of analytic continuation,[14] except for a shift of the path of integration. This shift is necessary to avoid the poles of the integrand which may otherwise lie on the path of integration. But the beauty of Barnes's theory is that the integrand has the *same* form in all regions of the complex x plane so that it defines a single analytic function. The same conclusion holds good for the CGC's $a(z)$ in our approach.

To evaluate the coefficient $a(z)$ we first expand the coupled state $e_{j\lambda}(x)$ in a Taylor series and rewrite the series as an integral over the imaginary axis. We start from the Taylor expansion inside the unit circle $|x|<1$. The HGF appearing in $e_{j\lambda}(x)$ has a branch point at the origin $x=0$. To get the Taylor expansion about the origin we therefore apply the following formula for analytic continuation[16]:

$$F(a,b,c,1-x)=\Gamma\begin{bmatrix}c, & c-a-b\\ c-a, & c-b\end{bmatrix}F(a,b,a+b-c+1,x)+\Gamma\begin{bmatrix}c, & a+b-c\\ a, & b\end{bmatrix}\times x^{c-a-b}F(c-a,c-b,c-a-b+1,x)\,. \tag{3.5}$$

Using Eq. (3.5) and the formula[17]

$$(1-x)^{\mu}F(a,b,c,x)=\sum_n\frac{(a)_n(b)_n}{(c)_n n!}\,{}_3F_2\times\begin{bmatrix}-\mu, & 1-c-n, & -n\\ 1-a-n, & 1-b-n\end{bmatrix}x^n\,, \tag{3.6}$$

we obtain after some calculation

$$\Gamma\begin{bmatrix}-j+i\lambda, & j_0-j, & -j-i\lambda, & -j_0-j\\ & -2j\end{bmatrix}e_{j\lambda}(x)=X_j^{j_0,i\lambda}+x^{-j_0-i\lambda}X_j^{-j_0,-i\lambda}\,, \tag{3.7}$$

where

$$X_j^{j_0,i\lambda}=\sum\frac{(-)^n}{n!}\times\Gamma\,[-j+i\lambda+n, j_0-j+n,\ -j_0-i\lambda-n]\times{}_3F_2\begin{bmatrix}j+1-\sigma, & -j_0-i\lambda-n, & -n\\ 1+j-i\lambda-n, & 1+j-j_0-n\end{bmatrix}x^n\,. \tag{3.8}$$

The various terms in the sum (3.7) are the residues at the simple poles $z=j_2-n$ and $z=j_1+i\lambda-n$ ($n=0,1,2,\ldots$) of the analytic functions

$$\chi(z)=\Gamma[-j+i\lambda+j_2-z,\ j_1-j-z,\ z-j_1-i\lambda,\ z-j_2]\times{}_3F_2\begin{bmatrix}j+1-\sigma, & -j_1-i\lambda+z, & -j_2+z\\ 1+j-i\lambda-j_2+z, & 1-j_1+j+z\end{bmatrix}x^{j_2-z}\,. \tag{3.9}$$

Besides the singularities at $z=j_2-n$, $z=j_1+i\lambda-n$, which lie on the left of the imaginary axis, the function $\chi(z)$ has simple poles at $z=-j+i\lambda+j_2+n$ and $z=j_1-j+n$, which are situated on the right. The function has, in addition, simple poles at the points where one of the denominator parameters of the generalized HGF becomes a negative integer (denominator catastrophe), i.e., at $d\equiv 1+j-i\lambda-j_2+z=-n$ and $e\equiv 1+j-j_1+z=-n$ ($n=0,1,2,\ldots$). It can be shown that the only possible singularities because of the denominator catastrophe lie on the right of the imaginary axis.

Let us now choose a contour C consisting of an infinite semicircle on the left and the pure imaginary axis. The singularities enclosed by the contour are the simple poles at $z=j_2-n$ and $z=j_1+i\lambda-n$. Therefore by Cauchy's theorem,

$$\frac{1}{2\pi i}\int_C\chi(z)dz=\sum_{n=0}^{\infty}\mathrm{Res}[\chi(z)]_{z=j_2-n}+\sum_{n=0}^{\infty}\mathrm{Res}[\chi(z)]_{z=j_1+i\lambda-n}\,, \tag{3.10}$$

and we obtain

$$\begin{bmatrix}-j+i\lambda, & j_0-j, & -j-i\lambda, & -j_0-j\\ & -2j\end{bmatrix}e_{j\lambda}(x)=\frac{1}{2\pi i}\int_C\chi(z)dz\,. \tag{3.11}$$

Using the asymptotic forms of the gamma function and the generalized HGF, it can be shown that the function $\chi(z)$ goes rapidly to zero as $|z|\to\infty$ in the region $\mathrm{Re}\,z<0$ provided $|\arg x|<\pi$. Thus

$$e_{j\lambda}(x)=\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}dz\,\Gamma\begin{bmatrix}-j+i\lambda+j_2-z, & j_1-j-z, & z-j_1-i\lambda, & z-j_2, & -2j\\ -j+i\lambda, & j_0-j, & -j-i\lambda, & -j_0-j\end{bmatrix}$$
$$\times{}_3F_2\begin{bmatrix}j+1-\sigma, & -j_1-i\lambda+z, & -j_2+z\\ 1+j-i\lambda-j_2+z, & 1+j-j_1+z\end{bmatrix}x^{j_2-z}, \tag{3.12}$$

where the path of integration is the entire imaginary axis. This integral representation was derived under the condition $|x|<1$. However, the integral represents a single analytic function $e_{j\lambda}(x)$. Using the principle of analytic continuation we can now assert that the integral representation outside the unit circle (i.e., $|x|>1$) must be identical and the condition $|x|<1$ can be dropped. It can in fact be verified by explicit calculation that the above integral when closed on the right will yield the Taylor expansion of $e_{j\lambda}(x)$ for $|x|>1$. The formula (3.12) is therefore valid in all regions of the complex x plane.

C. Evaluation of the normalized CGC's

Let us now introduce the normalized product and coupled states

$$f_{j_1\lambda_1}f_{j_2\lambda_2}=N_{j_1\lambda_1}N_{j_2\lambda_2}x_1^{j_1-i\lambda_1}x_2^{j_2-i\lambda_2}, \tag{3.13a}$$

$$f_{j\lambda}=N_{j\lambda}x_1^{j_1+j_2-i\lambda}e_{j\lambda}(x), \tag{3.13b}$$

where the normalization factors $N_{j_1\lambda_1}$, $N_{j_2\lambda_2}$, and $N_{j\lambda}$ are chosen in such a way that the product and the coupled states are orthonormal:

$$\begin{aligned}&(f_{j_1\lambda_1}f_{j_2\lambda_2},f_{j_1\lambda_1'}f_{j_2\lambda_2'})=\delta(\lambda_1-\lambda_1')\delta(\lambda_2-\lambda_2'),\\ &(f_{j\lambda},f_{j'\lambda'})=\delta_{jj'}\delta(\lambda-\lambda').\end{aligned} \tag{3.14}$$

We now write the expansion of $e_{j\lambda}(x)$ in the form

$$e_{j\lambda}=\int d\lambda_2'\,a(\lambda_2')x^{j_2-i\lambda_2'}, \tag{3.15}$$

where $a(\lambda_2')$ is given by Eq. (3.12). Using Eqs. (3.13)–(3.15) we get

$$C\begin{pmatrix}j_1 & j_2 & j\\ \lambda_1 & \lambda_2 & \lambda\end{pmatrix}=(f_{j\lambda},f_{j_1\lambda_1},f_{j_2\lambda_2})$$
$$=\delta(\lambda-\lambda_1-\lambda_2)(N_{j\lambda}/N_{j_1\lambda_1}N_{j_2\lambda_2})a(\lambda_2). \tag{3.16}$$

On the other hand, from the CG series we have

$$x^{j_2-i\lambda_2}=\sum_{j=-\infty}^{j_1+j_2}b(j)e_{j\tau}, \tag{3.17}$$

where $b(j)$ is given by Eq. (3.3). This yields

$$\bar{C}\begin{pmatrix}j_1 & j_2 & j\\ \lambda_1 & \lambda_2 & \lambda\end{pmatrix}=(f_{j_1\lambda_1}f_{j_2\lambda_2},f_{j\lambda})$$
$$=\delta(\lambda-\lambda_1-\lambda_2)(N_{j_1\lambda_1}N_{j_2\lambda_2}/N_{j\lambda})b(j). \tag{3.18}$$

Equations (3.16) and (3.18) require that

$$\left|\frac{N_{j\lambda}}{N_{j_1\lambda_1}N_{j_2\lambda_2}}\right|^2=\frac{\bar{b}(j)}{a(\lambda_2)}. \tag{3.19}$$

Thus $\bar{b}(j)/a(\lambda_2)$ must be a positive definite quantity. To ensure this we shall first show that the generalized HGF appearing in $a(\lambda_2)$ can be transformed into the complex conjugate of the one appearing in $b(j)$. We start from the Thomae–Whipple identity,[18]

$${}_3F_2\begin{bmatrix}a, & b, & c\\ d, & e\end{bmatrix}=(-)^{-c}\Gamma\begin{bmatrix}1+b-e & 1+a-e, & d, & e\\ d-e & e-c, & 1+b+c-e, & 1+a+c-e\end{bmatrix}$$
$$\times{}_3F_2\begin{bmatrix}c, & 1-s, & 1+c-e\\ 1+b+c-e, & 1+a+c-e\end{bmatrix}, \tag{3.20}$$

where $s=d+e-a-b-c$ and $c=j+1-\sigma$ is a negative integer. From this we obtain

$$\frac{\bar{b}(j)}{a(\lambda_2)}=2\pi\Gamma\begin{bmatrix}-j-i\lambda, & -j+i\lambda, & j_0-j, & -j_0-j, & -\sigma-j\\ -2j-1, & \sigma-j, & -2j, & -j_1-i\lambda+i\lambda_2, & -j_1+i\lambda-i\lambda_2, & -j_2+i\lambda_2, & -j_2-i\lambda_2\end{bmatrix} \tag{3.21}$$

which is a positive definite quantity. This ensures the correctness of our result. Combining all these results we now obtain the final expression for the CGC:

$$C\begin{pmatrix}j_1 & j_2 & j\\ \lambda_1 & \lambda_2 & \lambda\end{pmatrix}=\frac{(-)^{\sigma-j-1}}{\sqrt{2\pi}}\delta(\lambda-\lambda_1-\lambda_2)$$
$$\times\left\{\Gamma\begin{bmatrix}-j-i\lambda, & j_0-j, & -\sigma-j, & -2j, & -j_1-i\lambda_1, & -j_1+i\lambda_1, & -j_2+i\lambda_2, & -j_2-i\lambda_2\\ & & & -j+i\lambda, & -j_0-j, & -2j-1, & \sigma-j\end{bmatrix}\right\}^{1/2}$$
$$\times\Gamma\begin{bmatrix}1\\ -2j_2, & 1-\sigma-i\lambda\end{bmatrix}{}_3F_2\begin{bmatrix}j+1-\sigma, & -\sigma-j, & -j_2-i\lambda_2\\ & -2j_2, & 1-\sigma-i\lambda\end{bmatrix}, \tag{3.22}$$

where we have omitted a phase. This expression is essentially the analytic continuation of the corresponding formula for the SO(2) basis.[2,6,7]

IV. CONCLUSION

The principal advantage of the Gel'fand realization is that it provides a convenient starting point for many practical calculations particularly those requiring explicit reduction under the noncompact SO(1,1) or $E(1)$ subgroups. In this paper we have shown that this realization yields a generating function of the CGC's in the continuous SO(1,1) basis. The differential equation whose solution yields the generating function is the analog of the recurrence relation of HB. In this paper we have used this function to evaluate the CGC's of SL(2,R) in the SO(1,1) basis for the coupling of two UIR's of the positive discrete class. For other cases of coupling like $D_{j_1}^- \times D_{j_2}^+$ or $D_{j_1}^c \times D_{j_2}^+$, the CG series has contribution from the UIR's of the principal series. This presents an additional difficulty because the representations of the SO(1,1) subgroup within the principal series are doubly degenerate. However, the difficulty can be circumvented by taking suitable linear combinations of the solutions of the HG equation. Calculations for this case are under way and the results will be communicated shortly.

The Gel'fand realization may also turn out to be helpful for the reduction of the Kronecker product of three irreducible representations of SL(2,R). There are several sets of commuting operators which may be diagonalized simultaneously for the coupling of three UIR's D_{j_1}, D_{j_2}, and D_{j_3}. A particularly convenient set is

$$\mathbf{J}^{(1)^2},\ \mathbf{J}^{(2)^2},\ \mathbf{J}^{(3)^2},\ \mathbf{J}_{\text{int}}^2,\ \mathbf{J}^2,\ J_3\ ,$$

where $\mathbf{J}^{(1)^2} = J_3^{(1)^2} - J_1^{(1)^2} - J_2^{(1)^2}$, etc. In the above set $D_{j_{\text{int}}}$ can be the UIR contained in $D_{j_1} \times D_{j_2}$, $D_{j_2} \times D_{j_3}$, or $D_{j_3} \times D_{j_1}$. Following the notation of Rose[19] the connection between the first two couplings is given by

$$e_{j\lambda}(j') = \sum_{j''} R_{j''j'} e_{j\lambda}(j'')\ ,$$

where Σ stands for the summation over the discrete and integration over the continuous j''-values and $R_{j''j'}$ is the Racah coefficient of SL(2,R). Since $e_{j\lambda}(j')$ and $e_{j\lambda}(j'')$ are expressible in terms of HGF's, the problem essentially consists in expanding a HGF in terms of a series of other HGF's. A variety of formulas of this genre have been derived by Burchnall and Chaundy[20] and it is interesting to see whether one of them followed, if necessary, by a Sommerfeld Watson transformation yields the Racah coefficient of SL(2,R). This problem will be treated in a forthcoming paper.

ACKNOWLEDGMENT

The author wishes to thank Dr. G. P. Sastry for many helpful discussions.

[1] L. Pukanszky, Trans. Am. Math. Soc. **100**, 116 (1961).
[2] W. J. Holman and L. C. Biedenharn, Ann. Phys. (NY) **39**, 1 (1966); **47**, 205 (1968).
[3] I. Ferretti and M. Verde, Nuovo Cimento **55**, 110 (1968); M. Verde, *Proceedings of the Mendeleevian Conference Torino, 1969* (Torino Academy of Science, Italy, 1971), p. 305.
[4] K. H. Wang, J. Math. Phys. **11**, 2077 (1970).
[5] Y. A. Verdiev, G. A. Kerimov, and Ya. A. Samorodinskii, Yad. Fiz. Sov. J. Nucl. Phys. **20**, 827 (1974) [Sov. J. Nucl. Phys. **20**, 411 (1975)].
[6] A. O. Barut and R. Wilson, J. Math. Phys. **17**, 900 (1976).
[7] D. Basu and S. D. Majumdar, J. Math. Phys. **20**, 492 (1979); S. D. Majumdar, J. Math. Phys. **17**, 193 (1976).
[8] M. Andrews and J. Gunson, J. Math. Phys. **5**, 1391 (1964); S. Sannikov, Dokl. Akad. Nauk SSSR **171**, 1058 (1966) [Sov. Phys. Dokl. **11**, 1045 (1967)]; H. Ui, Ann. Phys. (NY) **49**, 69 (1969); E. Chacon, D. Levi, and M. Moshinsky, J. Math. Phys. **17**, 1919 (1976).
[9] D. Basu and K. B. Wolf, J. Math. Phys. **24**, 478 (1983).
[10] N. Mukunda and B. Radhakrishnan, J. Math. Phys. **15**, 1320, 1332, 1643, 1656 (1974).
[11] I. M. Gel'fand, M. I. Graev, and N. Ya. Vilenkin, *Generalized Functions* (Academic, New York, 1966), Vol. 5, Chap VII.
[12] D. Basu and T. Bhattacharya, J. Math. Phys. **26**, 12 (1985).
[13] V. Bargmann, Ann. Math. **48**, 568 (1947); A. O. Barut and C. Fronsdal, Proc. R. Soc. London Ser. A **287**, 532 (1965), D. Basu, J. Math. Phys. **18**, 742 (1977).
[14] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge U. P., Cambridge, 1969), Chap. XIV, pp. 286–291.
[15] See Ref. 7, p. 494.
[16] See Ref. 14, p. 291.
[17] *Higher Transcendental Functions*, edited by A. Erdelyi (McGraw-Hill, New York, 1953), Vol. I, p. 187.
[18] L. J. Slater, *Generalized Hypergeometric Functions* (Cambridge U. P., Cambridge, 1966), Chap. 4, pp. 114–120.
[19] M. E. Rose, *Elementary Theory of Angular Momentum* (Wiley, New York, 1957), p. 107.
[20] See Ref. 17, p. 187; J. L. Burchnall, J. Lond. Math. Soc. **23**, 253 (1948); J. L. Burchnall and T. W. Chaundy, Proc. Lond. Math. Soc. (2) **50**, 56 (1958); T. W. Chaundy, Q. J. Oxford Ser. **13**, 159 (1942); **14**, 55 (1943).

Nonlinear equations with superposition formulas and the exceptional group G_2. II. Classification of the equations

J. Beckers and V. Hussin[a)]
Physique théorique et mathématique, Université de Liège, Institut de Physique au Sart Tilman, B.5, B-4000 Liege 1, Belgium

P. Winternitz
Centre de recherches mathématiques, Université de Montréal, C.P. 6128, Succ. A, Montreal, Québec, H3C 3J7, Canada

(Received 18 June 1986; accepted for publication 22 October 1986)

Nonlinear equations with superposition formulas are obtained, corresponding to the action of the complex and real forms of the exceptional Lie group G_2 on the homogeneous spaces G_2/H. The isotropy group of the origin H is taken as one of the maximal parabolic subgroups of G_2, or as one of the maximal reductive subgroups, leaving some vector space $V\in\mathbb{C}^7$ (or $V\in\mathbb{R}^7$) invariant. The parabolic subgroups, as well as the simple subgroups $SL(3,\mathbb{C})$, $SU(3)$, $SL(3,\mathbb{R})$, or $SU(2,1)$ lead to equations with quadratic or quartic nonlinearities. The semisimple subgroups $SL(2,\mathbb{C})\otimes SL(2,\mathbb{C})$, $SU(2)\otimes SU(2)$, and $SU(1,1)\otimes SU(1,1)$ lead to equations with quadratic nonlinearities and additional nonlinear constraints on the independent variables.

I. INTRODUCTION

Part I of this series[1] (further referred to as I) was devoted to a classification of the maximal subalgebras of the complex and real forms of the exceptional Cartan Lie algebra g_2 and to an analysis of their matrix realizations.

In this paper we make use of the results of I to construct the homogeneous spaces G/H. We take G to be the complex exceptional Lie group $G_2(\mathbb{C})$, the real compact Lie group $G_2^C(\mathbb{R})$, or the real noncompact Lie group $G_2^{NC}(\mathbb{R})$, and H to be one of the corresponding maximal subgroups. The realizations of the homogeneous spaces are then used to obtain systems of nonlinear ordinary differential equations (ODE's) with superposition formulas, based on the action of the group G on the space G/H.

Such equations will in general have the form

$$\frac{dx^\mu}{dt}\equiv\dot{x}^\mu=\sum_{i=1}^{k}a_i(t)\hat{\xi}_i x^\mu,\quad \mu=i,...,n, \tag{1.1}$$

where the $a_i(t)$ are arbitrary functions of t and the $\hat{\xi}_i$ are vector fields representing the Lie algebra L of the Lie group G, when acting on the homogeneous space G/H:

$$\hat{\xi}_i=\sum_{\nu=1}^{n}\xi_i^\nu(x^1,...,x^n)\frac{\partial}{\partial x^\nu},\quad i=1,...,k. \tag{1.2}$$

The general solution of Eq. (1.1) can be written as a function of a finite number m of particular solutions and of n significant constants c_j:

$$x^\mu(t)=F^\mu(\mathbf{x}_1(t),...,\mathbf{x}_m(t),c_1,...,c_n),\quad \mu=1,...,n. \tag{1.3}$$

It is (1.3) that we call a "superposition formula" and $\mathbf{x}_1(t),...,\mathbf{x}_m(t)$ is a "fundamental set of solutions."

These concepts were originally introduced by Lie, who also gave the necessary and sufficient conditions for a system of ODE's to allow a (nonlinear) superposition formula.[2]

Out interest in ODE's with superposition formulas was motivated in earlier publications.[3-10] Let us mention, as far as mathematical interest is concerned, that the application of the theory of transitive primitive Lie algebras has made it possible to solve a problem posed by Lie, namely to classify the systems of "indecomposable" ODE's with superposition formulas.[6] From the practical point of view the superposition formulas provide a new method for obtaining analytical or numerical solutions of certain systems of nonlinear ODE's.[9,10] Finally, from the physical point of view, it should be stressed that ODE's of the type (1.1) occur in many applications.[5-9] A prime example are matrix Riccati equations,[5] occuring as Bäcklund transformations for the nonlinear σ model[11,12] and in many engineering applications,[13] specially in optimal control theory.[9]

Tables of all maximal subalgebras of the Lie algebras $g_2(\mathbb{C})$, $g_2^C(\mathbb{R})$, and $g_2^{NC}(\mathbb{R})$ are given in Ref. 1. Use was made of seven-dimensional irreducible representations of these algebras and the corresponding Lie groups. A given maximal subalgebra can be imbedded in this representation either reducibly or irreducibly. A subalgebra h imbedded reducibly in g, by definition leaves a proper nontrivial subspace of $\mathbb{C}^7$ (or $\mathbb{R}^7$) invariant. This makes the construction of the corresponding homogeneous space G/H much easier. In this article we restrict ourselves to the case of reducibly imbedded subalgebras and we obtain the ODE's for all such cases. The problem for irreducibly imbedded subgroups has only been partially solved, even for the classical groups.[6,7]

We shall see below that the homogeneous spaces we are interested in, i.e., $G_2(\mathbb{C})/H$, $G_2^C(\mathbb{R})/H$, and $G_2^{NC}(\mathbb{R})/H$, where H is any one of the corresponding maximal subgroups (I), can always be imbedded into $O(7,\mathbb{C})/\tilde{H}$, $O(7)/\tilde{H}$, or $O(4,3)/\tilde{H}$, respectively. Here $\tilde{H}$, is again a maximal subgroup of the corresponding compact or noncompact rotation group. For certain subgroups H and $\tilde{H}$ the spaces G_2/H

[a)] Chargé de recherches F.N.R.S., Belgium.

 0022-2488/87/030520-10$02.50

and $O(7,F)/\tilde{H}$ are actually locally diffeomorphic (and in particular have the same dimension). The G_2 equations for such spaces will be special cases of the $O(7,F)$ equations. For other subgroups we find $\dim[G_2/H] < \dim[O(7,F)/\tilde{H}]$. The G_2/H space is then properly contained in the other one and we must find the conditions that restrict the larger space to the lower-dimensional one. Typically the G_2 equations are then obtained as $O(7,F)$ equations with further constraints imposed on the dependent variables. In principle, if not necessarily in practice, these constraints can be solved and some redundant variables can be eliminated from the system. In the case of maximal reductive subalgebras the constraints are obtained via the invariance properties of the completely antisymmetric tensor T discussed in Ref. 1.

Section II is devoted to equations related to maximal parabolic subalgebras of $g_2(\mathbb{C})$ and $g_2^{NC}(\mathbb{R})$. In each case two different maximal parabolic subalgebras exist, one involving additional constraints with respect to the $O(7,\mathbb{C})$ or $O(4,3)$ case, the other not.

The simple maximal subgroups $SL(3,\mathbb{C}) \subset G_2(\mathbb{C})$, $SU(3) \subset G_2^C(\mathbb{R})$, $SL(3,\mathbb{R}) \subset G_2^{NC}(\mathbb{R})$, and $SU(2,1) \subset G_2^{NC}(\mathbb{R})$ are treated in Sec. III and are shown to lead to special cases of projective Riccati equations[3,4] (with no nonlinear constraints).

The reducibly imbedded semisimple subgroups $SL(2,\mathbb{C}) \otimes SL(2,\mathbb{C}) \subset G_2(\mathbb{C})$, $SU(2) \otimes SU(2) \subset G_2^C(\mathbb{R})$, $SU(2) \otimes SU(2) \subset G_2^{NC}(\mathbb{R})$, and $SU(1,1) \otimes SU(1,1) \subset G_2^{NC}(\mathbb{R})$ are shown in Sec. IV to lead to rectangular matrix Riccati equations with four additional nonlinear constraints leading to more complicated nonlinearities in the equations.

II. EQUATIONS RELATED TO MAXIMAL PARABOLIC SUBGROUPS

A. General form of the equations

It was shown in I that the noncompact groups $G_2(\mathbb{C})$ and $G_2^{NC}(\mathbb{R})$ have two mutually nonisomorphic maximal parabolic subgroups each. We denote them $P_{\alpha a}(F)$, with $a = 1$ or 2, and $F = \mathbb{C}$ or $\mathbb{R}$; the corresponding maximal parabolic subalgebras of $g_2(\mathbb{C})$ and $g_2^{NC}(\mathbb{R})$ are $P_{\alpha a}(F)$.

To simplify the presentation, we shall consider the case $F = \mathbb{C}$ explicitly. All formulas of this section are equally valid for $F = \mathbb{R}$, with the complex orthogonal group $O(7,\mathbb{C})$ replaced by the real pseudo-orthogonal group $O(4,3)$. Similarly, all subgroups of $O(7,\mathbb{C})$ restrict to the relevant subgroups of $O(4,3)$.

In order to obtain the nonlinear ODE's with superposition formulas we need to construct a coordinate realization of the homogeneous spaces $G_2(\mathbb{C})/P_{\alpha a}(\mathbb{C})$. To do this we first construct the corresponding homogeneous spaces $O(7,\mathbb{C})/P_a(\mathbb{C})$, where $P_a(\mathbb{C})$ is a maximal parabolic subgroup of $O(7,\mathbb{C})$ leaving an a-dimensional completely isotropic vector space invariant. We then restrict from $O(7,\mathbb{C})$ to $G_2(\mathbb{C})$ and impose further constraints whenever necessary. It was shown in I that we have $P_1(\mathbb{C}) = SIM(5,\mathbb{C})$ and $P_2(\mathbb{C}) = OPT(5,\mathbb{C})$, i.e., $P_1(\mathbb{C})$ and $P_2(\mathbb{C})$ are the similitude and "optical" groups[14] of $\mathbb{C}^7$, respectively.

The spaces $O(7,\mathbb{C})/P_a(\mathbb{C})$ can be realized as the Grassmannians of complex isotropic a-planes in $\mathbb{C}^{7\times a}$.

We use the "antidiagonal" metric, given by the symmetric form J_N with $(J_N)_{ik} = \delta_{i,N-k}$. The Lie algebra $o(N,\mathbb{C})$ is represented by matrices $M \in \mathbb{C}^{N\times N}$ satisfying $J_N M + M^T J_N = 0$ (T denotes transposition). We have

$$J = \begin{bmatrix} & & J_a \\ & J_{N-2a} & \\ J_a & & \end{bmatrix},$$
$$M = \begin{bmatrix} A & B^T & C \\ D & E & -J_{N-2a} B J_a \\ H & -J_a D^T J_{N-2a} & -J_a A^T J_a \end{bmatrix}, \tag{2.1}$$

$a = 1,2,\ldots,[N/2]$, $A,C,H \in \mathbb{C}^{a\times a}$; $B,D \in \mathbb{C}^{(N-2a)\times a}$, $E \in \mathbb{C}^{(N-2a)\times(N-2a)}$, $J_a C + C^T J_a = 0$, $J_a H + H^T J_a = 0$, $J_{N-2a} E + E^T J_{N-2a} = 0$. The Lie algebra $g_2(\mathbb{C})$ is obtained by setting $N = 7$ and imposing further conditions on the entries in (2.1). More specifically, M of (2.1) coincides with M'' of (I.4.21) and $A,\ldots,H$ can be read off from (I.4.21), separately for $a = 1$ and 2 (notice the correction of a sign misprint with respect to Ref. 8).

We shall use both homogeneous and affine coordinates on the Grassmannian $O(N,\mathbb{C})/P_a(\mathbb{C})$ of isotropic a-planes. Homogeneous coordinates are given by the matrix elements of the rectangular matrix

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix},$$

$$X_1, X_3 \in \mathbb{C}^{a\times a}, \quad X_2 \in \mathbb{C}^{(N-2a)\times a}, \quad \operatorname{rank} X = a, \tag{2.2}$$

satisfying the isotropy condition

$$X^T J X = X_1^T J_a X_3 + X_3^T J_a X_1 + X_2^T J_{n-2a} X_2 = 0. \tag{2.3}$$

As usual, the homogeneous coordinates are highly redundant, i.e., the matrices X and Xg with $g \in GL(a,\mathbb{C})$ describe the same point. To remove this redundancy we choose the point $X_1 = 0$, $X_2 = 0$, $X_3 = I_a$ as the origin on $O(N,\mathbb{C})/P_a(\mathbb{C})$ and introduce affine coordinates as components of the rectangular matrix

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix},$$
$$Z_1 = X_1 X_3^{-1} \in \mathbb{C}^{a\times a}, \quad Z_2 = X_2 X_3^{-1} \in \mathbb{C}^{(N-2a)\times a} \tag{2.4}$$

(in the neighborhood of the origin we have $\det X_3 \neq 0$). The isotropy condition (2.3) in affine coordinates is

$$Z_1^T J_a + J_a Z_1 = -Z_2^T J_{N-2a} Z_2. \tag{2.5}$$

Thus, the "J-symmetric" part of the matrix Z_1 is not independent. The coordinates on $O(N,\mathbb{C})/P_a(\mathbb{C})$ can be identified with components of the two matrices

$$Z_2 \quad \text{and} \quad R = Z_1^T J_a - J_a Z_1. \tag{2.6}$$

Following the usual procedure[3-8] we can now write the nonlinear ODE's corresponding to the action of $G_2(\mathbb{C})$ on $G_2(\mathbb{C})/P_a(\mathbb{C})$.

In homogeneous coordinates we have a set of linear equations

$$\dot{X} = MX, \quad \dot{X} \equiv \frac{dX}{dt}, \tag{2.7}$$

with the nonlinear constraint (2.3) and the redundancy $X \sim Xg, g \in \mathrm{GL}(a,\mathbb{C})$. Going over to affine coordinates, we get rid of the redundancy. Moreover, we can eliminate the J-symmetric part of Z_1 from the equations, using (2.5). In terms of the variables in (2.6) we obtain the final form of the $\mathrm{O}(N,\mathbb{C})/P_a(\mathbb{C})$ equations:

$$\begin{aligned}
\dot{R} &= 2C^T J_a + Z_2^T(BJ_a) - (J_a B^T)Z_2 + R(J_a A^T J_a) + (J_a A J_a)R \\
&\quad + \tfrac{1}{2}\{Z_2^T(J_{N-2a}DJ_a)R + R(J_a D^T J_{N-2a})Z_2\} + \tfrac{1}{2}R(HJ_a)R \\
&\quad + \tfrac{1}{2}\{Z_2^T J_{N-2a}(Z_2 J_a D^T - DJ_a Z_2^T)J_{N-2a}Z_2\} + \tfrac{1}{2}(Z_2^T J_{N-2a}Z_2)HJ_a(Z_2^T J_{N-2a}Z_2), \\
\dot{Z}_2 &= -J_{N-2a}BJ_A + EZ_2 + Z_2(J_a A^T J_a) - \tfrac{1}{2}(DJ_a)R + \tfrac{1}{2}Z_2(HJ_a)R \\
&\quad + Z_2(J_a D^T J_{N-2a})Z_2 - \tfrac{1}{2}(DJ_a)Z_2^T J_{N-2a} + \tfrac{1}{2}Z_2(HJ_a)Z_2^T J_{N-2a}Z_2.
\end{aligned} \tag{2.8}$$

The matrices A, B, C, D, E, and H are given functions of t, satisfying the conditions given in (2.1) (for all t).

Equations (2.8) can also be viewed as equations based on the action of the real group $\mathrm{O}(N/2,N/2)$ (N even) or $\mathrm{O}((N+1)/2,(N-1)/2)$ (N odd) on the corresponding Grassmannian of real isotropic a-planes (with $1 \leqslant a \leqslant [N/2]$). In this case the matrices $R,Z_2,A,\ldots,H$ are real.

In order to obtain the equations related to the action of the group G_2 on the space $\mathrm{G}_2/P_{\alpha a}$ we set $N=7$ in all the above equations and consider the cases $a=1$ and $a=2$ separately.

B. The $\mathrm{G}_2/P_{\alpha_1}$ equations

We have shown in I that we have

$$\begin{aligned}
&P_{\alpha 1}(\mathbb{C}) \sim \mathrm{G}_2(\mathbb{C}) \cap \mathrm{SIM}(5,\mathbb{C}), \\
&P_{\alpha 1}(\mathbb{R}) \sim \mathrm{G}_2^{\mathrm{NC}}(\mathbb{R}) \cap \mathrm{SIM}(3,2).
\end{aligned} \tag{2.9}$$

The corresponding homogeneous spaces are diffeomorphic, i.e., we have

$$\begin{aligned}
&\mathrm{G}_2(\mathbb{C})/P_{\alpha 1}(\mathbb{C}) \sim \mathrm{O}(7,\mathbb{C})/\mathrm{SIM}(5,\mathbb{C}), \\
&\mathrm{G}_2^{\mathrm{NC}}(\mathbb{R})/P_{\alpha 1}(\mathbb{R}) \sim \mathrm{O}(4,3)/\mathrm{SIM}(3,2),
\end{aligned} \tag{2.10}$$

[with $\mathrm{SIM}(5,\mathbb{C}) = P_1(\mathbb{C})$, $\mathrm{SIM}(3,2) = P_1(\mathbb{R})$]. Indeed, the dimensions satisfy

$$\begin{aligned}
&\dim \mathrm{G}_2/P_{\alpha 1} = 14 - 9 = 5, \\
&\dim \mathrm{O}(7,\mathbb{C})/\mathrm{SIM}(5,\mathbb{C}) = 21 - 15 = 5,
\end{aligned}$$

and it is easy to verify that $\mathrm{G}_2(\mathbb{C})$ acts transitively on $\mathrm{O}(7,\mathbb{C})/\mathrm{SIM}(5,\mathbb{C})$. The same is true in the real case. We concentrate on the complex case, all results are valid for $\mathrm{G}_2^{\mathrm{NC}}(\mathbb{R})$ as well.

In view of (2.10) we can directly use the coordinates (2.6) for $a=1$, $N=7$. In this case we have $J_a = 1$, $Z_1 \in \mathbb{C}$, hence $R=0$ and also $C=H=0$ in (2.1). Equations (2.8) reduce to complex conformal Riccati equations,[4] which in this case we write as

$$\begin{aligned}
\dot{Z}_2 &= -J_5 B + (E+A)Z_2 + Z_2(D^T J_5)Z_2 \\
&\quad - \tfrac{1}{2}D(Z_2^T J_5 Z_2), \\
&Z_2, B, D \in \mathbb{C}^5, \quad A \in \mathbb{C}, \quad E \in \mathbb{C}^{5\times 5}, \quad J_5 E + E^T J_5 = 0.
\end{aligned} \tag{2.11}$$

The $\mathrm{G}_2(\mathbb{C})/P_{\alpha 1}$ equations are a special case of (2.11), obtained by requiring that the matrix M of (2.1) be an element of $\mathrm{g}_2(\mathbb{C})$. Comparing with (I.4.21) we see that this implies

$$\begin{aligned}
&B^T = (-a_{10}, a_{20}, -\sqrt{2}a_{03}, -a_{13}, -a_{23}), \quad A = -a_2, \\
&D^T = (-a_{01}, a_{02}, -\sqrt{2}a_{30}, -a_{31}, -a_{32}),
\end{aligned}$$

$$E = \begin{pmatrix}
-a_1 + a_2 & a_{21} & \sqrt{2}a_{20} & a_{03} & 0 \\
a_{12} & a_1 - 2a_2 & \sqrt{2}a_{10} & 0 & -a_{03} \\
\sqrt{2}a_{02} & \sqrt{2}a_{01} & 0 & -\sqrt{2}a_{10} & -\sqrt{2}a_{20} \\
a_{30} & 0 & -\sqrt{2}a_{01} & -a_1 + 2a_2 & -a_{21} \\
0 & -a_{30} & -\sqrt{2}a_{02} & -a_{12} & a_1 - a_2
\end{pmatrix}. \tag{2.12}$$

Equations (2.11) and (2.12) provide the $\mathrm{G}_2(\mathbb{C})/P_{\alpha 1}(\mathbb{C})$ ODE's if all the entries Z_2,A,B,D,E are allowed to be complex. The $\mathrm{G}_2(\mathbb{R})/P_{\alpha 1}(\mathbb{R})$ equations are obtained by constraining the above matrices to be real.

C. The $\mathrm{G}_2/P_{\alpha 2}$ equations

This case is somewhat more complicated and more interesting. As shown in I, we have

$$P_{\alpha 2}(\mathbb{C}) \sim \mathrm{G}_2(\mathbb{C}) \cap \mathrm{OPT}(5,\mathbb{C}), \tag{2.13}$$

however, $\mathrm{G}_2(\mathbb{C})$ does not act transitively on $\mathrm{O}(7,\mathbb{C})/\mathrm{OPT}(5,\mathbb{C})$. Indeed, in this case we have

$$\begin{aligned}
&\dim \mathrm{G}_2/P_{\alpha 2} = 14 - 9 = 5, \\
&\dim \mathrm{O}(7,\mathbb{C})/\mathrm{OPT}(5,\mathbb{C}) = 21 - 14 = 7.
\end{aligned} \tag{2.14}$$

Our first task is to provide a model of the space $\mathrm{G}_2/P_{\alpha 2}$ as a

subspace of $O(7,\mathbb{C})/OPT(5,\mathbb{C})$ and to introduce an appropriate coordinate patch.

To do this we use a decomposition of the Lie algebra $g_2(C)$ in the form

$$g_2 = N + p_{\alpha 2}, \tag{2.15}$$

where N is a nilpotent algebra represented by matrices M of the form (2.1) satisfying

$$M = \begin{bmatrix} 0 & B^T & C \\ 0 & 0 & -J_5 B J_2 \\ 0 & 0 & 0 \end{bmatrix},$$

$$B^T = \begin{pmatrix} a_{20} & -\sqrt{2}a_{03} & -a_{13} \\ a_{21} & \sqrt{2}a_{20} & a_{03} \end{pmatrix}, \tag{2.16}$$

$$C = \begin{bmatrix} -a_{23} & 0 \\ 0 & a_{23} \end{bmatrix}$$

[see (I.4.21)]. We parametrize elements of the group $G_2(\mathbb{C})$ as

$$g = e^N g_p, \quad g_p \in P_{\alpha 2}. \tag{2.17}$$

We have

$$e^N = I + N + \tfrac{1}{2}N^2. \tag{2.18}$$

We act with the general element of $G_2(\mathbb{C})$ in the form (2.17) on the origin $(X_1^T, X_2^T, X_3^T) = (0,0,I_2)$ of the Grassmannian and use the fact that g_p leaves the origin invariant. The action of e^N then provides the required coordinates on the Grassmannian $G_2(\mathbb{C})/P_{\alpha 2}(\mathbb{C})$. Explicitly, in affine coordinates, we obtain

$$Z_1 = \begin{bmatrix} \frac{1}{2}v + \frac{1}{2}(xy - uz) & -xz - y^2 \\ -x^2 - uy & -\frac{1}{2}v + \frac{1}{2}(xy - uz) \end{bmatrix},$$

$$Z_2 = \begin{pmatrix} y & z \\ \sqrt{2}x & -\sqrt{2}y \\ u & x \end{pmatrix}. \tag{2.19}$$

The J-antisymmetric part of Z_1 is

$$R = Z_1^T J_2 - J_2 Z_1 = \begin{bmatrix} 0 & v \\ -v & 0 \end{bmatrix}. \tag{2.20}$$

Substituting Z_2 and R into Eqs. (2.8) and writing the elements of the matrix M of (2.1) in the form agreeing with (I.4.21), i.e.,

$$A = \begin{bmatrix} -a_2 & -a_{10} \\ -a_{01} & -a_1 + a_2 \end{bmatrix}, \quad B^T = \begin{bmatrix} a_{20} & -\sqrt{a_{03}} & -a_{13} \\ a_{21} & \sqrt{2}a_{20} & a_{03} \end{bmatrix}, \quad C = \begin{bmatrix} -a_{23} & 0 \\ 0 & a_{23} \end{bmatrix},$$

$$D = \begin{bmatrix} a_{02} & a_{12} \\ -\sqrt{2}a_{30} & \sqrt{2}a_{02} \\ -a_{31} & a_{30} \end{bmatrix}, \quad E = \begin{bmatrix} a_1 - 2a_2 & \sqrt{2}a_{10} & 0 \\ \sqrt{2}a_{01} & 0 & -\sqrt{2}a_{10} \\ 0 & -\sqrt{2}a_{01} & -a_1 + 2a_2 \end{bmatrix}, \quad H = \begin{bmatrix} -a_{32} & 0 \\ 0 & a_{32} \end{bmatrix}. \tag{2.21}$$

We obtain the $G_2(\mathbb{C})/P_{\alpha 2}(\mathbb{C})$ equations as

$$\begin{aligned}
\dot{x} &= \tfrac{1}{2}\{-a_{32}(x^2 y + 2y^2 u + xzu) + 2a_{02}x^2 + 2a_{31}y^2 + 5a_{30}xy + 2a_{12}xu \\
&\quad + a_{32}xv - 4a_{02}yu + a_{30}zu - 2(a_1 - a_2)x + 4a_{01}y - 2a_{10}u - a_{30}v - 2a_{20}\}, \\
\dot{y} &= \tfrac{1}{2}\{a_{32}(xy^2 + 2x^2 z + yzu) - 2a_{12}x^2 + 2a_{30}y^2 + 5a_{02}xy - 4a_{30}xz \\
&\quad - 2a_{31}yz + a_{32}yv + a_{02}zu + 4a_{10}x - 2a_2 y - 2a_{01}z + a_{02}v - 2a_{03}\}, \\
\dot{z} &= \tfrac{1}{2}\{-a_{32}(2y^3 - z^2 u + 3xyz) - 6a_{02}y^2 - 2a_{31}z^2 + 3a_{12}xy + 6a_{30}yz \\
&\quad - a_{12}zu + a_{32}zv - 6a_{10}y + 2(a_1 - 3a_2)z - a_{12}v + 2a_{13}\}, \\
\dot{u} &= \tfrac{1}{2}\{a_{32}(2x^3 - zu^2 + 3xyu) - 6a_{30}x^2 + 2a_{12}u^2 - 3a_{31}xy + 6a_{02}xu \\
&\quad + a_{31}zu + a_{32}uv - 6a_{01}x - 2(2a_1 - 3a_2)u - a_{31}v - 2a_{21}\}, \\
\dot{v} &= \tfrac{1}{2}\{-a_{32}(4x^3 z + 4y^3 u + 3x^2 y^2 - z^2 u^2 + 6xyzu) + a_{12}(2x^3 - zu^2 + 3xyu) \\
&\quad - 3a_{02}(x^2 y + 2y^2 u + xzu) + a_{31}(2y^3 - z^2 u + 3xyz) + 3a_{30}(2x^2 z + xy^2 + yzu) \\
&\quad + a_{32}v^2 + 3a_{02}xv + 3a_{30}yv - a_{31}zv + a_{12}uv - 6a_{03}x + 6a_{20}y - 2a_{21}z - 2a_{13}u - 2a_1 v - 4a_{23}\}.
\end{aligned} \tag{2.22}$$

Thus, we obtain a system of five coupled nonlinear ODE's with polynomial nonlinearities of degree 4. All the coefficients a_{ik} are arbitrary functions of t. Notice that in agreement with the general theory,[6] the cubic and quartic terms have coefficients that already occur in linear or quadratic terms.

The $G_2^{NC}(\mathbb{R})/P_{\alpha 2}(\mathbb{R})$ equations coincide with (2.22), but all coefficients a_{ik} as well as the variables $x,\ldots,v$ must be considered to be real functions of t.

Let us mention that the invariance of the completely antisymmetric tensor T of (I.2.11) under the group G_2 was not explicitly used in the section. It was however used implicitly. It is the invariance of T that imposes two constraints on the seven coordinates of $O(7,\mathbb{C})/OPT(5,\mathbb{C})$. A possible solution of these constraints would lead to the five coordinates x,y,z,u,v of (2.19).

III. EQUATIONS RELATED TO MAXIMAL SIMPLE SUBGROUPS LEAVING ONE-DIMENSIONAL SUBSPACES INVARIANT

We have seen in Ref. 1 that the maximal subgroups of the G_2 groups, leaving one-dimensional nonisotropic subspaces in a seven-dimensional space invariant, are all complex or real forms of SL(3,C). Four different cases occur, namely $G_2(\mathbb{C}) \supset SL(3,\mathbb{C})$, $G_2^{C}(\mathbb{R}) \supset SU(3)$, $G_2^{NC}(\mathbb{R}) \supset SL(3,\mathbb{R})$, and $G_2^{NC}(\mathbb{R}) \supset SU(2,1)$. We shall treat all these cases in a unified manner. To construct the corresponding homogeneous space we first construct a Grassmann manifold of one-planes $SL(7,F)/Aff(6,F)$, with $F = \mathbb{C}$ or $F = \mathbb{R}$, respectively. We then restrict $SL(7,F)$ to $O(7,\mathbb{C})$, $O(7)$, or $O(4,3)$, respectively, and introduce the corresponding metric on the Grassmannian. Finally we restrict to the corresponding G_2 subgroup. Since G_2 acts transitively on the appropriate Grassmannian of nonisotropic one-planes, no further constraints on the coordinates of the Grassmannian pertain and we obtain, in all four cases, special cases of projective Riccati equations.[4]

In order to preserve the unity of presentation, we always, in this section, make use of the diagonal metric.

A. The $G_2(\mathbb{C})/SL(3,\mathbb{C})$ equation

We first construct the $SL(7,\mathbb{C})/Aff(6,\mathbb{C})$ Grassmannian by introducing homogeneous coordinates in $\mathbb{C}^7$ as

$$\begin{pmatrix} \mathbf{x} \\ z \\ \mathbf{y} \end{pmatrix}, \quad \mathbf{x},\mathbf{y}\in\mathbb{C}^3, \quad z\in\mathbb{C}. \tag{3.1}$$

We choose the origin to be the point $\mathbf{x} = \mathbf{y} = 0$, $z = 1$ and remove the redundancy of the homogeneous coordinates by identifying any two points with coordinates satisfying

$$\begin{pmatrix} \mathbf{x} \\ z \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{x}g \\ zg \\ \mathbf{y}g \end{pmatrix}, \quad g\in\mathbb{C}, \quad g\neq 0. \tag{3.2}$$

Notice that the isotropy group of the origin $(0,0,0,1,0,0,0)^T$ is indeed the affine group $Aff(6,\mathbb{C})$ realized by the matrices

$$H \sim \begin{bmatrix} G_{11} & 0 & G_{13} \\ G_{21} & G_{22} & G_{23} \\ G_{31} & 0 & G_{33} \end{bmatrix},$$

$$G_{11},G_{13},G_{31},G_{33}\in\mathbb{C}^{3\times 3}, \quad G_{21},G_{23}\in\mathbb{C}^{1\times 3}, \quad G_{22}\in\mathbb{C}.$$

We restrict to $O(7,\mathbb{C})$ by requiring that the $SL(7,\mathbb{C})$ matrices G satisfy $G^T G = I_7$ (the seven-dimensional identity matrix). Acting on the origin of the Grassmannian constructed above, the $O(7,\mathbb{C})$ matrices sweep out a submanifold of nonisotropic one-planes with homogeneous coordinates satisfying

$$\mathbf{x}^2 + z^2 + \mathbf{y}^2 = 1. \tag{3.3}$$

The isotropy group for the origin reduces to $O(6,\mathbb{C})$.

Further restricting to $G_2(\mathbb{C})$ (in the I_7 realization, see Ref. 1) we notice that the isotropy group of the origin reduces to $SL(3,\mathbb{C})$ [indeed $SL(3,\mathbb{C})$ was obtained in I as the maximal subgroup of $G_2(\mathbb{C})$ leaving a nonisotropic one-dimensional vector space invariant]. Moreover $G_2(\mathbb{C})$ acts transitively on $O(7,\mathbb{C})/O(6,\mathbb{C})$. Comparing dimensions, we have

$$\begin{aligned} &\dim SL(7,\mathbb{C})/Aff(6,\mathbb{C}) = 48 - 42 = 6, \\ &\dim O(7,\mathbb{C})/O(6,\mathbb{C}) = 21 - 15 = 6, \\ &\dim G_2(\mathbb{C})/SL(3,\mathbb{C}) = 12 - 8 = 6. \end{aligned} \tag{3.4}$$

We can now write down the nonlinear equations with superposition formulas in the usual manner.[3,4,7]

In homogeneous coordinates we have the $SL(7,\mathbb{C})/Aff(6,\mathbb{C})$ equations

$$\begin{pmatrix} \dot{\mathbf{x}} \\ \dot{z} \\ \dot{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} R & \mathbf{m} & V \\ \mathbf{a}^T & \mu & \mathbf{n}^T \\ W & \mathbf{b} & U \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ z \\ \mathbf{y} \end{pmatrix},$$

$$R,V,W,U\in\mathbb{C}^{3\times 3}, \quad \mathbf{m},\mathbf{n},\mathbf{a},\mathbf{b}\in\mathbb{C}^3,$$

$$\mu\in\mathbb{C}, \quad \mu = -(\operatorname{Tr} R + \operatorname{Tr} U). \tag{3.5}$$

Removing the redundancy (3.2) by introducing affine coordinates

$$\xi = \mathbf{x}/z, \quad \eta = \mathbf{y}/z, \tag{3.6}$$

we obtain the projective Riccati equations

$$\begin{aligned} \dot{\xi} &= \mathbf{m} + (R - \mu)\xi + V\eta - \xi((\mathbf{a},\xi) + (\mathbf{n},\eta)), \\ \dot{\eta} &= \mathbf{b} + W\xi + (U - \mu)\eta - \eta((\mathbf{a},\xi) + (\mathbf{n},\eta)). \end{aligned} \tag{3.7}$$

The $O(7,\mathbb{C})/O(6,\mathbb{C})$ equations are obtained by requiring that the matrix $M\in\mathbb{C}^{7\times 7}$ in (3.5) should satisfy

$$M + M^T = 0. \tag{3.8}$$

Equations (3.7) simplify to

$$\begin{aligned} \dot{\xi} &= \mathbf{m} + R\xi + V\eta + \xi((\mathbf{m},\xi) - (\mathbf{n},\eta)), \\ \dot{\eta} &= -\mathbf{n} - V^T\xi + U\eta + \eta((\mathbf{m},\xi) - (\mathbf{n},\eta)). \end{aligned} \tag{3.9}$$

Finally, we obtain the $G_2(\mathbb{C})/SL(3,\mathbb{C})$ equations by restricting the matrix M in (3.8) to the Lie algebra $g_2(\mathbb{C})$. Following I we see that this is achieved by putting

$$M = \begin{pmatrix} R & \mathbf{m} & V \\ -\mathbf{m}^T & 0 & \mathbf{n}^T \\ -V^T & -\mathbf{n} & U \end{pmatrix},$$

$$\text{with} \quad \mathbf{m} = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}, \quad \mathbf{n} = \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix}, \tag{3.10}$$

$$R = \frac{1}{2}\begin{bmatrix} 0 & a_3 + m_3 & -a_2 - m_2 \\ -a_3 - m_3 & 0 & a_1 + m_1 \\ a_2 + m_2 & -a_1 - m_1 & 0 \end{bmatrix},$$

$$U = \frac{1}{2}\begin{bmatrix} 0 & a_3 - m_3 & -a_2 + m_2 \\ -a_3 + m_3 & 0 & a_1 - m_1 \\ a_2 - m_2 & -a_1 + m_1 & 0 \end{bmatrix},$$

$$V=\frac{1}{2}\begin{bmatrix} v_{11} & v_{12}+n_3 & v_{13}-n_2 \\ v_{12}-n_3 & v_{22} & v_{23}+n_1 \\ v_{13}+n_2 & v_{23}-n_1 & -v_{11}-v_{22} \end{bmatrix}, \quad (3.11)$$

$m_i,n_i,a_iv_{ik}=v_{ki}\in\mathbb{C}$ in M and in Eqs. (3.9).

B. The $G_2^C(\mathbb{R})/SU(3)$ equations

This case is completely analogous to the $G_2(\mathbb{C})/SL(3,\mathbb{C})$ case treated above. We again start from a Grassmannian, this time $SL(7,\mathbb{R})/Aff(6,\mathbb{R})$ and use homogeneous coordinates, as in (3.1) and (3.2), but with all entries real. The restriction to $O(7)/O(6)$ again leads to the condition (3.3) defining a sphere $S_6\in\mathbb{R}_7$. Finally the $G_2^C(\mathbb{R})/SU(3)$ equations coincide with the equations (3.9)–(3.11) with $m_i,n_i,a_i,v_{ik}\in\mathbb{R}$ in (3.10) and $\xi,\eta\in\mathbb{R}^3$.

C. The $G_2^{NC}(\mathbb{R})/SL(3,\mathbb{R})$ equations

We use the metric $I_{4,3}=\mathrm{diag}(1,1,1,1,-1,-1,-1)$. The $o(4,3)$ matrices satisfy

$$I_{4,3}M+M^TI_{4,3}=0$$

and $g_2^{NC}(\mathbb{R})\subset o(4,3)$ is represented by the matrices (I)

$$M=\begin{pmatrix} R & \mathbf{m} & V \\ -\mathbf{m}^T & 0 & \mathbf{n}^T \\ V^T & \mathbf{n} & U \end{pmatrix}, \quad (3.12)$$

with n, m, R, U, and V as in (3.11), but with real entries.

We introduce homogeneous and affine coordinates on the Grassmannian $SL(7,\mathbb{R})/Aff(6,\mathbb{R})$ as in Secs. III A and III B. Restricting to $O(4,3)$ we see that the orbit of the origin $(0,0,0,1,0,0,0)^T$ under this group is the hyperboloid

$$\mathbf{x}^2+z^2-\mathbf{y}^2=1, \quad (3.13)$$

diffeomorphic to $O(4,3)/O(3,3)$.

The $O(4,3)/O(3,3)$ equations in affine coordinates are

$$\begin{aligned} \dot{\xi}&=\mathbf{m}+R\,\xi+V\eta+\xi((\mathbf{m},\xi)-(\mathbf{n},\eta)), \\ \dot{\eta}&=\mathbf{n}+V^T\xi+U\eta+\eta((\mathbf{m},\xi)-(\mathbf{n},\eta)). \end{aligned} \quad (3.14)$$

The $G_2^{NC}(\mathbb{R})/SL(3,\mathbb{R})$ equations are obtained by taking the values of R, V, U, $\mathbf{m}$, and $\mathbf{n}$ as in (3.11) (and real).

D. The $G_2^{NC}(\mathbb{R})/SU(2,1)$ equations

The diffeomorphism that we are using in this case is $O(4,3)/O(4,2)\sim G_2^{NC}(\mathbb{R})/SU(2,1)$. In order to be able to use the same realization of $G_2^{NC}(\mathbb{R})$ as above and as in Ref. 1, we must choose the origin in $SL(7,\mathbb{R})/Aff(6,\mathbb{R})$ differently than in the previous sections. A convenient choice is the point $(0,0,0,0,0,0,1)^T$ in homogeneous coordinates. The group $SO(4,3)$, realized by matrices satisfying $G^TI_{4,3}G=I_{4,3}$ sweeps out the space

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ z \end{pmatrix}, \quad u_1^2+u_2^2+u_3^2+v_1^2-v_2^2-v_3^2-z^2=-1. \quad (3.15)$$

We introduce affine coordinates

$$\xi=\mathbf{u}/z, \quad \eta=\mathbf{v}/z,$$

and write an element of the $o(4,3)$ algebra as

$$M=\begin{pmatrix} A & B & \mathbf{p} \\ -I_{12}B^T & E & \mathbf{q} \\ \mathbf{p}^T & \mathbf{q}^TI_{12} & 0 \end{pmatrix},$$

$$I_{12}=\begin{pmatrix} 1 & & \\ & -1 & \\ & & -1 \end{pmatrix}, \quad A+A^T=0,$$

$$I_{12}E+E^TI_{12}=0, \quad A,B,E\in\mathbb{R}^{3\times3}, \quad \mathbf{p},\mathbf{q}\in\mathbb{R}^3. \quad (3.16)$$

The $O(4,3)/O(4,2)$ Riccati equations have the form

$$\begin{aligned} \dot{\xi}&=\mathbf{p}+A\,\xi+B\eta-\xi((\mathbf{p},\xi)+(\mathbf{q},I_{12}\eta)), \\ \dot{\eta}&=\mathbf{q}+I_{12}B^T\xi+E\eta-\eta((\mathbf{p},\xi)+(\mathbf{q},I_{12}\eta)). \end{aligned} \quad (3.17)$$

The $G_2^{NC}(\mathbb{R})/SU(2,1)$ equations are obtained by making (3.16) and (3.17) compatible with (3.12) and (3.11), i.e., putting

$$\begin{aligned} A&=R, \quad \mathbf{p}=\frac{1}{2}\begin{pmatrix} V_{13}-n_2 \\ V_{23}+n_1 \\ -V_{11}-V_{22} \end{pmatrix}, \\ \mathbf{q}&=\frac{1}{2}\begin{pmatrix} 2n_3 \\ -a_2+m_2 \\ a_1-m_1 \end{pmatrix}, \\ B&=\frac{1}{2}\begin{pmatrix} 2m_1 & V_{11} & V_{12}+n_3 \\ 2m_2 & V_{12}-n_3 & V_{22} \\ 2m_3 & V_{13}+n_2 & V_{23}-n_1 \end{pmatrix}, \\ E&=\frac{1}{2}\begin{pmatrix} 0 & 2n_1 & 2n_2 \\ 2n_1 & 0 & a_3-m_3 \\ 2n_2 & -a_3+m_3 & 0 \end{pmatrix}. \end{aligned} \quad (3.18)$$

IV. EQUATIONS RELATED TO MAXIMAL SEMISIMPLE SUBGROUPS LEAVING THREE-DIMENSIONAL NONDEGENERATE SUBSPACES INVARIANT

It was shown in Ref. 1 that the invariance of a nondegenerate three-dimensional subspace leads to semisimple subgroups of G_2. More specifically the corresponding subgroups are $SL(2,\mathbb{C})\otimes SL(2,\mathbb{C})\subset G_2(\mathbb{C})$, $SU(2)\otimes SU(2)\subset G_2^C(\mathbb{R})$, $SU(2)\otimes SU(2)\subset G_2^{NC}(\mathbb{R})$, and $SU(1,1)\otimes SU(1,1)\subset G_2^{NC}(\mathbb{R})$. The construction of the corresponding homogeneous spaces and systems of nonlinear ODE's with superposition formulas is of considerable interest, since the invariance of the alternating tensor T [see (I.2.12)] plays a crucial role here.

Let us consider the four cases separately.

A. The $G_2(\mathbb{C})/[SL(2,\mathbb{C})\otimes SL(2,\mathbb{C})]$ equations

It was shown in Ref. 1 that the subgroup $SL(2,\mathbb{C})\otimes SL(2,\mathbb{C})\subset G_2(\mathbb{C})$ can be realized as the intersection $[O(4,\mathbb{C})\otimes O(3,\mathbb{C})]\cap G_2(\mathbb{C})$. We make use of this fact to imbed the homogeneous space $G_2(\mathbb{C})/[SL(2,\mathbb{C})\otimes SL(2,\mathbb{C})]$ into $O(7,\mathbb{C})/[O(4,\mathbb{C})\otimes O(3,\mathbb{C})]$. Moreover, it was shown earlier[6] that this last space can be realized in terms of the Grassmannian of nondegenerate three-planes $G_3(\mathbb{C}^7)\sim SL(7,\mathbb{C})/Aff(4,3,\mathbb{C})$, where $Aff(4,3,\mathbb{C})$ is realized by the matrices

$$\mathrm{Aff}(4,3,\mathbb{C})\sim\begin{bmatrix} G_{11} & 0 \\ G_{21} & G_{22}\end{bmatrix},$$
$$G_{11}\in\mathbb{C}^{4\times 4},\quad G_{22}\in\mathbb{C}^{3\times 3},\quad G_{21}\in\mathbb{C}^{3\times 4},$$
$$\det G_{11}\det G_{22}=1. \tag{4.1}$$

Introducing the diagonal $O(7,\mathbb{C})$ metric I_7, and restricting from $SL(7,\mathbb{C})$ to $O(7,\mathbb{C})$, we see that $Aff(4,3,\mathbb{C})$ restricts to $O(4,\mathbb{C})\otimes O(3,\mathbb{C})$. The dimensions of the corresponding spaces satisfy

$$\begin{aligned}&\dim SL(7,\mathbb{C})/\mathrm{Aff}(4,3,\mathbb{C})=48-36=12,\\ &\dim O(7,\mathbb{C})/[O(4,\mathbb{C})\times O(3,\mathbb{C})]=21-9=12.\end{aligned} \tag{4.2}$$

Having in mind that we shall below wish to restrict to the $G_2(\mathbb{C})$ group, we define homogeneous coordinates on the Grassmannian of nondegenerate three-planes as the matrix elements of the matrices

$$\xi=\begin{pmatrix} X\\ \mathbf{Z}^T\\ Y\end{pmatrix},\quad X,Y\in\mathbb{C}^{3\times 3},\quad \mathbf{Z}^T\in\mathbb{C}^{1\times 3}. \tag{4.3}$$

We choose the origin to be $(0,0,I_3)^T$; the fact that $Aff(4,3,\mathbb{C})$ and $O(4,\mathbb{C})\otimes O(3,\mathbb{C})$ are the isotropy groups of the origin within $SL(7,\mathbb{C})$ and $O(7,\mathbb{C})$, respectively, is then manifest.

The corresponding $O(7,\mathbb{C})$ equations can be written in homogeneous coordinates as

$$\begin{pmatrix}\dot X\\ \dot{\mathbf{Z}}^T\\ \dot Y\end{pmatrix}=\begin{pmatrix} R & \mathbf{m} & V\\ -\mathbf{m}^T & 0 & \mathbf{n}^T\\ -V^T & -\mathbf{n} & U\end{pmatrix}\begin{pmatrix} X\\ \mathbf{Z}^T\\ Y\end{pmatrix},$$
$$R,U,V\in\mathbb{C}^{3\times 3},\quad \mathbf{m},\mathbf{n}\in\mathbb{C}^{3\times 1},$$
$$R^T+R=0,\quad U^T+U=0. \tag{4.4}$$

The homogeneous coordinates satisfy the $O(7,\mathbb{C})$ condition

$$X^TX+ZZ^T+Y^TY=I. \tag{4.5}$$

Introducing affine coordinates in the usual manner

$$W_1=XY^{-1},\quad \mathbf{W}_2^T=\mathbf{Z}^TY^{-1},\quad \det Y\neq 0, \tag{4.6}$$

we obtain a system of 12 nonlinear ODE's associated to the action of $O(7,\mathbb{C})$ on $O(7,\mathbb{C})/[O(4,\mathbb{C})\times O(3,\mathbb{C})]$:

$$\dot W_1=V+RW_1-W_1U+\mathbf{m}\mathbf{W}_2^T+W_1V^TW_1+W_1\mathbf{n}\mathbf{W}_2^T,$$
$$\dot{\mathbf{W}}_2^T=\mathbf{n}^T-\mathbf{m}^TW_1-\mathbf{W}_2^TU+\mathbf{W}_2^TV^TW_1+\mathbf{W}_2^T\mathbf{n}\mathbf{W}_2^T. \tag{4.7}$$

Notice that (4.5) does not imply any restrictions on the matrices of affine coordinates W_1 and W_2. Notice also that (4.7) could have been written in the form of one rectangular matrix Riccati equation[6] for the matrix $W\in\mathbb{C}^{4\times 3}$, $W^T=(W_1^T,W_2)$, however, (4.7) is more convenient in the G_2 context.

Consider now the homogeneous space $G_2(\mathbb{C})/[SL(2,\mathbb{C})\otimes SL(2,\mathbb{C})]$. We have

$$\dim G_2(\mathbb{C})/[SL(2,\mathbb{C})\otimes SL(2,\mathbb{C})]=14-6=8. \tag{4.8}$$

Thus, four supplementary conditions must be imposed on the components of W_1 and W_2 in (4.6). These must be consequences of the specific properties of the group $G_2(\mathbb{C})$, i.e., the invariance of the tensor T of (I.2.12). To see this, let us apply an element $G=\{g_{ij}\}\in G_2(\mathbb{C})$ to the origin. We have

$$\begin{bmatrix} X\\ \mathbf{Z}^T\\ Y\end{bmatrix}=G\begin{bmatrix} 0\\ 0\\ I\end{bmatrix}=\begin{bmatrix} g_{15} & g_{16} & g_{17}\\ g_{25} & g_{26} & g_{27}\\ g_{35} & g_{36} & g_{37}\\ g_{45} & g_{46} & g_{47}\\ g_{55} & g_{56} & g_{57}\\ g_{65} & g_{66} & g_{67}\\ g_{75} & g_{76} & g_{77}\end{bmatrix}. \tag{4.9}$$

We are using the $O(7,\mathbb{C})$ metric given by the identity matric I_7, hence the nonzero components of the tensor T are given by (I.2.12). The invariance conditions (I.2.11) relate the first column of (4.9) to the other two:

$$g_{i5}T_{576}=g_{i5}=g_{m7}\,g_{n6}\,T_{imn}. \tag{4.10}$$

The $O(7,\mathbb{C})$ conditions (4.5) can be rewritten as

$$g_{i6}\,g_{i6}=g_{i7}\,g_{i7}=1,\quad g_{i6}\,g_{i7}=0. \tag{4.11}$$

Thus, only $21-7-3=11$ of the 21 homogeneous coordinates are independent. To reduce further, namely to eight truly independent quantities, we must, as usual, go over to affine coordinates. In doing so, we automatically account for the equivalence $(X^T,Z,Y^T)\sim(G_0^TX^T,G_0^TZ,G_0^TY^T)$, where in the considered case we have $G_0=O(3,\mathbb{C})$. This will effectively remove three redundant coordinates, or provide three needed constraints on the 11 quantities that we have so far reduced to.

Using (4.6), we express

$$X=W_1Y,\quad Z^T=W_2^TY,$$
$$W_1\equiv\{w_{ik}\},\quad W_2^T\equiv\{v_1,v_2,v_3\},\quad i,k=1,2,3. \tag{4.12}$$

Using (4.9), we express X, Z^T, and Y in terms of the elements g_{ab} of a $G_2(\mathbb{C})$ group element. We then eliminate g_{a5} $(a=1,\ldots,7)$ using (4.10). Defining the minors S_{55}, S_{65}, and S_{75} as

$$S_{55}=g_{66}g_{77}-g_{67}g_{76},\quad S_{65}=g_{57}g_{76}-g_{56}g_{77},$$
$$S_{75}=g_{56}g_{67}-g_{57}g_{66}, \tag{4.13}$$

we see that (4.9), (4.10), and (4.12) provide a system of

four linear homogeneous equations for S_{55}, S_{65}, and S_{75}. Since this system must have a nonzero solution, the rank of the matrix of the system must be 2. This in turn requires that the determinants of four 3×3 matrices vanish. The matrix elements of these matrices are themselves third-order polynomials in the components w_{ik} and v_i of W_1 and W_2. Thus we obtain four nonlinear constraints

$$\Delta_\mu(w_{ik},v_i)=0, \quad \mu=1,...,4, \tag{4.14}$$

on the 12 components of the matrices W_1 and W_2, reducing the number of independent components to precisely 8, as required. We give the determinants Δ_μ in the Appendix.

The $G_2(\mathbb{C})/[SL(2,\mathbb{C})\otimes SL(2,\mathbb{C})]$ ODE's with superposition formulas are thus the matrix Riccati equations (4.7), subject to the following conditions.

(i) The coefficients R, V, U, $\mathbf{m}$, and $\mathbf{n}$ are such that the matrix M of (3.10) is an element of $g_2(\mathbb{C})$ for all times t [i.e., they satisfy (3.10) and (3.11)].

(ii) The matrix elements of W_1 and W_2 satisfy the constraints (4.14). These can be imposed at the initial time $t=t_0$ and they will then be satisfied for all times t [as a consequence of the above condition (i)].

We have not attempted to solve the constraints explicitly: this would involve solving cubic equations and would lead to very complicated explicit formulas. The nonlinearities would, in general be irrational, involving square and cubic roots of the dependent variables.

B. The $G_2^C(\mathbb{R})/[SU(2)\circledast SU(2)]$ and $G_2^{NC}(\mathbb{R})/[SU(2)\circledast SU(2)]$ equations

The nonlinear ODE's in these two cases are intimately related to those obtained in Sec. IV A.

Consider first the compact case $G_2^C(\mathbb{R})/[SU(2)\otimes SU(2)]$. The metric is again given by the identity matrix I_7 (see Table III of I) and the tensor T is exactly the same as in the complex case. The nonlinear ODE's associated to the action of $G_2^C(\mathbb{R})$ on $G_2^C(\mathbb{R})/[SU(2)\otimes SU(2)]$ are hence the same equations (4.7) with the same g_2 constraints on the coefficients R, V, U, $\mathbf{m}$, and $\mathbf{n}$, as above, and the some nonlinear constraints (4.14) on the matrices $W_1(t)$ and $W_2(t)$. The only difference is that both the coefficients in the equations, and the matrices of dependent variables are restricted to be real.

The noncompact case $G_2^{NC}(\mathbb{R})/[SU(2)\otimes SU(2)]$ is only slightly different. The appropriate metric is given by the matrix $I_{4,3}$ (see Table IV of I). We have

$$I_{4,3}=H^T I_7 H, \quad H=\begin{pmatrix} I_4 & 0 \\ 0 & iI_3 \end{pmatrix}. \tag{4.15}$$

For the elements of Lie group $G_2(\mathbb{C})$, Lie algebra $g_2(\mathbb{C})$ and the tensor T, we then have

$$g'=H^{-1}gH, \quad M'=H^{-1}MH, \tag{4.16}$$

$$T'_{imn}=(H^{-1})_{ia}T_{abc}H_{bm}H_{cn}. \tag{4.17}$$

The conditions (4.10) are not affected by this change, hence the constraints (4.14) remain the same as in the complex and compact cases.

The equations themselves are modified in that the matrix M of (4.4) is replaced by (3.12), i.e.,

$$\begin{aligned} \dot W_1 &= V+RW_1-W_1U+\mathbf{m}\mathbf{W}_2^T-W_1V^TW_1-W_1\mathbf{n}\mathbf{W}_2^T, \\ \dot{\mathbf{W}}_2^T &= \mathbf{n}^T-\mathbf{m}W_1-\mathbf{W}_2^TU-\mathbf{W}_2^TV^TW_1-\mathbf{W}_2^T\mathbf{n}\mathbf{W}_2^T, \end{aligned} \tag{4.18}$$

where all entries are real.

C. The $G_2^{NC}(\mathbb{R})/[SU(1,1)\circledast SU(1,1)]$ equations

We shall again make use of the metric given by the matrix $I_{4,3}$. The subgroup $SU(1,1)\otimes SU(1,1)$ was identified in Ref. 1 as the maximal subgroup of $G_2^{NC}(\mathbb{R})$ leaving a nondegenerate three-dimensional subspace with signature $(+\,+\,-)$ invariant. In keeping with this fact, and in analogy with our procedure in the complex case, we choose the homogeneous coordinates of the origin in $G_2^{NC}(\mathbb{R})/[SU(1,1)\otimes SU(1,1)]$ to be

$$U_0=\begin{bmatrix} 0&0&0\\ 0&0&0\\ 1&0&0\\ 0&1&0\\ 0&0&0\\ 0&0&0\\ 0&0&1 \end{bmatrix}. \tag{4.19}$$

Correspondingly, the homogeneous coordinates of an arbitrary point in this space are

$$U'=GU_0=\begin{pmatrix} g_{13}&g_{14}&g_{17}\\ g_{23}&g_{24}&g_{27}\\ \hline g_{33}&g_{34}&g_{37}\\ g_{43}&g_{44}&g_{47}\\ \hline g_{53}&g_{54}&g_{57}\\ g_{63}&g_{64}&g_{67}\\ \hline g_{73}&g_{74}&g_{77} \end{pmatrix}=\begin{pmatrix} X_1\\ \hline Y_1\\ \hline X_2\\ \hline Y_2^T \end{pmatrix}$$

$$X_1,Y_1,X_2\in\mathbb{R}^{2\times 3}, \quad Y_2^T\in\mathbb{R}^{1\times 3}, \quad G\in G_2^{NC}(\mathbb{R}). \tag{4.20}$$

The invariance of the tensor T' under the action of $G_2^{NC}(\mathbb{R})$ allows us to express the first column in terms of the other two:

$$g_{i3}=-iT'_{ijk}\,g_{j7}\,g_{k4}, \tag{4.21}$$

in analogy to (4.10). In order to reduce to the required number of real coordinates, namely eight, we again introduce affine coordinates

$$W_1 = X_1 Y^{-1}, \quad W_2 = X_2 Y^{-1}, \quad Y \equiv \begin{pmatrix} Y_1 \\ Y_2^T \end{pmatrix}, \quad \det Y \neq 0 \tag{4.22}$$

(notice that in this formalism we have $W_1, W_2 \in \mathbb{R}^{2\times 3}$).

Proceeding as in the complex case we can again obtain, from (4.20)–(4.22), four nonlinear constraints on 12 components of W_1 and W_2. In view of their length we do not reproduce them here (they are available from the authors upon request).

The $G_2^N(\mathbb{R})/[SU(1,1) \otimes SU(1,1)]$ equations in homogeneous coordinates are

$$\begin{pmatrix} \dot{X}_1 \\ \dot{Y}_1 \\ \dot{X}_2 \\ \dot{Y}_2^T \end{pmatrix} = \begin{pmatrix} A & B & C & p \\ -B^T & D & E & q \\ C^T & E^T & F & r \\ p^T & q^T & -r^T & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ Y_1 \\ X_2 \\ Y_2^T \end{pmatrix}, \tag{4.23}$$

where, in agreement with (3.12), we have

$$A = \frac{1}{2}\begin{pmatrix} 0 & a_3 + m_3 \\ -a_3 - m_3 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & m_3 \\ -m_3 & 0 \end{pmatrix}, \quad F = \frac{1}{2}\begin{pmatrix} 0 & a_3 - m_3 \\ -a_3 + m_3 & 0 \end{pmatrix},$$

$$B = \begin{pmatrix} (-a_2 - m_2)/2 & m_1 \\ (a_1 + m_1)/2 & m_2 \end{pmatrix}, \quad C = \frac{1}{2}\begin{pmatrix} v_{11} & v_{12} + n_3 \\ v_{12} - n_3 & v_{22} \end{pmatrix}, \quad E = \begin{pmatrix} (v_{12} + n_2)/2 & (v_{23} - n_2)/2 \\ n_1 & n_2 \end{pmatrix}, \tag{4.24}$$

$$p = \frac{1}{2}\begin{pmatrix} v_{13} - n_2 \\ v_{23} + n_1 \end{pmatrix}, \quad q = \begin{pmatrix} (-v_{11} - v_{22})/2 \\ n_3 \end{pmatrix}, \quad r = \frac{1}{2}\begin{pmatrix} -a_2 + m_2 \\ a_1 - m_1 \end{pmatrix}.$$

Rewriting (4.23) in the affine coordinates (4.22) we obtain

$$\begin{aligned} \dot{W}_1 &= (B,p) + AW_1 - W_1\begin{pmatrix} D & q \\ q^T & 0 \end{pmatrix} + CW_2 - W_1\begin{pmatrix} -B^T \\ p^T \end{pmatrix} W_1 - W_1\begin{pmatrix} E \\ -r^T \end{pmatrix} W_2, \\ \dot{W}_2 &= (E^T, r) + C^T W_1 + FW_2 - W_2\begin{pmatrix} D & q \\ q^T & 0 \end{pmatrix} - W_2\begin{pmatrix} -B^T \\ p^T \end{pmatrix} W_1 - W_2\begin{pmatrix} E \\ -r^T \end{pmatrix} W_2. \end{aligned} \tag{4.25}$$

Thus (4.25), together with four constraints of the type (4.14) provide the nonlinear ODE's corresponding to the action of $G_2^{NC}(\mathbb{R})$ on the space $G_2^{NC}(\mathbb{R})/[SU(1,1) \otimes SU(1,1)]$.

V. CONCLUSIONS

We have shown that nonlinear ordinary differential equations with superposition formulas can be associated with the exceptional Lie group G_2 in a manner quite similar to that used for the classical Lie groups. Since G_2 is simple and since we have only used homogeneous spaces G_2/H, where H is a maximal subgroup of G_2, the obtained systems of equations are all indecomposable.[6]

In all cases we have made use of the imbedding of a seven-dimensional representation of $G_2(\mathbb{C})$, $G_2^C(\mathbb{R})$, or $G_2^{NC}(\mathbb{R})$ into $O(7,\mathbb{C})$, $O(7)$, or $O(4,3)$, respectively. The ODE's for some subgroups H turned out to be special cases of $O(7,\mathbb{C})$, $O(7)$, or $O(4,3)$ equations. For other subgroups new features appeared, due to the existence of an invariant antisymmetric tensor T.

A mathematical by-product of our analysis is the construction of quite a few models and coordinate systems for various homogeneous spaces for the complex and real forms of G_2. These can also be used for other purposes than those of the present article.

The emphasis in this article has been on deriving the ODE's themselves. In a forthcoming article we shall present the superposition formulas. This is of interest for both equations specific to the G_2 group, and for those that are restrictions of $O(7)$ type equations. In the latter case the superposition formulas for the G_2 equations are more efficient than for the $O(7)$ ones, in that they make use of a smaller number of particular solutions to express the general one.

ACKNOWLEDGMENTS

Financial support from the "Accords culturels Belgique-Québec" making possible mutual visits was greatly appreciated. The research of one of the authors (P.W.) is partially sponsored by the Natural Sciences and Engineering Research Council of Canada and the "Fonds FCAR du Gouvernement du Québec."

APPENDIX: THE CONSTRAINTS FOR THE $G_2(\mathbb{C})/[SL(2,\mathbb{C}) \otimes SL(2,\mathbb{C})]$, $G_2^C(\mathbb{R})/[SU(2) \otimes SU(2)]$, AND $G_2^{NC}/[SU(2) \otimes SU(2)]$ EQUATIONS

The explicit form of the constraints (4.14) imposed on the affine coordinates $W_1 = \{w_{ik}\}$, $W_2^T = \{v_1, v_2, v_3\}$ are obtained by requiring that the determinants of all four 3×3 submatrices of the following 4×3 matrix should vanish:

$$
\begin{bmatrix}
\begin{aligned}
&w_{11}+w_{22}+w_{33}+w_{11}(w_{23}w_{32}-w_{22}w_{33})\\
&\quad+w_{12}(w_{12}w_{33}-w_{13}w_{32})\\
&\quad+w_{13}(w_{13}w_{22}-w_{12}w_{23})\\
&\quad+(w_{11}w_{12}+w_{12}w_{22}+w_{13}w_{32})v_3\\
&\quad-(w_{11}w_{13}+w_{12}w_{23}+w_{13}w_{33})v_2
\end{aligned}
&
\begin{aligned}
&w_{12}-w_{21}-v_3+w_{11}(w_{21}w_{33}-w_{23}w_{31})\\
&\quad+w_{12}(w_{13}w_{31}-w_{11}w_{33})\\
&\quad+w_{13}(w_{11}w_{23}-w_{13}w_{21})\\
&\quad+(w_{11}w_{13}+w_{12}w_{23}+w_{13}w_{33})v_1\\
&\quad-(w_{11}^2+w_{12}w_{21}+w_{13}w_{31})v_3
\end{aligned}
&
\begin{aligned}
&w_{13}-w_{31}+v_2+w_{11}(w_{22}w_{31}-w_{21}w_{32})\\
&\quad+w_{12}(w_{11}w_{32}-w_{12}w_{31})\\
&\quad+w_{13}(w_{12}w_{21}-w_{11}w_{22})\\
&\quad+(w_{11}^2+w_{12}w_{21}+w_{13}w_{31})v_2\\
&\quad-(w_{11}w_{12}+w_{12}w_{22}+w_{13}w_{32})v_1
\end{aligned}
\\
\begin{aligned}
&w_{21}-w_{12}+v_3+w_{21}(w_{23}w_{32}-w_{22}w_{33})\\
&\quad+w_{22}(w_{12}w_{33}-w_{13}w_{32})\\
&\quad+w_{23}(w_{13}w_{22}-w_{12}w_{23})\\
&\quad+(w_{21}w_{12}+w_{22}^2+w_{23}w_{32})v_3\\
&\quad-(w_{21}w_{13}+w_{22}w_{23}+w_{23}w_{33})v_2
\end{aligned}
&
\begin{aligned}
&w_{11}+w_{22}+w_{33}+w_{21}(w_{21}w_{33}-w_{23}w_{31})\\
&\quad+w_{22}(w_{13}w_{31}-w_{11}w_{33})\\
&\quad+w_{23}(w_{11}w_{23}-w_{13}w_{21})\\
&\quad+(w_{21}w_{13}+w_{22}w_{23}+w_{23}w_{33})v_1\\
&\quad-(w_{21}w_{11}+w_{22}w_{21}+w_{23}w_{31})v_3
\end{aligned}
&
\begin{aligned}
&w_{23}-w_{32}-v_1+w_{21}(w_{22}w_{31}-w_{21}w_{32})\\
&\quad+w_{22}(w_{11}w_{32}-w_{12}w_{31})\\
&\quad+w_{23}(w_{12}w_{21}-w_{11}w_{22})\\
&\quad+(w_{21}w_{11}+w_{22}w_{21}+w_{23}w_{31})v_2\\
&\quad-(w_{21}w_{12}+w_{22}^2+w_{23}w_{32})v_1
\end{aligned}
\\
\begin{aligned}
&w_{31}-w_{13}-v_2+w_{31}(w_{23}w_{32}-w_{22}w_{33})\\
&\quad+w_{32}(w_{12}w_{33}-w_{13}w_{32})\\
&\quad+w_{33}(w_{13}w_{22}-w_{12}w_{23})\\
&\quad+(w_{31}w_{12}+w_{32}w_{22}+w_{33}w_{32})v_3\\
&\quad-(w_{31}w_{13}+w_{32}w_{23}+w_{33}^2)v_2
\end{aligned}
&
\begin{aligned}
&w_{32}-w_{23}+v_1+w_{31}(w_{21}w_{33}-w_{23}w_{31})\\
&\quad+w_{32}(w_{13}w_{31}-w_{11}w_{33})\\
&\quad+w_{33}(w_{11}w_{23}-w_{13}w_{21})\\
&\quad+(w_{31}w_{13}+w_{32}w_{23}+w_{33}^2)v_1\\
&\quad-(w_{31}w_{11}+w_{32}w_{21}+w_{33}w_{31})v_3
\end{aligned}
&
\begin{aligned}
&w_{11}+w_{22}+w_{33}+w_{31}(w_{22}w_{31}-w_{21}w_{32})\\
&\quad+w_{32}(w_{11}w_{32}-w_{12}w_{31})\\
&\quad+w_{33}(w_{12}w_{21}-w_{11}w_{22})\\
&\quad+(w_{31}w_{11}+w_{32}w_{21}+w_{33}w_{31})v_2\\
&\quad-(w_{31}w_{12}+w_{32}w_{22}+w_{33}w_{32})v_1
\end{aligned}
\\
\begin{aligned}
&v_1-w_{23}+w_{32}+(w_{23}w_{32}-w_{22}w_{33})v_1\\
&\quad+(w_{12}w_{33}-w_{13}w_{32})v_2\\
&\quad+(w_{13}w_{22}-w_{12}w_{23})v_3\\
&\quad+(v_1w_{12}+v_2w_{22}+v_3w_{32})v_3\\
&\quad-(v_1w_{13}+v_2w_{23}+v_3w_{33})v_2
\end{aligned}
&
\begin{aligned}
&v_2+w_{13}-w_{31}+(w_{21}w_{33}-w_{31}w_{23})v_1\\
&\quad+(w_{13}w_{31}-w_{11}w_{33})v_2\\
&\quad+(w_{11}w_{23}-w_{13}w_{21})v_3\\
&\quad+(v_1w_{13}+v_2w_{23}+v_3w_{33})v_1\\
&\quad-(v_1w_{11}+v_2w_{21}+v_3w_{31})v_3
\end{aligned}
&
\begin{aligned}
&v_3+w_{21}-w_{12}+(w_{22}w_{31}-w_{21}w_{32})v_1\\
&\quad+(w_{11}w_{32}-w_{12}w_{31})v_2\\
&\quad+(w_{12}w_{21}-w_{11}w_{22})v_3\\
&\quad+(v_1w_{11}+v_2w_{21}+v_3w_{31})v_2\\
&\quad-(v_1w_{12}+v_2w_{22}+v_3w_{32})v_1
\end{aligned}
\end{bmatrix}
$$

[1]J. Beckers, V. Hussin, and P. Winternitz, J. Math. Phys. **27**, 2217 (1986).
[2]S.Lie and G. Scheffers, *Vorlesungen über continuierlichen Gruppen mit geometrischen und anderen Anwendungen* (Teubner, Leipzig, 1893) (reprinted by Chelsea, New York, 1967).
[3]R. L. Anderson, Lett. Math. Phys. **4**, 1 (1980).
[4]R. L. Anderson, J. Harnad, and P. Winternitz, Lett. Math. Phys. **5**, 143 (1981); Physica D **4**, 164 (1982).
[5]R. L. Anderson, J. Harnad, and P. Winternitz, J. Math. Phys. **24**, 1062 (1982).
[6]S. Shnider and P. Winternitz, Lett. Math. Phys. **8**, 69 (1984); J. Math. Phys. **25**, 3155 (1984).
[7]M. del Olmo, M. A. Rodriguez, and P. Winternitz, J. Math. Phys. **27**, 14 (1986).
[8]J. Beckers, V. Hussin, and P. Winternitz, Lett. Math. Phys. **11**, 81 (1986).
[9]M. Sorine and P. Winternitz, IEEE Trans. Autom. Control **AC-30**, 266 (1985).
[10]D. Rand and P. Winternitz, Comp. Phys. Comm. **33**, 305 (1984).
[11]A. T. Ogielski, M. K. Prasad, A. Sinha, and L. L. Chau-Wang, Phys. Lett. B **91**, 387 (1980).
[12]J. Harnad, Y. Saint-Aubin, and S. Shnider, Commun. Math. Phys. **92**, 329 (1984); **93**, 33 (1984).
[13]W. T. Reid, *Riccati Differential Equations* (Academic, New York, 1972).
[14]J. Beckers, J. Harnad, M. Perroud, and P. Winternitz, J. Math. Phys. **19**, 2126 (1978).

Superposition formulas for rectangular matrix Riccati equations

M. A. del Olmo,[a] M. A. Rodríguez,[b] and P. Winternitz
Centre de Recherches Mathématiques, Université de Montréal, C.P. 6128, Succ. A, Montréal, Québec, Canada H3C 3J7

(Received 21 November 1985; accepted for publication 22 October 1986)

A system of nonlinear ordinary differential equations allowing a superposition formula can be associated with every Lie group–subgroup pair $G \supset G_0$. We consider the case when $G = \mathrm{SL}(n+k,\mathbb{C})$ and $G_0 = P(k)$ is a maximal parabolic subgroup of G, leaving a k-dimensional vector space invariant ($1 \leqslant k \leqslant n$). The nonlinear ordinary differential equations (ODE's) in this case are rectangular matrix Riccati equations for a matrix $W(t) \in \mathbb{C}^{n \times k}$. The special case $n = rk$ ($n,r,k \in \mathbb{N}$) is considered and a superposition formula is obtained, expressing the general solution in terms of $r+3$ particular solutions for $r \geqslant 2$, $k \geqslant 2$. For $r = 1$ (square matrix Riccati equations) five solutions are needed, for $r = n$ (projective Riccati equations) the required number is $n+2$.

I. INTRODUCTION

Superposition formulas for ordinary differential equations (ODE's) are based on the following theorem, due to Lie and Scheffers.[1]

The general solution $\mathbf{x}(t)$ of the system of equations

$$\frac{dx^\mu}{dt} \equiv \dot{x}^\mu = f^\mu(\mathbf{x},t), \quad \mu = 1,\ldots,n, \tag{1.1}$$

can be expressed as a function of m particular solutions and n significant constants

$$\mathbf{x}(t) = S(\mathbf{x}_1(t),\ldots,\mathbf{x}_m(t),c_1,\ldots,c_n) \tag{1.2}$$

if and only if the right-hand side of (1.1) has the form

$$f^\mu(\mathbf{x},t) = \sum_{j=1}^{r} Z_j(t)\xi_j^\mu(\mathbf{x}), \tag{1.3}$$

and the differential operators

$$X_j \equiv \sum_{\mu=1}^{n} \xi_j^\mu(\mathbf{x}) \frac{\partial}{\partial x^\mu}, \quad j = 1,\ldots,r \tag{1.4}$$

generate a Lie algebra of finite dimensions r under commutation.

We shall call expression (1.2) a "superposition formula" and the solutions $\mathbf{x}_1(t),\ldots,\mathbf{x}_m(t)$ a "fundamental set of solutions." Their number m and the independence conditions which they must satisfy have to be established in each specific case.

Given an arbitrary Lie group G and a Lie subgroup $G_0 \subset G$, we can always, at least in principle, construct the homogeneous space $M \sim G/G_0$. The infinitesimal action of G on M (in some coordinates) will give us the vector fields X_j of (1.4) and from these we can read off the ODE's (1.1). It has recently been shown[2] that if the Lie algebras L and L_0, corresponding to the Lie groups G and G_0, respectively, form a transitive primitive Lie algebra[3,4] then the corresponding system of ODE's with a superposition formula will be indecomposable. We recall here that "indecomposable" in this context means that it is not possible to decouple a proper subsystem of equations from (1.1) that will have its own superposition formula.

Restricting ourselves to the indecomposable case, when (L,L_0) do form a transitive primitive Lie algebra, we note that the following possibilities occur[2].

(1) L is not semisimple. The ODE's in this case are linear (in general inhomogeneous).

(2) L is semisimple, but not simple. The equations are, *a priori*, linear, but the dependent variables are subject to nonlinear constraints.

(3) L is simple and L_0 is a maximal reductive subalgebra. Examples of such ODE's have recently been studied,[5] corresponding to the pairs $\mathrm{sl}(n,\mathbb{C}) \supset \mathrm{o}(n,\mathbb{C})$ and $\mathrm{sl}(2n,\mathbb{C}) \supset \mathrm{sp}(2n,\mathbb{C})$. The equations have rational but nonpolynomial nonlinearities.

(4) L is simple and L_0 is a maximal parabolic subalgebra.[6–9] The equations in this case have polynomial nonlinearities.

This paper is devoted to the last of the above cases. In particular, the Lie algebra L of Lie's theorem is chosen to be $\mathrm{sl}(N,\mathbb{C})$ and the subalgebra L_0 of vector fields vanishing at the origin is a maximal parabolic subalgebra $p(k)$ of $\mathrm{sl}(N,\mathbb{C})$. The corresponding maximal parabolic subgroup $P(k) \subset \mathrm{SL}(N,\mathbb{C})$ leaves a k-dimensional vector space invariant ($1 \leqslant k \leqslant [N/2]$).

It has already been shown that this case leads to interesting ODE's, namely rectangular matrix Riccati equations (MRE's) for a $n \times k$-dimensional real or complex matrix $W(t)$.

MRE's occur in many applications; e.g., as Bäcklund transformations in the study of integrable systems, as special cases of Volterra–Lotke equations in population dynamics, in optimal control theory and elsewhere.[10]

Explicit superposition formulas have so far been obtained in two special cases only. The first is the case $k = 1$, when the matrix reduces to a single column. The equations were called projective Riccati equations[6,7] and the superposition formula involves $n+2$ particular solutions. The other case is $k = n$, i.e., square MRE's. The superposition formula requires just five particular solutions (for any $n \geqslant 2$).[8]

It is convenient to relate the dimensions n and k of W by the formula

$$n = rk + l, \quad 0 \leqslant l \leqslant k-1, \quad n,r,k,l \in \mathbb{N}. \tag{1.5}$$

[a] On leave of absence from Departamento de Física Teórica, Facultad de Ciencias, Universidad de Valladolid, Valladolid, Spain.

[b] Present address: Departamento de Métodos Matemáticos de la Física, Facultad de Físicas, Universidad Complutense, Madrid, Spain.

 0022-2488/87/030530-06$02.50

In this paper we concentrate on the simplest case, when $l = 0$ in (1.5), i.e., $n = rk$. We take $r \geqslant 2$, $k \geqslant 2$ since $r = 1$ corresponds to square MRE's and $k = 1$ to projective Riccati equations, both of which have already been treated.[6–8] The case $l \neq 0$ is more complicated and has so far not been treated.

The general form of the MRE and of its solution is presented in Sec. II. The properties of a fundamental set of solutions are established in Sec. III, where we also present a "standard" form of the initial conditions. In Sec. IV we obtain the superposition formula and also show how r particular solutions can be used to linearize the MRE and to partly decouple it.

II. THE RECTANGULAR MATRIX RICCATI EQUATIONS

By definition, a maximal parabolic subalgebra p of a complex simple Lie algebra L is a subalgebra $p \subset L$ that is maximally contained in L and contains the Borel subalgebra.[11] A maximal parabolic subalgebra $p(k)$ of $\mathrm{sl}(n+k,\mathbb{C})$ can be characterized by the fact that it is the largest subalgebra that leaves a k-dimensional subspace of $\mathbb{C}^N$ invariant $(N = n + k,\ 1 \leqslant k \leqslant N-1)$.[12]

We have

$$\mathrm{SL}(n+k,\mathbb{C})/P(k) \approx G_k(\mathbb{C}^{n+k}), \tag{2.1}$$

where $G_k(\mathbb{C}^{n+k})$ is the Grassmannian[13] of k planes in $\mathbb{C}^{n+k}$. Homogeneous coordinates on this space are given by the matrix elements of the matrix

$$\begin{pmatrix} X \\ Y \end{pmatrix}, \quad X \in \mathbb{C}^{n\times k}, \quad Y \in \mathbb{C}^{k\times k}, \quad \mathrm{rank}\begin{pmatrix} X \\ Y \end{pmatrix} = k. \tag{2.2}$$

Two matrices of the form (2.2) describe the same point if they satisfy

$$\begin{pmatrix} \tilde{X} \\ \tilde{Y} \end{pmatrix} = \begin{pmatrix} XG \\ YG \end{pmatrix}, \quad G \in \mathrm{GL}(k,\mathbb{C}), \tag{2.3}$$

for some nonsingular matrix G. In these (redundant) coordinates the action of $\mathrm{SL}(n+k,\mathbb{C})$ is linear and the associated ODE's are

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \end{pmatrix} = \begin{pmatrix} C & A \\ -D & -B \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}. \tag{2.4}$$

For points satisfying $\det Y \neq 0$ we can introduce affine coordinates on $G_k(\mathbb{C}^{n+k})$, thus removing the redundancy (2.3),

$$W = XY^{-1}. \tag{2.5}$$

The action of $\mathrm{SL}(n+k,\mathbb{C})$ in affine coordinates is a matrix fractional linear one; the vector fields representing the infinitesimal action are

$$\begin{aligned} &\hat{A}_{\mu\nu} = \frac{\partial}{\partial w_{\mu\nu}}, \quad \hat{B}_{\nu\nu'} = \sum_{\mu=1}^{n} w_{\mu\nu}\frac{\partial}{\partial w_{\mu\nu'}}, \\ &\hat{C}_{\mu\mu'} = \sum_{\nu=1}^{k} w_{\mu'\nu}\frac{\partial}{\partial w_{\mu\nu}}, \\ &\hat{D}_{\nu\mu} = \sum_{\alpha=1}^{n}\sum_{\beta=1}^{k} w_{\alpha\nu}w_{\mu\beta}\frac{\partial}{\partial w_{\alpha\beta}}, \\ &1 \leqslant \mu,\mu' \leqslant n, \quad 1 \leqslant \nu,\nu' \leqslant k, \end{aligned} \tag{2.6}$$

where $w_{\mu\nu}$ are the matrix elements of W.

The corresponding system of ODE's are the rectangular MRE's mentioned in the Introduction[8]:

$$\begin{aligned} &\dot{W} = A + WB + CW + WDW, \\ &W,A \in \mathbb{C}^{n\times k}, \quad B \in \mathbb{C}^{k\times k}, \quad C \in \mathbb{C}^{n\times n}, \quad D \in \mathbb{C}^{k\times n}, \end{aligned} \tag{2.7}$$

where $A,\ldots,D$ are given matrix functions of time t.

The right-hand side of (2.7) represents a curve in the Lie algebra $\mathrm{sl}(n+k,\mathbb{C})$. The general solution of (2.7) is given by the corresponding action of $\mathrm{SL}(n+k,\mathbb{C})$,

$$W(t) = [G_{11}(t)U + G_{12}(t)][G_{21}(t)U + G_{22}(t)]^{-1}, \tag{2.8}$$

where $U \in \mathbb{C}^{n\times k}$ is a constant matrix, specifying the initial conditions for $W(t)$ and

$$G(t) = \begin{pmatrix} G_{11}(t) & G_{12}(t) \\ G_{21}(t) & G_{22}(t) \end{pmatrix} \tag{2.9}$$

is a curve in $\mathrm{SL}(n+k,\mathbb{C})$, to be determined in terms of a sufficient number of particular solutions $W_i(t)$ of (2.7).

With no loss of generality we can assume $n \geqslant k$, since the case $n < k$ can be reduced to the considered one by transposing the MRE. It is convenient to put $n = rk + l$ as in (1.5). Here we restrict ourselves to $l = 0$. Moreover, $r = 1$ corresponds to the square MRE, $k = 1$ to the projective Riccati equations. Both have been treated earlier.[6–8]

III. A FUNDAMENTAL SET OF SOLUTIONS

Let us assume that m solutions of the MRE (2.7) are known. They provide $m \cdot n \cdot k$ equations for the matrix elements of $G(t)$, when substituted into (2.8). This provides a lower limit on m, namely,

$$mnk \geqslant (n+k)^2 - 1, \tag{3.1}$$

since $(n+k)^2 - 1$ is the number of independent matrix elements in $G(t) \in \mathrm{SL}(n+k,\mathbb{C})$. According to the general theory,[7,8] a set of solutions $W_1(t),\ldots,W_m(t)$ of the considered MRE will suffice, at least locally, to determine $G(t)$ if any subgroup of $\mathrm{SL}(n+k,\mathbb{C})$ leaving the m initial values $W_i(t_0)$ on the product of m copies of the Grassmannian $G_k(\mathbb{C}^{n+k})$ invariant is contained in the center of $\mathrm{SL}(n+k,\mathbb{C})$. We shall construct such a fundamental set of solutions explicitly and then show that a generically chosen set of m solutions can be transformed into this "standard set."

Let us now restrict to the case (1.5) with $l = 0$, i.e., put

$$n = rk, \quad r \geqslant 2, \quad k \geqslant 2. \tag{3.2}$$

A point on $G_k(\mathbb{C}^{n+k})$ can be given as

$$\xi = \begin{pmatrix} X_1 \\ \vdots \\ X_r \\ Y \end{pmatrix} \quad \text{or} \quad W = \begin{pmatrix} W_1 \\ \vdots \\ W_r \end{pmatrix}, \quad X_i, Y, W_i \in \mathbb{C}^{k\times k}, \quad i = 1,\ldots,r, \tag{3.3}$$

in homogeneous or affine coordinates, respectively. Correspondingly, we shall write the elements of $G \in \mathrm{SL}(n+k,\mathbb{C})$ of (2.9) as

$$G_{11}=\begin{pmatrix} M_{11},\dots,M_{1r} \\ \vdots \\ M_{r1},\dots,M_{rr}\end{pmatrix},\quad G_{12}=\begin{pmatrix} N_1 \\ \vdots \\ N_r\end{pmatrix},$$
$$G_{21}=(P_1,\dots,P_r),\quad G_{22}=Q, \tag{3.4}$$
$$M_{ij},N_i,P_i,Q\in\mathbb{C}^{k\times k},\quad i,j=1,\dots,r.$$

Theorem 1: The following standard set of $r+3$ initial conditions for solutions of the MRE (2.7), given in homogeneous coordinates, has only the center of $SL(n+k,\mathbb{C})$ as its isotropy group:

$$\{\xi^S_1,\dots,\xi^S_{r+3}\}=\left\{\begin{pmatrix} I_k \\ 0 \\ \vdots \\ 0 \\ \hline 0\end{pmatrix},\ \begin{pmatrix} 0 \\ I_k \\ \vdots \\ 0 \\ \hline 0\end{pmatrix},\dots,\begin{pmatrix} 0 \\ 0 \\ \vdots \\ I_k \\ \hline 0\end{pmatrix},\right.$$
$$\left.\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \hline I_k\end{pmatrix},\ \begin{pmatrix} I_k \\ I_k \\ \vdots \\ I_k \\ \hline I_k\end{pmatrix},\ \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \\ \vdots \\ \Lambda_r \\ \hline I_k\end{pmatrix}\right\}. \tag{3.5}$$

The blocks $\Lambda_i\in\mathbb{C}^{k\times k}$ are such that one of them, say Λ_1, satisfies $\Lambda_1=\mathrm{diag}(\lambda_1,\lambda_2,\dots,\lambda_k)$, with $\lambda_i\in\mathbb{C}$, $\lambda_i\neq\lambda_j$ for $i\neq j$ and another one, say Λ_2, has no irreducible invariant subspaces in common with Λ_1.

Proof: A simple calculation shows that the conditions $G\xi^S_i\sim\xi^S_i$ for $i=1,\dots,r+1$, imply $M_{ij}=0$ for $i\neq j$, $N_i=0$, $P_i=0$ in (3.4). Further, imposing $G\xi^S_{r+2}\sim\xi^S_{r+2}$ we obtain

$$M_{11}=M_{22}=\cdots=M_{rr}=Q,\quad \det Q\neq 0.$$

The last condition $G\xi^S_{r+3}\sim\xi^S_{r+3}$ implies

$$Q\Lambda_iQ^{-1}=\Lambda_i,\quad i=1,\dots,r,$$

and in view of the conditions on Λ_1 and Λ_2 we find $Q=\lambda I$, $\lambda^{n+k}=1$. Q.E.D.

Notice that for $n=rk$ relation (3.1) yields

$$m\geqslant r+3-[1-(1/r)(1-1/k^2)],$$

so that the relation

$$m=r+3 \tag{3.6}$$

actually saturates (3.1).

Theorem 2: Given a set of $r+3$ initial conditions for solutions of the MRE (2.7) in affine coordinates $\{U_1,\dots,U_{r+3}\}\subset\mathbb{C}^{rk\times k}$ satisfying the conditions

$$\text{(i)}\quad \det\begin{pmatrix} U_1 & \cdots & U_r & U_{r+1} \\ I_k & \cdots & I_k & I_k\end{pmatrix}\neq 0; \tag{3.7}$$

$$\text{(ii)}\quad \det\begin{pmatrix} U_1 & \cdots & U_{j-1} & U_{j+1} & \cdots & U_{r+1} & U_{r+2} \\ I_k & \cdots & I_k & I_k & \cdots & I_k & I_k\end{pmatrix}\neq 0,\quad j=2,\dots,r; \tag{3.8}$$

$$\text{(iii)}\quad \det\begin{pmatrix} U_1 & \cdots & U_r & U_{r+3} \\ I_k & \cdots & I_k & I_k\end{pmatrix}\neq 0; \tag{3.9}$$

(iv) the matrices

$$T_i=S_iR_i(S_{r+1}R_{r+1})^{-1}\in\mathbb{C}^{k\times k},\quad i=1,2, \tag{3.10}$$

have no common nontrivial irreducible eigenspaces and one of them, say T_1, has k distinct eigenvalues, where S_i and R_i are defined by

$$U\equiv\begin{pmatrix} U_1 & \cdots & U_{r+1} \\ I_k & \cdots & I_k\end{pmatrix}, \tag{3.11}$$

$$\begin{pmatrix} (S_1)^{-1} \\ \vdots \\ (S_{r+1})^{-1}\end{pmatrix}\equiv U^{-1}\begin{pmatrix} U_{r+2} \\ I_k\end{pmatrix},\quad \begin{pmatrix} R_1 \\ \vdots \\ R_{r+1}\end{pmatrix}\equiv U^{-1}\begin{pmatrix} U_{r+3} \\ I_k\end{pmatrix}. \tag{3.12}$$

Then, there exists a transformation $G\in SL(n+k,\mathbb{C})$ transforming the set

$$\{\xi_i\}=\left\{\begin{pmatrix} U_i \\ I_k\end{pmatrix}\right\},\quad i=1,\dots,r+3, \tag{3.13}$$

into the standard set ξ^S_i of (3.5).

Proof: Put

$$G=\Gamma U^{-1},\quad \Gamma=\begin{pmatrix} \Gamma_1 & & \\ & \ddots & \\ & & \Gamma_{r+1}\end{pmatrix},$$
$$\Gamma_i\in GL(k,\mathbb{C}),\quad 1\leqslant i\leqslant r+1,$$

where U^{-1} exists in view of (3.7). By construction, we have

$$G\xi_i=\Gamma\xi^S_i\sim\xi^S_i,\quad i=1,\dots,r+1.$$

Moreover

$$G\xi_{r+2}=\begin{pmatrix} \Gamma_1S_1^{-1} \\ \vdots \\ \Gamma_{r+1}S_{r+1}^{-1}\end{pmatrix}=\begin{pmatrix} \Gamma_{r+2} \\ \vdots \\ \Gamma_{r+2}\end{pmatrix}\sim\begin{pmatrix} I_k \\ \vdots \\ I_k\end{pmatrix},$$

where $\Gamma_i=\Gamma_{r+2}S_i$. The existence of $S_i\in GL(k,\mathbb{C})$ and $\Gamma_{r+2}\in GL(k,\mathbb{C})$ follows from (3.8). Finally we have

$$G\xi_{r+3}=\Gamma\begin{pmatrix} R_1 \\ \vdots \\ R_{r+1}\end{pmatrix}=\begin{pmatrix} \Gamma_{r+2}S_1R_1 \\ \vdots \\ \Gamma_{r+2}S_{r+1}R_{r+1}\end{pmatrix}\sim\begin{pmatrix} \Lambda_1 \\ \vdots \\ \Lambda_r \\ I_k\end{pmatrix}$$

with

$$\Lambda_i=\Gamma_{r+2}[S_iR_iR_{r+1}^{-1}S_{r+1}^{-1}]\Gamma_{r+2}^{-1}$$
$$=\Gamma_{r+2}T_i\Gamma_{r+2}^{-1},\quad i=1,\dots,r+1.$$

The condition $\det R_{r+1}\neq 0$ is assured by (3.9). Condition (iv) finally assures that Λ_1 and Λ_2 have the properties required in Theorem 1. Q.E.D.

Notice that sets of $r+3$ initial conditions not satisfying conditions (i)–(iv) form a set of measure zero in all $(r+3)$ tuplets of matrices in $\mathbb{C}^{r(k+1)\times k}$.

IV. THE SUPERPOSITION FORMULA AND LINEARIZATION OF THE MRE

A. Reconstruction of the group element

Let us now turn formula (2.8) into a superposition formula by reconstructing the group element $G(t)$ in (2.9) in terms of $r+3$ particular solutions. In view of Theorem 2 we can restrict ourselves to the case when the initial conditions for our solutions are given, in homogeneous coordinates, by the standard set ξ_i^S of (3.5).

We parametrize the group element $G(t)$ as in (2.9) and (3.4). Writing (2.8) for the first $r+1$ standard solutions $W_i(t)$ we obtain

$$M_{ij} = W_{ji}P_j, \quad N_i = W_{r+1,i}Q, \quad i,j = 1,...,r, \tag{4.1}$$

where we put

$$W_j(t) = \begin{pmatrix} W_{j1}(t) \\ \vdots \\ W_{jr}(t) \end{pmatrix}, \quad j = 1,...,r+3. \tag{4.2}$$

Using $W_{r+2}(t)$ we obtain a system of inhomogeneous linear equations for P_i in terms of the known solutions $W_j(t)$ $(j = 1,...,r+2)$ and the still unknown matrix $Q(t) \in \mathbb{C}^{k\times k}$:

$$\begin{aligned} &\widetilde{W}\begin{pmatrix} P_1 \\ \vdots \\ P_r \end{pmatrix} = (W_{r+1} - W_{r+2})Q, \\ &\widetilde{W} = \begin{pmatrix} W_{r+2,1} - W_{11},...,W_{r+2,1} - W_{r1} \\ \vdots \\ W_{r+2,r} - W_{1r},...,W_{r+2,r} - W_{rr} \end{pmatrix}. \end{aligned} \tag{4.3}$$

The solution exists and is unique as long as

$$\det \widetilde{W} \neq 0. \tag{4.4}$$

Finally, to determine Q we use the remaining solution $W_{r+3}(t)$:

$$\widetilde{\widetilde{W}}\begin{pmatrix} F_1Q\Lambda_1 \\ \vdots \\ F_rQ\Lambda_r \end{pmatrix} = (W_{r+1} - W_{r+3})Q, \tag{4.5}$$

where

$$\begin{pmatrix} F_1 \\ \vdots \\ F_r \end{pmatrix} = \widetilde{W}^{-1}(W_{r+1} - W_{r+2}),$$

$$\widetilde{\widetilde{W}} = \begin{pmatrix} W_{r+3,1} - W_{11} & \cdots & W_{r+3,1} - W_{r1} \\ \vdots & & \vdots \\ W_{r+3,r} - W_{1r} & \cdots & W_{r+3,r} - W_{rr} \end{pmatrix},$$

and

$$\begin{pmatrix} \Lambda_1 \\ \vdots \\ \Lambda_r \end{pmatrix}$$

is defined in Theorem 1.

Using (4.5) and

$$\begin{pmatrix} H_1 \\ \vdots \\ H_r \end{pmatrix} = \widetilde{\widetilde{W}}^{-1}(W_{r+1} - W_{r+3}),$$

we can write the following equations:

$$Q\Lambda_iQ^{-1} = (F_i)^{-1}H_i, \quad i = 1,...,r, \tag{4.6}$$

which determine Q. Note that the matrices $F_i^{-1}H_i$ are conjugated to constant matrices. The existence of $\widetilde{\widetilde{W}}^{-1}$ is assured by the conditions imposed in Theorem 2. For the same reasons F_i^{-1} exists, $i = 1,...,r$ [Theorem 2, (ii)].

B. Linearization of the matrix Riccati equation

An alternative approach to the solution of MRE (2.7) is that of using $r+2$ particular solutions (belonging to a fundamental set of solutions) to transform this equation into a decoupled system of r identical linear homogeneous matrix equations, expressed in the form of commutators. The $(r+3)$rd solution can then be used to express the general solution of this decoupled linear system explicitly.

To obtain this system we perform a series of invertible transformations of the dependent variables. Most of our deliberations will make use of homogeneous coordinates. Thus the MRE (2.7) will, at least temporarily, be replaced by the associated linear equations (2.4), which we rewrite as

$$\begin{pmatrix} \dot{X} \\ \dot{Y} \end{pmatrix} = \Phi\begin{pmatrix} X \\ Y \end{pmatrix}, \quad \Phi = \begin{pmatrix} C_{11} & \cdots & C_{1r} & A_1 \\ \vdots & & & \vdots \\ C_{r1} & \cdots & C_{rr} & A_r \\ -D_1 & \cdots & -D_r & -B \end{pmatrix}. \tag{4.7}$$

Each block in Φ belongs to $\mathbb{C}^{k\times k}$.

We first use a particular solution $W_{(1)}$ of the MRE (2.7) to define an invertible transformation

$$\begin{pmatrix} X^1 \\ Y^1 \end{pmatrix} = \theta_1\begin{pmatrix} X \\ Y \end{pmatrix}, \tag{4.8}$$

with

$$\theta_1 = \begin{pmatrix} 0 & 0 & 0 & I \\ I & 0 & 0 & -W_{(1)1} \\ \vdots & \ddots & & \vdots \\ 0 & 0 & I & -W_{(1)r} \end{pmatrix}, \quad \det\theta_1 = (-1)^{rk}. \tag{4.9}$$

The transformed variables satisfy a simpler equation, namely,

$$\begin{pmatrix} \dot{X}^1 \\ \dot{Y}^1 \end{pmatrix} = \Phi^{(1)}\begin{pmatrix} X^1 \\ Y^1 \end{pmatrix}, \tag{4.10}$$

with

$$\Phi^{(1)} = [\theta_1\Phi + \dot{\theta}_1]\theta_1^{-1}. \tag{4.11}$$

More explicitly, we have

$$\Phi^{(1)} = \begin{pmatrix} -B - \sum_a D_aW_{(1)a} & -D_1 & \cdots & -D_r \\ 0 & C_{11} + W_{(1)1}D_1 & \cdots & C_{1r} + W_{(1)1}D_r \\ \vdots & \vdots & & \vdots \\ 0 & C_{r1} + W_{(1)r}D_1 & \cdots & C_{rr} + W_{(1)r}D_r \end{pmatrix}. \tag{4.12}$$

In affine coordinates we denote

$$W^1 = \theta_1(W) = X^1(Y^1)^{-1}$$

$$= \begin{pmatrix} I \\ W_1 - W_{(1)1} \\ \vdots \\ W_{r-1} - W_{(1)r-1} \end{pmatrix} (W_r - W_{(1)r})^{-1}, \quad (4.13)$$

and the transformation exists for all W such that $\det(W_r - W_{(1)r}) \neq 0$. The transformed quantity W^1 satisfies a MRE of the form (2.7) with coefficients determined by the entries in $\Phi^{(1)}$. To simplify further we use a second solution, say $W_{(2)}$, transform it into $W^1_{(2)} = \theta_1(W_{(2)})$ as in (4.13) and define

$$\Phi^{(2)} = \begin{pmatrix} \phi^{(1)}_{r+1,r+1} + \sum_{a=2}^{r} \phi^{(1)}_{r+1\,a} W^1_{(2)a} & 0 & * & \cdots & * \\ 0 & \phi^1_{11} & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{pmatrix} \quad (4.15)$$

(the stars denote quantities that are, in general, nonvanishing).

In a similar manner we use the first $r+1$ particular solutions to construct the transformation

$$W^{r+1} = \theta_{r+1}[\theta_r[\cdots[\theta_1(W)]\cdots]], \quad (4.16)$$

with

$$\theta_j = \begin{pmatrix} 0 & 0 & \cdots & 0 & I \\ I & 0 & & 0 & -W^{j-1}_{(j)1} \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & \cdots & & I & -W^{j-1}_{(j)r} \end{pmatrix},$$

$$j = 1,\ldots,r+1, \quad (4.17)$$

and $\Phi^{(r+1)}$ diagonal

$$\theta_{r+2} = \begin{pmatrix} 0 & W^{r+1}_{(r+2)1}[W^{r+1}_{(r+2)2}]^{-1} & 0 & \cdots & 0 \\ 0 & 0 & W^{r+1}_{(r+2)1}[W^{r+1}_{(r+2)3}]^{-1} & & \vdots \\ & & & \ddots & \\ \vdots & & & W^{r+1}_{(r+2)1}[W^{r+1}_{(r+2)r}]^{-1} & 0 \\ 0 & 0 & 0 & \cdots & W^{r+1}_{(r+2)} \\ I & 0 & 0 & \cdots & 0 \end{pmatrix}. \quad (4.21)$$

The transformed quantities $W^{r+2}_i = X^{r+2}(Y^{r+2})^{-1}$ satisfy

$$\dot{W}^{r+2}_i = [\tilde{C}, W^{r+2}_i], \quad i = 1,\ldots,r, \quad (4.22)$$

i.e., each component satisfies the same equation (4.22) (the known matrix $\tilde{C}$ does not depend on the label i) and the right-hand side has the form of a commutator. Given one more solution, $W_{(r+3)}$, we use all the previous ones to transform it into $W^{r+2}_{(r+3)}$, satisfying (4.22). The general solution of (4.22) can be written as

$$\begin{pmatrix} X^2 \\ Y^2 \end{pmatrix} = \theta_2 \begin{pmatrix} X^1 \\ Y^1 \end{pmatrix},$$

$$\theta_2 = \begin{pmatrix} 0 & 0 & \cdots & 0 & I \\ I & 0 & \cdots & 0 & -W^1_{(2)1} \\ \vdots & \ddots & & & \vdots \\ 0 & \cdots & & I & -W^1_{(2)r} \end{pmatrix}. \quad (4.14)$$

The transformed quantity

$$W^2 = \theta_2(W^1) = \theta_2[\theta_1(W)]$$

will satisfy a MRE with coefficients determined by the matrix

$$\Phi^{(r+1)} = \begin{pmatrix} \phi^{(r)}_{r+1\,r+1} & & & \\ & \phi^{(r)}_{11} & & \\ & & \ddots & \\ & & & \phi^{(r)}_{rr} \end{pmatrix}. \quad (4.18)$$

If W satisfies the MRE (2.7) then the transformed quantity W^{r+1} satisfies the linear decoupled system

$$\dot{W}^{r+1}_i = W^{r+1}_i \tilde{B} + \tilde{C}_{ii} W^{r+1}_i, \quad i = 1,\ldots,r. \quad (4.19)$$

So far, each component W^{r+1}_i satisfies a different equation. To simplify further we use one more solution, $W_{(r+2)}$, and construct a different transformation, namely,

$$\begin{pmatrix} X^{r+2} \\ Y^{r+2} \end{pmatrix} = \theta_{r+2} \begin{pmatrix} X^{r+1} \\ Y^{r+1} \end{pmatrix}, \quad (4.20)$$

with

$$W^{r+2}_i = G(t) U_i G^{-1}(t), \quad i = 1,\ldots,r, \quad (4.23)$$

where $U_i \in \mathbb{C}^{k\times k}$ is a constant matrix. Choosing the initial conditions for $W_{(r+3)}$ such that we have

$$U_{(r+3)} = \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \\ \vdots \\ \Lambda_r \end{pmatrix},$$

where Λ_1 and Λ_2 have no common nontrivial irreducible eigenspaces and Λ_1 is diagonalizable with all eigenvalues different, we can completely reconstruct $G(t)$.

Without proof we state that if the $r+3$ solutions used above satisfy the conditions of Theorem 2, then all the transformations θ_i $(i=1,...,r+2)$ exist and are invertible. For a general r the explicit formulas are quite complicated and it is best to follow the described procedure as a recursive algorithm.

V. CONCLUSIONS

The problem posed in the Introduction, namely that of obtaining the general solution of a rectangular MRE for a matrix $W \in \mathbb{C}^{n\times k}$ with $n=rk$ $(r\geqslant 2,\ k\geqslant 2)$ in terms of $r+3$ particular solutions has been solved. If $r+3$ particular solutions, satisfying the conditions discussed in Sec. III, are known analytically, then the superposition formula of Sec. IV amounts to a general analytical solution. If the required particular solutions are not available, then the superposition formula, or the linearization technique, can be viewed as a numerical method. Thus, a fundamental set of $r+3$ particular solutions can be obtained numerically, starting from well chosen initial conditions, such that the solutions have no singularities in the considered region of t. Further solutions, corresponding to other initial conditions, can then be obtained via the superposition formula.

Such a procedure has so far been implemented for square MRE's only.[14,15] It is particularly efficient when large matrices are involved, when we are interested in solutions that have singularities for real values of t, or when a large number of solutions, corresponding to different initial values is required.

Let us mention that the results of Sec. IV provide insight into the properties of the solution set of rectangular MRE. In particular they imply that the MRE (2.7) has the Painlevé property: the only moving singularities that can develop in the solutions are poles.

The reason why we restricted ourselves to the case $n=rk$ in this paper is that this allowed us to present all formulas and arguments in terms of the matrices $W_i(t)$ of (3.3). More generally, for $n=rk+l$, $1\leqslant l\leqslant r-1$, we have found it necessary to proceed differently and to argue in terms of the matrix elements of $W(t)$ directly.

ACKNOWLEDGMENTS

The research reported in this paper was supported in part by the Natural Sciences and Engineering Research Council of Canada, the "Fonds FCAR du Gouvernement du Québec," and the "Ministerio de Educación y Ciencia" of Spain (Grants to M. A. d. O. and M. A. R. from the "Plan de Formación de Personal Investigador").

[1] S. Lie and G. Scheffers, *Vorlesungen über continuierlichen Gruppen mit geometrischen und anderen Anwendungen* (Teubner, Leipzig, 1893) (reprinted by Chelsea, New York, 1967).
[2] S. Shnider and P. Winternitz, Lett. Math. Phys. **8**, 69 (1984); J. Math. Phys. **25**, 3155 (1984).
[3] V. W. Guillemin and S. Sternberg, Bull. Am. Math. Soc. **70**, 16 (1964).
[4] M. Golubitsky, J. Diff. Geom. **7**, 175 (1972).
[5] M. A. del Olmo, M. A. Rodriguez, and P. Winternitz, J. Math. Phys. **27**, 14 (1986).
[6] R. L. Anderson, Lett. Math. Phys. **4**, 1 (1980).
[7] R. L. Anderson, J. Harnad, and P. Winternitz, Physica D **4**, 164 (1982).
[8] J. Harnad, P. Winternitz, and R. L. Anderson, J. Math. Phys. **24**, 1062 (1983).
[9] J. Beckers, V. Hussin, and P. Winternitz, Lett. Math. Phys. **11**, 81 (1986).
[10] W. T. Reid, *Riccati Differential Functions* (Academic, New York, 1972).
[11] T. Ochiai, Trans. Am. Math. Soc. **124**, 313 (1966).
[12] J. Wolf, Mem. Am. Math. Soc. **180**, 1 (1976).
[13] S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces* (Academic, New York, 1978).
[14] D. Rand and P. Winternitz, Comput. Phys. Comm. **33**, 305 (1984).
[15] M. Sorine and P. Winternitz, IEEE Trans. Autom. Control AC **30**, 266 (1985).

On a property of a classical solution of the nonlinear mass transport equation $u_t = u_{xx}/1 + u_x^2$. II

Akihiko Kitada
Liberal Arts, Nippon College of Physical Education, Fukazawa, Setagaya-ku, Tokyo 158, Japan
Hiroyuki Umehara
Industrial Products Research Institute, Yatabe-machi Higashi, Tsukuba-gun, Ibaraki 305, Japan

(Received 7 August 1986; accepted for publication 5 November 1986)

A mechanism of smoothing due to evaporation condensation of the roughly perturbed surface of solid is formulated by Mullins [W. W. Mullins, J. Appl. Phys. **28**, 333 (1957); **30**, 77 (1959)] as a certain Cauchy problem for a nonlinear parabolic equation which describes the evolution of the profile of the surface. In the preceding paper [A. Kitada, J. Math. Phys. **27**, 1391 (1986)], through the careful investigations of the Cauchy problem, it was demonstrated that each peak in the initial surface did *not increase* in height with time. In the present paper, by slightly limiting the set of functions to which the classical solutions of the Cauchy problem belong, it is demonstrated that each peak height *decreases* with time in the strict sense.

I. INTRODUCTION

In the preceding paper,[1] we proposed the relation

$$u(x,t) \leqslant \alpha(x_0), \quad (x,t) \in C - \{(x_0,0)\}, \tag{1}$$

which estimates the variation with time, due to evaporation condensation, in height of a peak in a roughly perturbed surface of solid. Here, $u(x,t)$ is a classical solution[2] of the Cauchy problem (P) (Mullins' model[3]) in the real line $\mathbb{R}^1$,

$$\begin{aligned} &u_t = u_{xx}/1 + u_x^2, \quad (x,t) \in \mathbb{R}^1 \times (0,\infty), \\ &u(x,0) = \alpha(x), \quad x \in \mathbb{R}^1, \end{aligned} \tag{P}$$

describing the evolution of the profile of the surface of solid; and the subset C of the real plane $\mathbb{R}^2$, which is a graph of a differentiable function $g(t)$ defined on some closed interval $[0,t_f]$ in $\mathbb{R}^1$, forms a part of a trajectory in the x,t plane drawn by the migration with time of a peak top initially located at the point $(x_0,0)$. That is, the curve C is characterized by

$$C = \{(x,t);\ x = g(t)\ (x_0 = g(0)),\ t \in [0,t_f]\}, \tag{2a}$$

$$u_x(x,t) = 0, \quad u_{xx}(x,t) < 0, \quad (x,t) \in C. \tag{2b}$$

In the present paper, by slightly limiting the set of function to which the classical solutions of the Cauchy problem (P) belong, we show that the relation (1) holds without sign of equality, that is, the peak height decreases with time in the strict sense. It is what the Mullins model has desired for the estimate without sign of equality to hold.

II. AN ESTIMATE DESCRIBING THE STRICTLY MONOTONE DECREASE OF THE PEAK HEIGHT

By demonstrating the more general estimate

$$u(x_2,t_2) < u(x_1,t_1), \quad (x_i,t_i) \in C\ (i = 1,2), \quad t_1 < t_2, \tag{3}$$

we will show the validities of the relation (1) without sign of equality.

The following theorem guarantees this strictly monotone decrease with time of the peak height.

Theorem: Consider a Cauchy problem (P*)

$$\begin{aligned} &u_t = F(u_x, u_{xx}), \quad (x,t) \in \mathbb{R}^1 \times (0,\infty), \\ &u(x,0) = \alpha(x), \quad x \in \mathbb{R}^1. \end{aligned} \tag{P*}$$

Let the conditions

$$\begin{aligned} &F \in C^2(\mathbb{R}^2)\ \text{(Ref. 4)}, \\ &F_q(p,q) > 0\ \text{(Ref. 5)}, \\ &F(0,0) = 0 \end{aligned} \tag{C}$$

hold for the right-hand side $F(p,q)$ of the nonlinear equation in (P*). Suppose, for such a classical solution $u(x,t)$ of (P*) that $u \in C^3(\mathbb{R}^1 \times (0,\infty))$, there exists a set C characterized by (2a) and (2b). Then the relation (3) holds for such a solution of (P*).

First of all, from our discussions in the preceding paper,[1] it is evident[6] that at least the relation (3'), i.e., the relation (3) with sign of equality,

$$u(x_2,t_2) \leqslant u(x_1,t_1) \quad (x_i,t_i) \in C\ (i = 1,2), \quad t_1 < t_2, \tag{3'}$$

must hold even for the ordinary classical solution of (P*).

In order to demonstrate the above theorem, we prepare a well-known lemma due to Nirenberg[7] for a linear parabolic equation.

Lemma: Let D be a bounded connected open set in $\mathbb{R}^1 \times (0,\infty)$ and let the coefficients $a(x,t)$ and $b(x,t)$ of a linear equation

$$a(x,t)u_{xx} + b(x,t)u_x - u_t = 0, \quad (x,t) \in \mathbb{R}^1 \times (0,\infty) \tag{4}$$

obey the conditions (5) in D

$$|a(x,t), b(x,t)| < \infty, \quad a(x,t) \geqslant \mu, \tag{5}$$

where μ is some positive constant. If there exists a constant M such that

$$u(x,t) \leqslant M, \quad (x,t) \in D,$$

and there exists a point $\xi \in D$ such that

$$u(\xi) < M,$$

then the relation

 0022-2488/87/030536-02$02.50

$$u(\gamma) < M$$

holds. Here, γ is a horizontal line segment (line segment which is parallel to the x axis) containing the point ξ as an internal point and is itself contained in D.

Proof of Theorem: As is pointed out in our preceding paper,[1] the solution of (P*) satisfies the following homogeneous linear equation under the conditions (C):

$$u_{xx}\int_0^1 F_q(hu_x(x,t),hu_{xx}(x,t))dh + u_x\int_0^1 F_p(hu_x(x,t),hu_{xx}(x,t))dh - u_t = 0. \tag{6}$$

If the coefficients in (6) are continuous in $\mathbb{R}^1\times(0,\infty)$, the first condition in (5) is well satisfied in any bounded open connected set D whose closure is contained in the set $\mathbb{R}^1\times(0,\infty)$ because of the compactness of the closure of D. Then the second condition in (5) is also satisfied in any compact set in $\mathbb{R}^1\times(0,\infty)$ because the function F_q is everywhere positive as is indicated in (C). As the solution $u(x,t)$ is assumed to belong to the set of function $C^3(\mathbb{R}^1\times(0,\infty))$, the difference between the value of the function $a(x,t)$ at the point (x^*,t^*) and the value at any point (x,t) which is close enough to the point (x^*,t^*) is estimated as follows with some positive constant L, under the first condition in (C):

$$\begin{aligned}
&|a(x,t)-a(x^*,t^*)|\\
&\leqslant\int_0^1 |F_q(hu_x(x,t),hu_{xx}(x,t))\\
&\quad - F_q(hu_x(x^*,t^*),hu_{xx}(x^*,t^*))|dh\\
&=\int_0^1 h\,|F_{qp}(\tau(h),\eta(h))\{u_{xx}(\theta,\xi)(x-x^*)\\
&\quad + u_{xt}(\theta,\xi)(t-t^*)\}\\
&\quad + F_{qq}(\tau(h),\eta(h))\{u_{xxx}(\theta',\xi')(x-x^*)\\
&\quad + u_{xxt}(\theta',\xi')(t-t^*)\}|dh\\
&\leqslant L(|x-x^*|+|t-t^*|),
\end{aligned}$$

where $\tau(h)$ is some value between $hu_x(x,t)$ and $hu_x(x^*,t^*)$, $\eta(h)$ between $hu_{xx}(x,t)$ and $hu_{xx}(x^*,t^*)$, θ, θ' between x and x^*, and ξ, ξ' between t and t^*. As the same is true for the coefficient $b(x,t)$, all the coefficients in (6) are continuous at any point in consideration. Since all the requirements for the coefficients of linear equation are satisfied, we can apply the Lemma to the solution of the linear equation (6), that is, to the solution in $C^3(\mathbb{R}^1\times(0,\infty))$ of the problem (P*).

Let (x_1,t_1) and (x_2,t_2) $(t_1<t_2)$ be two arbitrary points in the set $C-(\{(x_0,0)\}\cup\{(g(t_f),t_f)\})$. Since $u_x(x,t)$ and $u_{xx}(x,t)$ are continuous and the relations $u_x(x_2,t_2)=0$ and $u_{xx}(x_2,t_2)<0$ hold, we can define a continuous implicit function $f(t)$ such that $u_x(f(t),t)=0$ on some interval in the t axis which contains the point t_2 as an internal point. It is clear from the elementary proof of the implicit function theorem that there exists an open rectangle $\Omega=(x',x'')\times(t',t'')$ such that $u_x(x,t)>0$ at any point of the set

$$\{(x,t);\ x'<x<f(t),\ t\in(t',t'')\}$$

and $u_x(x,t)<0$ at any point of the set

$$\{(x,t);\ f(t)<x<x'',\ t\in(t',t'')\}.$$

Here, we may take t' and t'' as $t_1<t'<t_2<t''$. Then, in the open set Ω, the estimate

$$u(x,t)\leqslant\sup_{(t',t'')}u(f(t),t)$$

holds. Now, since the implicit function is uniquely determined in Ω as is well known in the elementary differential calculus, the function $f(t)$ must be equal to the function $g(t)$ given in (2a) on the open interval (t',t'').[8] Thus taking the estimate (3′) into account, we obtain

$$u(x,t)\leqslant\sup_{(t',t'')}u(g(t),t)\leqslant u(x_1,t_1),\quad (x,t)\in\Omega.$$

If we take the open set Ω as the set D and the value $u(x_1,t_1)$ as the constant M in the Lemma, we obtain the fact that the function $u(x,t)$ cannot have the value $u(x_1,t_1)$ at the point (x_2,t_2). Therefore we get the following relation:

$$\begin{aligned}
&u(x_2,t_2)<u(x_1,t_1),\\
&(x_i,t_i)\in C-(\{(x_0,0)\}\cup\{(g(t_f),t_f)\})\\
&\qquad (i=1,2),\quad t_1<t_2.
\end{aligned}$$

The relation (3′) guarantees that $u(x_0,0)$ is not less than the value of $u(x,t)$ at any point in $C-\{(x_0,0)\}$ and the relation $u(g(t_f),t_f)\leqslant u(x,t)$ must be valid at any point (x,t) in $C-\{(g(t_f),t_f)\}$. Therefore we can concludingly obtain the desired estimate (3) on the whole trajectory C. □

Since the right-hand side $F(p,q)=q/1+p^2$ of the nonlinear mass transport equation satisfies all the conditions required in (C), the relation (3), that is, the relation (1) without sign of equality, holds for the classical solution $u(x,t)$ in $C^3(\mathbb{R}^1\times(0,\infty))$ of the problem (P).

[1] A. Kitada, J. Math. Phys. **27**, 1391 (1986).

[2] O. A. Ladyzenskaja, V. A. Solonikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type* (Am. Math. Soc., Providence, RI, 1968), p. 12.

[3] W. W. Mullins, J. Appl. Phys. **28**, 333 (1957); **30**, 77 (1959).

[4] The symbol $C^m(\Omega)$ denotes the set of all functions defined on Ω whose partial derivatives of order $\leqslant m$ are all continuous.

[5] The partial derivative $\partial F/\partial q$ is abbreviated as F_q. In the same manner, for example, we write $\partial/\partial p(F_q)$ as F_{qp}.

[6] For the point $(x_1,t_1)\in C-\{(g(t_f),t_f)\}$, the estimate (5) in the preceding paper,[1] can be easily generalized to

$$u(x^*,t^*)<\max\{\sup_{C'-\{(x_1,t_1)\}}[f(x,t)\exp\lambda(t^*-t)/\lambda],\ u(x_1,t_1)\exp\lambda(t^*-t_1)\},$$

where $C'=\{(x,t)\in C;\ t_1\leqslant t\leqslant t_f\}$ and $(x^*,t^*)\in C'-\{(x_1,t_1)\}$.

[7] L. Nirenberg, Commun. Pure Appl. Math. **6**, 167 (1956).

[8] Therefore no crossing of the two different trajectories takes place.

The evolution partial differential equation $u_t = u_{xxx} + 3(u_{xx}u^2 + 3u_x^2 u) + 3u_x u^4$

F. Calogero

Centro Linceo Interdisciplinare di Scienze Matematiche e loro Applicazioni, Accademia Nazionale dei Lincei, Roma, Italy[a]*; Dipartimento di Fisica, Università di Roma "La Sapienza," 00185 Roma, Italy*[b]*; and Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Roma, Italy*

(Received 24 June 1986; accepted for publication 22 October 1986)

The evolution equation $u_t = u_{xxx} + 3(u_{xx}u^2 + 3u_x^2 u) + 3u_x u^4$, $u = u(x,t)$, is integrable; it can be (exactly) linearized by an appropriate change of (dependent) variable. Hence several explicit solutions of this partial differential equation can be exhibited; some of them display a remarkable solitronic phenomenology.

I. INTRODUCTION

This paper is devoted to a study of the evolution partial differential equation (PDE)

$$u_t = u_{xxx} + 3(u_{xx}u^2 + 3u_x^2 u) + 3u_x u^4. \tag{1.1}$$

Here, and throughout this paper, $u \equiv u(x,t)$. This equation can be linearized by an appropriate change of dependent variable; hence it is integrable, and indeed several of its solutions can be explicitly exhibited.

Many other nonlinear PDE's that can also be linearized by appropriate changes of variables are known. A classic example is Burger's equation,[1]

$$u_t = u_{xx} + u_x u, \tag{1.2}$$

together with a few of its variants, for instance,

$$u_t = u_{xx}u^2 + 1, \tag{1.2'}$$

$$u_t = u_{xx}u^2 + u^2, \tag{1.2''}$$

$$u_t = u_{xx}u^2 + u_x^2. \tag{1.3'''}$$

These are all second-order PDE's of parabolic type (with one dependent, and two independent, variables, as are all the PDE's mentioned in this paper).

Another integrable PDE of second order and of parabolic type reads

$$u_t = (u_{xx}/u_x^2) f_1(u) + u_x f_2(u) + f_3(u). \tag{1.3}$$

Here, as in all the equations of this section, the functions f_m are arbitrary (they could generally depend on t, in addition to the argument shown explicitly, without spoiling the integrability). A detailed analysis of this equation, and of some of those listed above and below, shall perhaps be published elsewhere.

A second-order PDE of hyperbolic type that is also integrable by quadratures reads[2]

$$u_{xt} = u_{xx}u + f(u_x). \tag{1.4}$$

It is a special case of the following equation of third order, which is also integrable:

$$u_{xt} = (u_{xxx}/u_{xx}^2) f_1(u_x) + u_{xx} f_2(u_x) u + u_{xx} f_3(u_x) + f_4(u_x). \tag{1.5}$$

Many other evolution equations that are also integrable by quadratures can be manufactured, such as the following ones:

$$u_t = (u_{xxx}/u_{xx}^3) f_1(u_x) + (1/u_{xx}) f_2(u_x) + f_3(u_x)u + f_4(u_x), \tag{1.6}$$

$$u_t = [(u_{xxx}/u_x^3) - 3(u_{xx}^2/u_x^4)] f_1(u) + (u_{xx}/u_x^2) f_2(u) + f_3(u), \tag{1.7}$$

$$u_t = [(u_{xxx}/u_x^3) - \tfrac{9}{4}(u_{xx}^2/u_x^4)] f_1(u) + \tfrac{1}{2}(u_{xx}/u_x^2) f_1'(u) + f_2(u), \tag{1.8}$$

$$u_t = u_{xxx}u^3 + c u_{xx} u_x u^2, \quad c = 3 \text{ or } c = \tfrac{3}{2}. \tag{1.9}$$

Note that, for $c = 3$, the last equation can be recast, via the change of variable $u = -(2)^{1/3} v^{-1}$, into the form

$$v_t = (v^{-2})_{xxx}; \tag{1.9'}$$

while for $c = 0$ the PDE (1.9), which in this case cannot, of course, be integrated by quadratures, can be transformed, via the change of variable $v = -(2)^{1/3} u^{-2}$, into the Harry Dym equation,

$$v_t = (v^{-1/2})_{xxx}. \tag{1.9''}$$

As is well known (see, for instance, Ref. 3, p. 290ff), this PDE can be reduced, by a nontrivial change of dependent and independent variables, to the Korteweg–de Vries equation.

The motivation to focus in this paper on the evolution equation (1.1) rests on its resemblance to the Korteweg–de Vries and modified Korteweg–de Vries equations, on the simplicity of the linearizing transformation (see Sec. III), and on the remarkable solitronic (rather than solitonic; for this terminology see Ref. 3, p. 132ff) phenomenology displayed by some of its solutions (see Sec. V); indeed (1.1) supports kinklike solitrons of three different kinds, as well as a periodic traveling wave solution and semi-infinite traveling wave solutions; and explicit solutions can be exhibited that display inelastic collisions of these objects.

II. PRELIMINARIES

The nonlinear evolution equation (1.1) is clearly invariant under translations of the time variable t and the space variable x; moreover, a term cu_x could be added in the rhs by the (Galileian) change of variable $x \to x' = x + ct$. Under the rescaling transformation

$$u(x,t) = a u'(x',t'), \quad x' = bx, \quad t' = ct, \tag{2.1}$$

the PDE (1.1) goes into

[a] For the academic years 1983–1984, 1984–1985 and 1985–1986.
[b] Permanent address.

 0022-2488/87/030538-18$02.50

$$u'_{t'} = Au'_{x'x'x'} + 3B(u'_{x'x'}u'^2 + 3u'^2_{x'}u') + 3B^2u'_{x'}u'^4, \quad A = b^3/c, \quad B = a^2/b; \tag{2.2}$$

in particular it is invariant under the transformation

$$u(x,t) = au'(x',t'), \quad x' = a^2x, \quad t' = a^6t. \tag{2.3}$$

Two special cases of this transformation are worth noticing:

$$u(x,t) = -u'(x,t) \quad (a = -1), \tag{2.3'}$$

$$u(x,t) = iu'(-x,-t) \quad (a = i). \tag{2.3''}$$

In this paper, however, attention will be generally confined to real (and nonsingular) solutions; note that, in such a context, (1.1) is not invariant under time and/or space reversal.

Other avatars of the PDE (1.1) may be obtained by changes of variables (see also Sec. III). For instance

$$w(x,t) = u^2(x,t), \tag{2.4a}$$

$$w_t = (w_{xx} - \tfrac{3}{4}w_x^2/w + 3w_xw + w^3)_x, \tag{2.4b}$$

and

$$U_x(x,t) = w(x,t) = u^2(x,t), \tag{2.5a}$$

$$U_t = U_{xxx} - \tfrac{3}{4}U_{xx}^2/U_x + 3U_{xx}U_x + U_x^3. \tag{2.5b}$$

Clearly the PDE (1.1) possesses traveling wave solutions

$$u(x,t) = g(x - Vt), \tag{2.6a}$$

$$-Vg' = g''' + 3(g''g + 3g'^2g) + 3g'g^4. \tag{2.6b}$$

This ordinary differential equation (ODE) can be easily integrated once, after multiplication by g. The fact that it can be explicitly integrated two more times is less obvious (see Appendix A). The solutions of type (2.6) of (1.1) that represent kinklike solitrons or periodic traveling waves are discussed below (see Sec. V and Appendix A).

The nonlinear PDE (1.1) also possesses similarity solutions of the following type:

$$u(x,t) = a(t)f[b(t)x], \tag{2.7a}$$

$$a(t) = [(t - t_0)/c]^{-1/6}, \tag{2.7b}$$

$$b(t) = a^2(t) = [(t - t_0)/c]^{-1/3}, \tag{2.7c}$$

$$2yf' + f + 6c[f''' + 3(f''f^2 + 3f'^2f) + 3f'f^4] = 0. \tag{2.7d}$$

Again, it is clear that the ODE (2.7d) can be integrated once; less trivial is the possibility to integrate it two more times (see Appendix B).

III. SOLUTION BY LINEARIZATION

Let $v(x,t)$ satisfy the linear PDE

$$v_t(x,t) = v_{xxx}(x,t), \tag{3.1}$$

and set

$$u(x,t) = v(x,t)/[2V(x,t)]^{1/2} \tag{3.2}$$

with

$$V_x(x,t) = [v(x,t)]^2 \tag{3.3a}$$

and

$$V_t(x,t) = 2v_{xx}(x,t)v(x,t) - [v_x(x,t)]^2. \tag{3.3b}$$

It is then easily seen that $u(x,t)$ satisfies the nonlinear PDE (1.1). Note the consistency of (3.3a) and (3.3b) with (3.1).

Note that (3.3a) and (3.3b) imply that $V(x,t)$ satisfies the nonlinear PDE

$$V_t = V_{xxx} - \tfrac{3}{4}V_{xx}^2/V_x, \tag{3.4}$$

while (3.2) and (3.3a) yield

$$u(x,t) = [V_x(x,t)/2V(x,t)]^{1/2}. \tag{3.5}$$

Hence the nonlinear PDE (3.4) can be seen as another avatar of (1.1), obtained by the "change of variable" (3.5); indeed, the linear PDE (3.1) could itself be interpreted as an avatar of (1.1), generated by the nonlinear transformation (3.2) with (3.3).

Consider the class of real solutions of (3.1) such that $v(x,t)$ vanishes as $x \to -\infty$ faster than $(-x)^{-1/2}$ and is regular for real x. It is then convenient to write

$$V(x,t) = \int_{-\infty}^{x} dx'[v(x',t)]^2 + \frac{1}{2}C^2, \tag{3.6}$$

so that (3.2) yields

$$u(x,t) = \frac{v(x,t)}{\{C^2 + 2\int_{-\infty}^{x} dx'[v(x',t)]^2\}^{1/2}}. \tag{3.7a}$$

This equation can be inverted, via (3.2), (3.3a), and (3.5), and one finds

$$v(x,t) = Cu(x,t)\exp\left\{\int_{-\infty}^{x} dx'[u(x',t)]^2\right\}. \tag{3.7b}$$

These two formulas, (3.7a) and (3.7b) (with, say, $C = 1$), provide the basis for solving the Cauchy problem for (1.1): given $u(x,0)$ one computes $v(x,0)$ from (3.7b), then $v(x,t)$ following the linear evolution (3.1), and finally one obtains $u(x,t)$ from $v(x,t)$ via (3.7a). Note that this technique of solution implies that, if $u(x,0)$ belongs to the class of real functions that vanish faster than $(-x)^{-1/2}$ as $x \to -\infty$ and are regular for real x, then $u(x,t)$ belongs to the same class for all values of t. If moreover $u(x,0)$ is Fourier expandable [for which it is required that it vanish at both ends, $u(\pm\infty,0) = 0$], then $u(x,t)$ is also Fourier expandable; as well of course as $v(x,t)$,

$$v(x,t) = (2\pi)^{-1}\int_{-\infty}^{+\infty} dk \exp(ikx)\hat{v}(k,t), \tag{3.8a}$$

$$\hat{v}(k,t) = \int_{-\infty}^{+\infty} dx \exp(-ikx)v(x,t). \tag{3.8b}$$

And of course in this "localized" case the time evolution of the Fourier transform $\hat{v}(k,t)$ of $v(x,t)$ is quite trivial,

$$\hat{v}(k,t) = \hat{v}(k,0)\exp(-ik^3t); \tag{3.9}$$

and this fact, together with the direct and inverse Fourier transform formulas (3.8b) and (3.8a), yield in the standard manner the solution of the Cauchy problem for the linear evolution equation (3.1).

It is moreover plain how to obtain, via (3.7a) and (3.7b), from the *linear superposition formula* [according to which, if $v_1(x,t)$ and $v_2(x,t)$ are solutions of (3.1), their linear superposition,

$$v(x,t) = C_1v_1(x,t) + C_2v_2(x,t) \tag{3.10}$$

also satisfies (3.1)], the following *nonlinear superposition formula* according to which, if $u_1(x,t)$ and $u_2(x,t)$ are two solutions of the nonlinear PDE (1.1), then

$$u(x,t) = \{C_1 u_1(x,t)\exp[U_1(x,t)] + C_2 u_2(x,t)\exp[U_2(x,t)]\}\Big\{C_3 + C_1^2 \exp[2U_1(x,t)] + C_2^2 \exp[2U_2(x,t)]$$
$$+ 4C_1C_2 \int_{-\infty}^{x} dx'\, u_1(x',t)u_2(x',t)\exp[U_1(x',t) + U_2(x',t)]\Big\}^{-1/2} \tag{3.11}$$

is a third solution of (3.1). Here C_1, C_2, and $C_3 = C^2 - C_1^2 - C_2^2$ are three arbitrary constants and we have used the convenient notation [see (2.5a)]

$$U_j(x,t) = \int_{-\infty}^{x} dx' [u_j(x',t)]^2, \quad j = 1,2. \tag{3.12}$$

IV. CONSERVED QUANTITIES

The possibility to linearize the nonlinear PDE (1.1) (see preceding section) implies that an infinity of conservation laws can be associated with this evolution equation. To present these results in the simplest setting, let us limit our consideration, in this section, to localized solutions, namely regular solutions that vanish asymptotically ($x \to \pm\infty$) sufficiently fast to guarantee the convergence of all the integrals written below. We moreover restrict our treatment to the exhibition of space integrals of $u(x,t)$ that remain constant, or evolve simply with time, as $u(x,t)$ evolves according to (1.1). The formulation of these results in terms of local conservation laws is an easy task that is left for the diligent reader; of course such a formulation has a broader validity than the results reported below, since it is applicable also to solutions which do not vanish asymptotically or are not regular for some real value of the space variable x.

All these results obtain easily, via (3.7b), from the analogous results for the solutions $v(x,t)$ of the linear evolution equation (3.1). Let us therefore begin with a terse review of these elementary results, whose proof is, for the sake of completeness, outlined in Appendix C.

Of course, if $v(x,t)$ is Fourier expandable, see (3.8a) and (3.8b), then, for all values of the Fourier parameter k, the modulus of the Fourier component $\hat{v}(k,t)$ is time independent,

$$|\hat{v}(k,t)| = |\hat{v}(k,0)| \tag{4.1}$$

[see (3.9)]. But it is more convenient to focus attention on the following infinite but denumerable set of conserved (i.e., time-independent) quantities:

$$C_m = \int_{-\infty}^{+\infty} dx [v^{(m)}(x,t)]^2, \quad m = 0,1,2,,\ldots . \tag{4.2}$$

Here and below we use the shorthand notation

$$v^{(m)}(x,t) \equiv \frac{\partial^m v(x,t)}{\partial x^m} . \tag{4.3}$$

Indeed, it is also of interest to introduce the more general set

$$X_{n,m}(t) = (n!)^{-1} \int_{-\infty}^{+\infty} dx\, x^n [v^{(m)}(x,t)]^2,$$
$$n = 0,1,2,\ldots, \quad m = 0,1,2,\ldots, \tag{4.4}$$

and to note that $X_{n,m}(t)$ evolves in time as a polynomial of degree n,

$$X_{n,m}(t) = \sum_{s=0}^{n} \left[X_{n,m}^{(s)} \frac{t^s}{s!}\right] . \tag{4.5}$$

The constant coefficients $X_{n,m}^{(s)}$ satisfy the recursion relations

$$X_{n,m}^{(s+1)} = 3X_{n-1,m+1}^{(s)} - X_{n-3,m}^{(s)}, \quad s = 0,1,\ldots,n-1 . \tag{4.6}$$

Here we use the convention

$$X_{n,m}^{(s)} = 0, \quad \text{if } n<0 \text{ or } s>n, \tag{4.7}$$

while of course

$$X_{0,m}^{(0)} = X_{0,m} = C_m \tag{4.8}$$

[see (4.4) and (4.2)]. Hence

$$X_{1,m}(t) = X_{1,m}(0) + 3C_{m+1}t, \tag{4.9a}$$
$$X_{2,m}(t) = X_{2,m}(0) + 3X_{1,m+1}(0)t + \tfrac{3}{2}C_{m+2}t^2, \tag{4.9b}$$
$$X_{3,m}(t) = X_{3,m}(0) + 3X_{2,m+1}(0) - C_m t + \tfrac{3}{2}X_{1,m+2}(0)t^2 + \tfrac{3}{2}C_{m+3}t^3, \tag{4.9c}$$

and so on.

Another interesting set of moments is given by the definition

$$Y_n(t) = (-)^n (n!)^{-1} \int_{-\infty}^{+\infty} dx\, x^n v(x,t), \quad n = 0,1,2,\ldots, \tag{4.10}$$

for it is easily shown (see Appendix C) that $Y_n(t)$ evolves in time as a polynomial of degree $[(n/3)]$ [here and below $[(n/3)]$ indicates the integral part of $n/3$]:

$$Y_n(t) = \sum_{s=0}^{[(n/3)]} \left[Y_n^{(s)} \frac{t^s}{s!}\right] . \tag{4.11}$$

Moreover the constant coefficients $Y_n^{(s)}$ satisfy the recursion relation

$$Y_n^{(s+1)} = Y_{n-3}^{(s)} , \tag{4.12}$$

where we assume of course that

$$Y_n^{(s)} = 0, \quad \text{if } n<0 \text{ or } s>[(n/3)]. \tag{4.13}$$

Hence

$$Y_n(t) = Y_n(0) = Y_n, \quad n = 0,1,2, \tag{4.14a}$$
$$Y_n(t) = Y_n(0) + Y_{n-3}t, \quad n = 3,4,5, \tag{4.14b}$$
$$Y_n(t) = Y_n(0) + Y_{n-3}(0)t + \tfrac{1}{2}Y_{n-6}t^2, \quad n = 6,7,8, \tag{4.14c}$$

and so on.

Let us note that these results imply that there exist, in addition to the set (4.2), many other constants of the motion, such as the first three elements of the set (4.10) [see (4.14a)], or appropriate combinations of the quantitites defined above, for instance $Y_0Y_4 - Y_1Y_3$ [see (4.14a) and (4.14b)] or the set $Y_0X_{1,m} - 3C_{m+1}Y_3$ [see (4.9a), (4.14a), and (4.14b)]; and of course many more.

It is now easy to obtain analogous results for the nonlin-

ear evolution Eq. (1.1), since they are obtained directly by replacing in the preceding formulas the field $v(x,t)$ by its expression in terms of the solution $u(x,t)$ of (1.1),

$$v(x,t) = u(x,t)\exp[U(x,t))] \tag{4.15a}$$

[see (3.7b)]. Here and below we use the convenient notation

$$U(x,t) = \int_{-\infty}^{x} dx'[u(x',t)]^2 \tag{4.15b}$$

[see (2.5a) and (3.12)]. Let us emphasize that, by redefining in this manner [via (4.15)] the quantities $C_m, X_{n,m}(t)$, and $Y_n(t)$ in terms of the solution $u(x,t)$ of (1.1), rather than the solution $v(x,t)$ of (3.1), one does not modify their (simple) time evolution, which has been detailed above.

For instance the first three conserved quantities C_m, see (4.2), may be written in terms of $u(x,t)$ as follows:

$$C_0 = \tfrac{1}{2}\{\exp[2U(\infty,t) - 1]\}, \tag{4.16a}$$

$$C_1 = \int_{-\infty}^{+\infty} dx[u_x(x,t)]^2 \exp[2U(x,t)], \tag{4.16b}$$

$$C_2 = \int_{-\infty}^{+\infty} dx\{[u_{xx}(x,t)]^2 - 8[u_x(x,t)]^3 u(x,t) - 2[u_x(x,t)]^2[u(x,t)]^4\} \exp[2U(x,t)]. \tag{4.16c}$$

Note that the first of these formulas implies that the quantity

$$\tilde{C} = \int_{-\infty}^{+\infty} dx[u(x,t)]^2 = U(\infty,t) = U(\infty,0) \tag{4.17}$$

is a constant of the motion for the nonlinear evolution (1.1) [a finding that can also be read directly from (2.4b)]; while to obtain the last two formulas we have integrated by parts, to simplify the expression of the integrand.

We end this section displaying a convenient expression of the moments $X_{n,0}(t)$, see (4.4). It reads

$$X_{n+1,0}(t) = (n!)^{-1}\left[\int_0^{\infty} dx\, x^n\{\exp[U(\infty,t)] - \exp[U(x,t)]\} + \int_{-\infty}^{0} dx\, x^n\{1 - \exp[U(x,t)]\}\right], \quad n = 0,1,2,..., \tag{4.18}$$

and it follows, after one integration by parts, from (4.4) via (4.15).

V. EXPLICIT SOLUTIONS

In this section we exhibit and discuss some explicit solutions of the nonlinear evolution equation (1.1).

The standard technique to obtain such solutions is to start from some simple solution $v(x,t)$ of the linear equation (3.1) and to evaluate the corresponding solution $u(x,t)$ of (1.1), as given by (3.2) with (3.3) or by (3.7a).

Hereafter we focus on solutions $u(x,t)$ that are real and regular for all real x. This generally requires that the function $V(x,t)$, see (3.2) and (3.3), be positive definite for all real (finite) values of x. Hence solutions $v(x,t)$ of (3.1) that are polynomials in x [of which the simpler one is $v(x,t) = A_0 + A_1 x + A_2 x^2$ with A_0, A_1, and A_2 arbitrary constants] are excluded from consideration, since the corresponding $V(x,t)$ cannot be positive definite, being a polynomial in x of odd degree [see (3.3a)].

A. Solitron of the first kind

Let

$$v(x,t) = A\exp[p(x + p^2 t)], \quad p > 0, \quad A = A^*. \tag{5.1}$$

Then (3.7a) yields

$$u(x,t) = \mathrm{sgn}(A)p^{1/2}h[2p(x - x_0 - Vt)], \tag{5.2}$$

$$h(y) = [1 + 2\exp(-y)]^{-1/2}, \tag{5.3}$$

$$x_0 = (2p)^{-1}\ln[pC^2/(2A^2)], \tag{5.4}$$

$$V = -p^2. \tag{5.5}$$

Note that this kinklike solution depends [apart from the trivial parameter x_0, whose arbitrariness reflects the translation invariance of (1.1)] on the single (positive) parameter p, that characterizes both its shape and the (negative) speed with which it travels [see (5.5), and Appendix A]. For future reference, it is convenient to introduce the notation

$$u_x(x,t) = \mathrm{sgn}(A)S_1(x - x_0, t; p) = \mathrm{sgn}(A)p^{3/2}S[p(x - x_0 + p^2 t)], \tag{5.6}$$

so that

$$S(y) = 2[\exp(\tfrac{4}{3}y) + 2\exp(-\tfrac{2}{3}y)]^{-3/2}, \tag{5.7}$$

and to refer to this function as representing a "solitron of the first kind"; it is preferable in this context to focus on u_x rather than u, since the fact that u_x is localized while u is kinklike will prove advantageous to discuss solutions with several solitrons present, see below. A graph of the function $S(y)$ is given in Fig. 1.

B. Solitrons of the second and third kind

Let

$$v(x,t) = A_1\exp[p_1(x + p_1^2 t)] + A_2\exp[p_2(x + p_2^2 t)], \quad p_2 > p_1 > 0, \quad A_1 = A_1^* \neq 0, \quad A_2 = A_2^* \neq 0. \tag{5.8}$$

Then (3.7a) with $C = 0$ yields

$$u(x,t) = \mathrm{sgn}(A_2)p_2^{1/2}H_s[(p_2 - p_1)(x - \bar{x} - Vt); p_1/p_2], \tag{5.9}$$

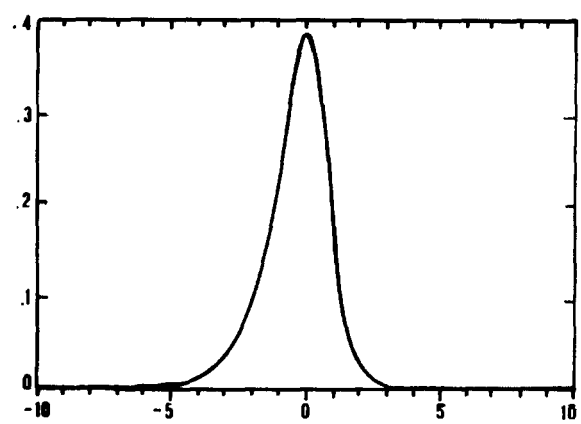


FIG. 1. Graph of the function $S(y)$, see (5.7), representing a solitron of first kind. Note that the profile is not symmetrical.

where (see Appendix A),

$$H_s(y;a) = [s + \exp(y)]/[a^{-1} + 4(1+a)^{-1} s \exp(y) + \exp(2y)]^{1/2}, \tag{5.10}$$

$$\bar{x} = (p_2 - p_1)^{-1} \ln|A_1/A_2|, \tag{5.11}$$

$$s = \mathrm{sgn}(A_1/A_2), \tag{5.12}$$

$$V = -(p_1^2 + p_2^2 + p_1 p_2). \tag{5.13}$$

Note that each of these kinklike solutions depends on two positive parameters, p_1 and p_2 (or, equivalently, on $p = p_2$ and $a = p_1/p_2$), in addition to the trivial constant $\bar{x}$ that accounts for the translation invariance of (1.1). For future reference, it is expedient to introduce the notation

$$u_x(x,t) = \mathrm{sgn}(A_2) S_s(x - \bar{x}, t; p_1, p_2), \tag{5.14}$$

so that (see Appendix A)

$$S_s(y,t;p_1,p_2) = (p_1 p_2)^{1/2}[(p_2 - p_1)^2/(p_2 + p_1)] \times [s p_1 Z^{1/2} + p_2/Z^{1/2}][p_1 Z + p_2/Z + 4s p_1 p_2/(p_1 + p_2)]^{-3/2} \tag{5.15a}$$

with

$$Z = \exp\{(p_2 - p_1)[y + (p_2^2 + p_1^2 + p_1 p_2)t]\}. \tag{5.15b}$$

We will refer to $S_+(y,t;p_1,p_2)$ as representing a "solitron of second kind" and to $S_-(y,t;p_1,p_2)$ as representing a "solitron of third kind"; we display these functions in Figs. 2 and 3, having set for this purpose

$$S_s(y,t;p_1,p_2) = p_2^{3/2} F_s(x,a), \tag{5.16a}$$

with

$$x = p_2[y + (p_2^2 + p_1^2 + p_1 p_2)t], \tag{5.16b}$$

$$a = p_1/p_2, \quad 0 < a < 1, \tag{5.16c}$$

$$F_s(x,a) = a^{1/2}[(1-a)^2/(1+a)](s + aZ) \times \{Z^{-2/3} + aZ^{4/3} + [4sa/(1+a)]Z^{1/3}\}^{-3/2}, \tag{5.16d}$$

$$Z = \exp[(1-a)x]. \tag{5.16e}$$

For an analytic analysis of the behavior of the function $F_s(x,a)$, see (A24c) and the discussion preceding it.

Note that the treatment of Appendix A implies that there are no other solitrons beside the three types obtained so far (and of course the three corresponding "antisolitrons," that obtain by changing the overall sign of each solution). This conclusion is of course based on the convention to reserve the term "solitron" (or "antisolitron") for solutions of (1.1) that are real and regular for all (real) values of x, whose time evolution consists of a mere translation with constant speed [that turns always out to be negative; see (5.5) and (5.13)], and that are "localized" at least in the sense that $u_x(x,t)$ vanishes asymptotically ($x \to \pm\infty$).

Note finally that the relation (3.7b) is not applicable to the solution (5.9); indeed $u(-\infty,t)$ does not vanish [see (A23a)]. This is of course a feature of all solutions of (1.1) obtained from solutions of (3.1) via (3.7a) with $C = 0$ (see below).

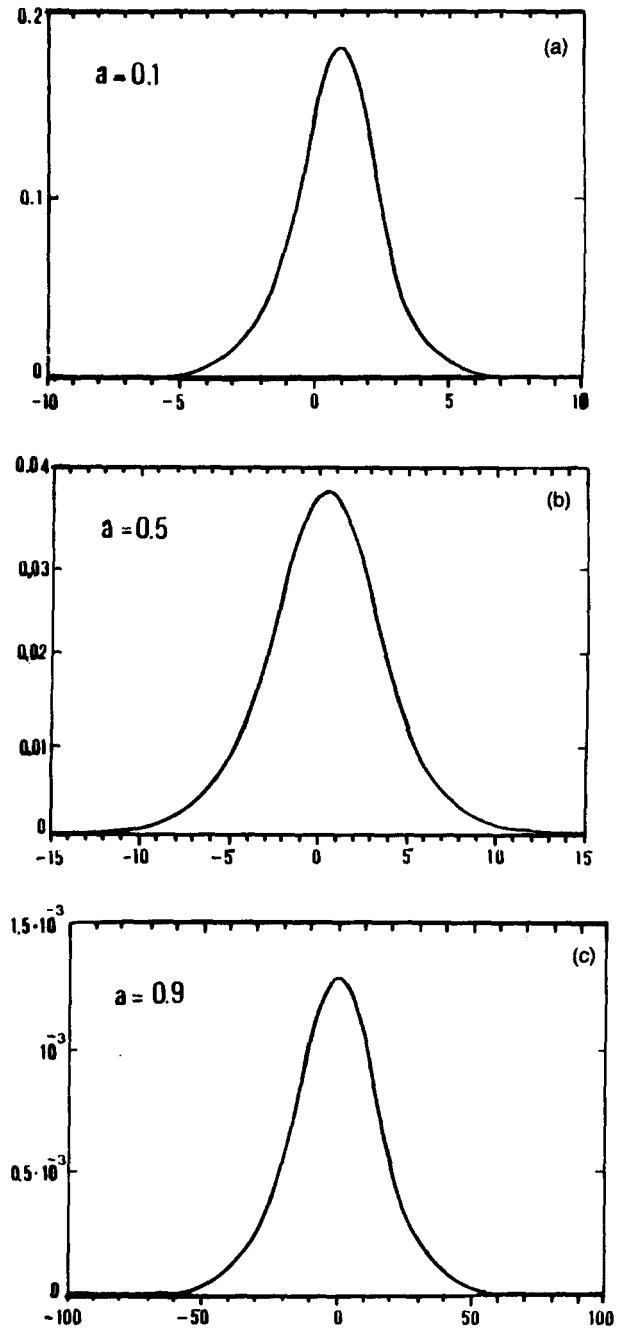


FIG. 2. Graphs of the function $F_+(x,a)$, see (5.16d), representing a solitron of the second kind.

C. Periodic traveling wave

Let

$$v(x) = A_1 \exp[p_1(x + p_1^2 t)] + A_2 \exp[p_2(x + p_2^2 t)] \tag{5.17a}$$

with

$$p_1 = r + iq, \quad p_2 = r - iq, \quad r > 0, \quad q > 0, \tag{5.17b}$$

$$A_1 = A \exp(ib), \quad A_2 = A \exp(-ib),$$

$$b = b^*, \quad A = A^* \neq 0. \tag{5.17c}$$

Then (3.7a) with $C = 0$ yields

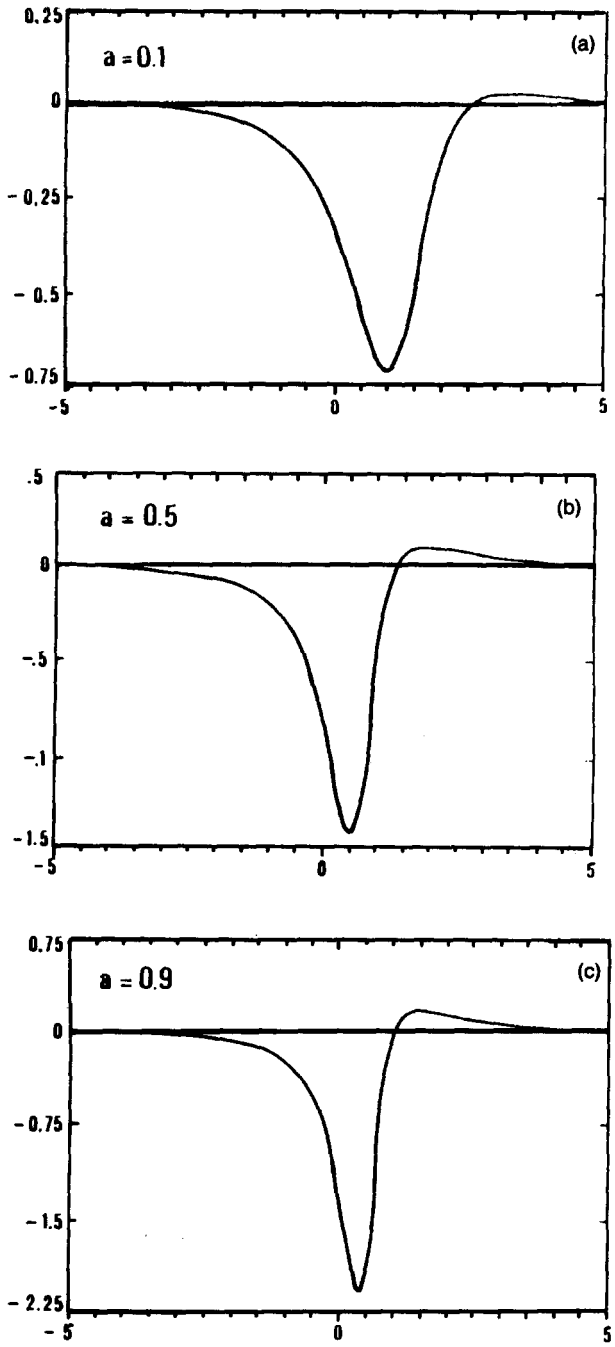


FIG. 3. Graphs of the function $F_{-}(x,a)$, see (5.16d), representing a solitron of the third kind.

$$u(x,t) = \mathrm{sgn}(A)(2r)^{1/2}\cos(y) \times [1 + \sin(a)\sin(2y + a)]^{-1/2}, \tag{5.18}$$

$$y = b + q(x - Vt), \tag{5.19}$$

$$V = q^2 - 3r^2, \tag{5.20}$$

$$\tan(a) = r/q. \tag{5.21}$$

Note that this periodic traveling wave depends on two parameters, r and q [in addition to b, that accounts trivially for the translation invariance of (1.1)]; and that in this case the speed with which it translates may be positive or negative, or it may vanish [see (5.20)].

For future reference, let us introduce the notation

$$T(x,t;r,q;b) = (2r)^{1/2}\cos(y)/[1 + \sin(a)\sin(2y + a)]^{1/2}, \tag{5.22}$$

with y and a defined by (5.19)–(5.21), as the function representing a periodic traveling wave. Note that here, in contrast to the case of the solitrons treated in the preceding two subsections (V A and V B), the notation refers directly to the solution $u(x,t)$ of the evolution PDE (1.1), rather than its x derivative. To display its shape, we also introduce the function $F(x,a)$,

$$F(x,a) = \cos(x)/[1 + \sin(a)\sin(2x + a)]^{1/2}, \tag{5.23}$$

and exhibit some graphs of it in Fig. 4. Note that, for all values of the parameter a (in the range $0 \leqslant a < \pi/2$), the periodic function $F(x,a)$ oscillates between the values -1 and $+1$ (see the end of Appendix A).

The results described so far have merely reproduced the findings reported in Appendix A; note that the treatment given there implies that no other real and regular solution exists, besides those described above, whose time evolution reduces merely to a translation with constant speed (without change of shape).

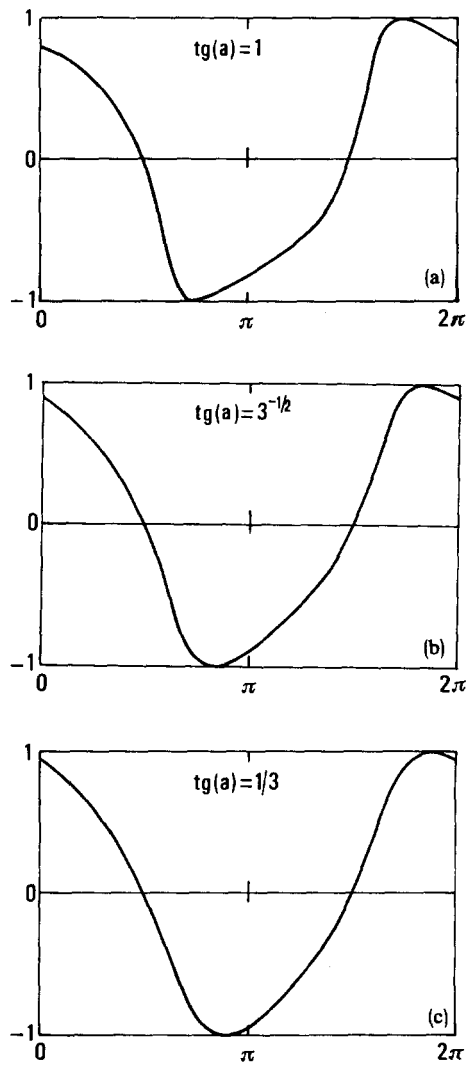


FIG. 4. Graphs of the function $F(x,a)$ see (5.23): (a) $\tan(a) = 1$, ($q = r$, $V = -2r^2 < 0$); see (5.20); (b) $\tan(a) = 3^{-1/2}$ ($q^2 = 3r^2$, $V = 0$); (c) $\tan(a) = \frac{1}{3}$ ($q = 3r$, $V = 6r^2 > 0$).

Let us now consider some other solutions, that are more complex but are nevertheless susceptible of explicit display. We exhibit and discuss firstly some simpler cases, and we consider subsequently more general instances.

D. Semi-infinite traveling wave

Let $v(x,t)$ be again given by (5.17a) with (5.17b) and (5.17c) and use (3.7a), but now with $C^2>0$. We obtain

$$u(x,t)=\mathrm{sgn}(A)(2r)^{1/2}\cos(y)\times[1+\sin(a)\sin(2y+a)+\exp(-2z)]^{-1/2}, \tag{5.24}$$

with [see (5.19 – 5.21)]

$$y=b+q(x-Vt), \tag{5.25}$$

$$V=q^2-3r^2, \tag{5.26}$$

$$\tan(a)=r/q, \tag{5.27}$$

and

$$z=r(x-\bar{x}-Wt), \tag{5.28}$$

$$W=3q^2-r^2, \tag{5.29}$$

$$\bar{x}=(2r)^{-1}\ln(rC^2/A^2). \tag{5.30}$$

Clearly $u(x,t)$ is exponentially small for $x\ll\bar{x}+Wt$ and it reduces to a periodic traveling wave (see Sec. V C) for $x\gg\bar{x}+Wt$. Note that the two speeds V and W can have any sign (or one of them could vanish); on the other hand the difference $W-V$ is positive,

$$W-V=2(q^2+r^2)>0. \tag{5.31}$$

E. Inelastic collision of a solitron of the first kind with one of the second or third kind

Let $v(x,t)$ be again given by (5.8) and use (3.7a), but now with $C^2>0$. We obtain

$$u(x,t)=\mathrm{sgn}(A_2)p_2^{1/2}[s+\exp(y_2)]\times\{a^{-1}+4(1+a)^{-1}s\exp(y_2)+\exp(2y_2)+\exp[-2(y_1-\bar{y}_1)]\}^{-1/2}, \tag{5.32}$$

$$y_2=(p_2-p_1)(x-\bar{x}-Wt), \tag{5.33}$$

$$\bar{x}=(p_2-p_1)^{-1}\ln|A_1/A_2|, \tag{5.34}$$

$$W=-(p_1^2+p_2^2+p_1p_2), \tag{5.35}$$

$$y_1=p_1(x-Vt), \tag{5.36}$$

$$V=-p_1^2, \tag{5.37}$$

$$\bar{y}_1=\tfrac{1}{2}\ln(p_2C^2/A_1^2), \tag{5.38}$$

$$s=\mathrm{sgn}(A_1/A_2). \tag{5.39}$$

To interpret this solution of (1.1) it is expedient to investigate the behavior of $u_x(x,t)$,

$$u_x(x,t)=\mathrm{sgn}(A_2)(p_1p_2)^{1/2}\times\{[(p_1-p_2)^2/(p_1+p_2)](sp_1Z_2^{1/2}+p_2Z_2^{-1/2})+Z_1(p_2Z_2^{1/2}+sp_1Z_2^{-1/2})\}\times[p_1Z_2+p_2/Z_2+4sp_1p_2/(p_1+p_2)+Z_1]^{-3/2}, \tag{5.40a}$$

$$Z_2=|A_2/A_1|\exp\{(p_2-p_1)[x+(p_1^2+p_2^2+p_1p_2)t]\}, \tag{5.40b}$$

$$Z_1=p_1p_2(C^2/|A_1A_2|)\exp\{-(p_1+p_2)\times[x+(p_1^2+p_2^2-p_1p_2)t]\}, \tag{5.40c}$$

in the remote past ($t\to-\infty$) and future ($t\to+\infty$). To do this we set, in (5.40a), $x=x'+V't$, we consider the limits as $t\to-\infty$ and as $t\to+\infty$ with x' and V' fixed, and we write the nonvanishing contributions that obtain for all (appropriately chosen) V'. In this manner we find, as $t\to-\infty$,

$$u_x(x,t)\simeq\mathrm{sgn}(A_1)S_1(x-x_1;p_1)+\mathrm{sgn}(A_2)S_s(x-\bar{x};p_1,p_2) \tag{5.41a}$$

with

$$x_1=(2p_1)^{-1}\ln[p_1C^2/(2A_1^2)], \tag{5.41b}$$

$$\bar{x}=(p_2-p_1)^{-1}\ln|A_1/A_2|, \tag{5.41c}$$

while for $t\to+\infty$ we find

$$u_x(x,t)\simeq\mathrm{sgn}(A_2)S_1(x-x_2;p_2) \tag{5.42a}$$

with

$$x_2=(2p_2)^{-1}\ln[p_2C^2/(2A_2^2)]. \tag{5.42b}$$

In these formulas the two functions S_1 and S_s are defined of course by (5.6) with (5.7) and by (5.15a) with (5.15b).

The interpretation of these findings is clear. The solution (5.32) describes in the remote past a solitron (if $A_1>0$) or an antisolitron (if $A_1<0$) of the first kind and parameter p_1, localized at $x\simeq x_1-p_1^2t$, and a solitron (if $A_2>0$) or antisolitron (if $A_2<0$) of the second kind (if $A_1/A_2>0$) or the third kind (if $A_1/A_2<0$) and parameters p_1 and p_2 (with $p_2>p_1$), localized at $x\simeq\bar{x}-(p_1^2+p_2^2+p_1p_2)t$. Both objects move of course towards the left; the solitron or antisolitron of the second or third kind moves faster (indeed, more than three times faster) and is therefore, in the remote past, farther to the right. As time goes by, the faster solitron or antisolitron of the second or third kind approaches the slower solitron or antisolitron of the first kind, and eventually the two coalesce into a single solitron or antisolitron of first kind and parameter p_2, that emerges alone in the remote future, moving with its characteristic speed that is intermediate between those of the two initial objects, since clearly $p_1^2<p_2^2<p_1^2+p_2^2+p_1p_2$.

Note that (5.41b), (5.41c), and (5.42b) imply the following relation between the parameters x_1, $\bar{x}$ and x_2 that characterize the asymptotic location of these objects [see (5.41a) and (5.42a)]:

$$p_2x_2=p_1x_1+(p_2-p_1)\bar{x}+\tfrac{1}{2}\ln(p_2/p_1). \tag{5.43}$$

Let us also point out that the solution discussed in this subsection does not describe the most general collision between a solitron (or an antisolitron) of the first kind and one of the second or third kind, but only one between a (generic) solitron (or antisolitron) of the first kind with parameter p and a solitron (or antisolitron) of the second or third kind characterized by parameters p_1 and p_2, with $p_2>p_1$ and $p_1=p$.

F. Inelastic collision of two solitrons of the second or third kind

Let

$$v(x,t) = \sum_{n=1}^{3} [A_n \exp(y_n)], \tag{5.44}$$

with the three constants A_n real (and nonvanishing), the three parameters p_n positive and different, say

$$0 < p_1 < p_2 < p_3, \tag{5.45}$$

and

$$y_n = p_n(x + p_n^2 t). \tag{5.46}$$

Then use (3.7a) with $C = 0$. We obtain

$$u(x,t) = \sum_{n=1}^{3} \frac{[A_n \exp(y_n)]}{[g(x,t)]^{1/2}}, \tag{5.47a}$$

$$g(x,t) = 2 \sum_{n=1}^{3} \sum_{m=1}^{3} \left\{ \frac{A_n A_m}{(p_n + p_m)} \exp(y_n + y_m) \right\}. \tag{5.47b}$$

To analyze the significance of this solution of the evolution PDE (1.1), it is again convenient to focus on the derivative $u_x(x,t)$ and to look at its behavior in the remote past and future. One finds, for $t \to -\infty$,

$$u_x(x,t) \approx \operatorname{sgn}(A_2) S_{s_1}(x - \bar{x}_1, t; p_1, p_2) + \operatorname{sgn}(A_3) S_{s_2}(x - \bar{x}_2, t; p_2, p_3), \tag{5.48a}$$

$$s_n = \operatorname{sgn}(A_n / A_{n+1}), \quad n = 1,2, \tag{5.48b}$$

$$\bar{x}_n = (p_{n+1} - p_n)^{-1} \ln|A_n / A_{n+1}|, \quad n = 1,2, \tag{5.48c}$$

and for $t \to +\infty$

$$u_x(x,t) \approx \operatorname{sgn}(A_3) S_s(x - \bar{x}, t; p_1, p_3), \tag{5.49a}$$

with

$$s = \operatorname{sgn}(A_1 / A_3), \tag{5.49b}$$

$$\bar{x} = (p_3 - p_1)^{-1} \ln|A_1 / A_3|. \tag{5.49c}$$

Here of course the function $S_s(y,t;p,p')$ is defined by (5.15).

The interpretation of these findings is clear. The solution (5.44) describes in the remote past two solitrons or antisolitrons [as the case may be; see (5.48a)] of the second or third kind [as the case may be; see (5.48b)], and in the remote future a single solitron or antisolitron of the second or third kind [as the case may be; the diligent reader may figure out the "selection rules" implied by (5.48a), (5.48b), and (5.49a)]. Note that (5.45) implies that the speed, $V_3 = -(p_3^2 + p_1^2 + p_1 p_3)$, of the final object is intermediate between the speeds, $V_1 = -(p_2^2 + p_1^2 + p_2 p_1)$ and $V_2 = -(p_3^2 + p_2^2 + p_2 p_3)$, of the two initial ones; while the parameters that characterize their asymptotic positions are related by the formula

$$(p_3 - p_1)\bar{x} = (p_3 - p_2)\bar{x}_2 + (p_2 - p_1)\bar{x}_1. \tag{5.50}$$

Let us, however, again emphasize that the solution (5.44) does not describe the most general collision between two solitrons or antisolitrons of second or third kind, but only a collision among two such solitrons or antisolitrons characterized by two pairs of parameters, say p_1, p_2 (with $p_2 > p_1$) and p_1', p_2' (with $p_2' > p_1'$), such that, say, $p_2 = p_1'$.

G. Inelastic collision of N solitrons of the second and third kind

Let

$$v(x,t) = \sum_{n=0}^{N} [A_n \exp(y_n)] \tag{5.51}$$

with the $N+1$ parameters A_n real and nonvanishing, the $N+1$ parameters p_n positive and different, say

$$0 < p_0 < p_1 < \cdots < p_N \tag{5.52}$$

and

$$y_n = p_n(x_n + p_n^2 t). \tag{5.53}$$

Then use (3.7a) with $C = 0$. We obtain

$$u(x,t) = \sum_{n=0}^{N} \frac{A_n \exp(y_n)}{[g(x,t)]^{1/2}}, \tag{5.54a}$$

$$g(x,t) = 2 \sum_{n=0}^{N} \sum_{m=0}^{N} \left\{ \frac{A_n A_m}{p_n + p_m} \exp(y_n + y_m) \right\}. \tag{5.54b}$$

The significance of this solution of the evolution equation (1.1) is apparent from the following findings (proved in Appendix D): as $t \to -\infty$,

$$u_x(x,t) \approx \sum_{n=1}^{N} [\operatorname{sgn}(A_n) S_{s_n}(x - \bar{x}_n, t; p_{n-1}, p_n)], \tag{5.55a}$$

with

$$\bar{x}_n = (p_n - p_{n-1})^{-1} \ln|A_{n-1} / A_n|, \tag{5.55b}$$

$$s_n = \operatorname{sgn}(A_{n-1} / A_n); \tag{5.55c}$$

as $t \to +\infty$,

$$u_x(x,t) \approx \operatorname{sgn}(A_N) S_s(x - \bar{x}, t; p_0, p_N), \tag{5.56a}$$

with

$$\bar{x} = (p_N - p_0)^{-1} \ln|A_0 / A_N|, \tag{5.56b}$$

$$s = \operatorname{sgn}(A_0 / A_N). \tag{5.56c}$$

Note that (5.55b) and (5.56b) yield

$$(p_N - p_0)\bar{x} = \sum_{n=1}^{N} [(p_n - p_{n-1})\bar{x}_n]. \tag{5.57}$$

These findings include of course those of the preceding subsection, to which we also refer for their interpretation; that should be sufficiently obvious not to warrant any additional comment here.

H. Inelastic collision of one solitron of the first kind and N solitrons of the second or third kind

Let $v(x,t)$ be again given by (5.51) with (5.52) and (5.53), but now use (3.7a) with $C^2 > 0$. We obtain

$$u(x,t) = \sum_{n=0}^{N} \frac{A_n \exp(y_n)}{[C^2 + g(x,t)]^{1/2}}, \tag{5.58}$$

with $g(x,t)$ defined by (5.54b).

The significance of this solution of the evolution equation (1.1) is apparent from the following results (proved in Appendix D): as $t \to -\infty$,

$$u_x(x,t) \approx \operatorname{sgn}(A_0) S_1(x - x_0, t; p_0) + \sum_{n=1}^{N} [\operatorname{sgn}(A_n) S_{s_n}(x - \bar{x}_n, t; p_{n-1}, p_n)], \tag{5.59a}$$

with $\bar{x}_n$ and s_n defined by (5.55b) and (5.55c) and

$$x_0 = (2p_0)^{-1} \ln[p_0 C^2/(2A_0^2)]; \tag{5.59b}$$

as $t \to +\infty$,

$$u_x(x,t) \approx \operatorname{sgn}(A_N) S_1(x - x_N, t; p_N), \tag{5.60a}$$

with

$$x_N = (2p_N)^{-1} \ln[p_N C^2/(2A_N^2)]. \tag{5.60b}$$

Note that these results imply the relation

$$p_N x_N = p_0 x_0 + \sum_{n=1}^{N} [(p_n - p_{n-1})\bar{x}_n] + \frac{1}{2} \ln\left(\frac{p_N}{p_0}\right). \tag{5.61}$$

These findings include those of Sec. V E, to which we also refer for their interpretation, which should be sufficiently obvious not to warrant any additional comment here.

I. Traveling wave and kink

Let $v(x,t)$ be given again by (5.44) with (5.46), but assume now

$$p_1 = p > 0, \quad p_2 = r + iq, \quad p_3 = r - iq, \quad r > 0, \quad q > 0, \tag{5.62}$$

$$A_1 = A_1^* = A \neq 0, \quad A_2 = B \exp(ib), \quad A_3 = B \exp(-ib), \quad B = B^* \neq 0, \quad b = b^*. \tag{5.63}$$

Then use (3.7a) with $C = 0$. We obtain

$$u(x,t) = \operatorname{sgn}(A) p^{1/2} [1 + sZ \cos(y)]/[f(x,t)]^{1/2}, \tag{5.64a}$$

$$f(x,t) = 1 + 4s(p/Q)Z \sin(y + a') + \tfrac{1}{2}(p/r)Z^2[1 + \sin(a)\sin(2y + a)], \tag{5.64b}$$

$$Z = 2|B/A| \exp[(r-p)(x - Wt)] = \exp[(r-p)(x - x_0 - Wt)], \tag{5.65}$$

$$x_0 = (p - r)^{-1} \ln|2B/A|, \tag{5.66}$$

$$W = -[p^2 + r^2 + pr + 3rq^2/(p - r)], \tag{5.67}$$

$$y = q(x - Vt) + b, \tag{5.68}$$

$$V = q^2 - 3r^2, \tag{5.69}$$

$$\tan(a') = (p + r)/q, \tag{5.70}$$

$$\tan(a) = r/q, \tag{5.71}$$

$$Q = [(p + r)^2 + q^2]^{1/2}, \tag{5.72}$$

$$s = \operatorname{sgn}(B/A). \tag{5.73}$$

If $p > r$, clearly for $x \gg x_0 + Wt$ the solution $u(x,t)$ is constant,

$$u(x,t) \approx \operatorname{sgn}(A) p^{1/2}, \tag{5.74}$$

while for $x \ll x_0 + Wt$ it approximates the periodic traveling wave of Sec. V C,

$$u(x,t) \approx \operatorname{sgn}(B) T(x - x_0, t; r, q; b) \tag{5.75}$$

[see (5.22)]. Note that in this case the speed W, with which moves the boundary layer between the two zones, is negative [see (5.67)].

If $p < r$, the situation is reversed, namely for $x \ll x_0 + Wt$ the solution $u(x,t)$ is constant, see (5.74), while for $x \gg x_0 + Wt$ it approximates the periodic traveling wave, see (5.75). Note that in this case, in contrast to the preceding one, the speed W, with which moves the boundary between the two zones, may have either sign, or it may vanish [see (5.67)].

Finally, in the marginal case $p = r$, $u(x,t)$ is a periodic function of x (with period $2\pi/q$), since in this special case the quantity Z, see (5.65) and (5.67), becomes independent of x

$$Z = 2|B/A| \exp(-3pq^2 t). \tag{5.76}$$

Hence this case provides another instance of periodic solution of the evolution PDE (1.1); but it has a more complicated time dependence than the periodic traveling wave of Sec. V C. Note, however, that, as $t \to -\infty$, this solution goes indeed over into the traveling wave solution of Sec. V C [see (5.64) and (5.76)],

$$u(x,t) \approx \operatorname{sgn}(B) T(x - x_0, t; r, q; b), \tag{5.77}$$

while as $t \to +\infty$ it becomes constant,

$$u(x,t) \approx \operatorname{sgn}(A) p^{1/2}. \tag{5.78}$$

J. Traveling wave that becomes a solitron of the first kind

Let $v(x,t)$ be again given by (5.44) with (5.46), (5.62), and (5.63), and use (3.7a), but now with $C^2 > 0$, to evaluate $u(x,t)$. We obtain

$$u(x,t) = \operatorname{sgn}(A) p^{1/2} [1 + sZ \cos(y)]/[f(x,t) + Z_1^2]^{1/2}, \tag{5.79}$$

with

$$Z_1 = \exp[-p(x - x_1 - V_1 t)], \tag{5.80}$$

$$x_1 = (2p)^{-1} \ln(pC^2/A^2), \tag{5.81}$$

$$V_1 = -p^2, \tag{5.82}$$

and the remaining notation as in the preceding subsection, see (5.64b)–(5.73).

To analyze the shape, at any given (fixed) time, of this solution of the evolution PDE (1.1), let us consider first the case

$$p > r > 0. \tag{5.83}$$

It is then easily seen that

$$u(-\infty, t) = 0, \tag{5.84a}$$

$$u(+\infty, t) = \operatorname{sgn}(A) p^{1/2}, \tag{5.84b}$$

implying

$$u_x(\pm\infty, t) = 0. \tag{5.84c}$$

It is also clear that $u_x(x,t)$ vanishes proportionally to $\exp(rx)$ (times an oscillatory factor) as $x \to -\infty$ and proportionally to $\exp[(r-p)x]$ (times another oscillatory factor) as $x \to +\infty$). Hence this solution is "localized," in the sense used above.

If instead

$$r \geqslant p > 0 \tag{5.85}$$

the solution $u(x,t)$, while still vanishing as $x \to -\infty$, see

(5.84a), approximates the solution discussed in Sec. V I as $x \to +\infty$; and in particular, if $r > p$, as $x \to +\infty$ it approximates the periodic traveling wave of Sec. V C,

$$u(x,t) \approx \mathrm{sgn}(B)\, T(x - x_0, t; r, q; b). \tag{5.85'}$$

Let us now discuss the behavior over time of the solution (5.79). The analysis here is limited to the "localized" case characterized by the inequality (5.83); the treatment of the case (5.85) is left as an exercise for the diligent reader.

Clearly the behavior of $u(x,t)$ depends on whether the (positive) quantities Z and Z_1, see (5.65) and (5.80), are much larger or much smaller than unity and, moreover, if they are much larger than unity, on their relative magnitude. Thus three speeds play an important role, namely W [see (5.67)], V_1 [see (5.82)], and

$$V_2 = 3q^2 - r^2. \tag{5.86}$$

Note that, as a consequence of (5.83), the inequality

$$W < V_1 < V_2 \tag{5.87}$$

holds, since both differences,

$$V_2 - V_1 = 3q^2 + p^2 - r^2 \tag{5.88a}$$

and

$$V_1 - W = r^2 + pr + 3rq^2/(p - r), \tag{5.88b}$$

are clearly positive. Note moreover that W and V_1 are negative, while V_2 may have any sign, or it might even vanish.

It is then easily seen that, as $t \to -\infty$,

$$u(x,t) \approx 0, \quad x \ll X_1(t), \tag{5.89a}$$

$$u(x,t) \approx \mathrm{sgn}(B)\, T(x - x_0, t; r, q; b),$$

$$X_1(t) \ll x \ll X_2(t), \tag{5.89b}$$

$$u(x,t) \approx \mathrm{sgn}(A) p^{1/2}, \quad X_2(t) \ll x, \tag{5.89c}$$

where

$$X_1(t) = x_0 + (p/r)(x_1 - x_0) + V_2 t, \tag{5.90a}$$

$$X_2(t) = x_0 + Wt, \tag{5.90b}$$

and of course x_0 and x_1 are defined by (5.67) and (5.81). Note that these "boundary layers," $X_1(t)$ and $X_2(t)$, move with constant speed, and that, as $t \to -\infty$, $X_2(t) \to +\infty$ and $X_2(t) - X_1(t) \to +\infty$. Thus, as $t \to -\infty$, the region occupied by the periodic traveling wave [see (5.89b) and (5.22)] becomes infinitely extended. Note however that, as $t \to -\infty$, $X_1(t)$ may diverge to positive or to negative infinity, or remain constant, depending on the value of V_2.

The behavior of the solution $u(x,t)$, or rather its x derivative, as $t \to +\infty$ [in the case (5.83) to which our attention is confined] is instead very simple:

$$u_x(x,t) \approx \mathrm{sgn}(A) S_1(x - x_1, t; p), \tag{5.91}$$

with S_1 defined by (5.6) with (5.7), and x_1 defined by (5.81).

These findings justify the title of this subsection.

K. Inelastic collision of several solitrons and wave trains

Let us finally consider a solution $u(x,t)$ of the evolution equation (1.1) that includes all those considered above (in this section). It obtains via (3.7a) from the following solution $v(x,t)$ of (3.1):

$$v(x,t) = \sum_{n=0}^{N} [A_n \exp(p_n x + p_n^3 t)] + \sum_{m=1}^{M} \{B_m \exp[ib_m + (r_m + iq_m)x + (r_m + iq_m)^3 t] + \text{c.c.}\} \tag{5.92a}$$

or equivalently

$$v(x,t) = \sum_{n=0}^{N} [A_n \exp(y_n)] + 2 \sum_{m=1}^{M} [B_m \exp(z_m)\cos(w_m)], \tag{5.92b}$$

with

$$y_n = p_n(x + p_n^2 t), \tag{5.93}$$

$$z_m = r_m[x + (r_m^2 - 3q_m^2)t], \tag{5.94}$$

$$w_m = b_m + q_m[x + (3r_m^2 - q_m^2)t]. \tag{5.95}$$

We assume of course that the quantities A_n are real and nonvanishing, that the quantities b_m are non-negative, and that the quantities B_m, p_n, r_m, and q_m are positive. We moreover assume, without loss of generality, that the inequalities

$$0 < p_0 < p_1 < \cdots < p_N, \tag{5.96a}$$

$$0 < r_1 < r_2 \cdots < r_m \tag{5.96b}$$

hold.

The corresponding expression of $u(x,t)$ reads

$$u(x,t) = v(x,t)/[g(x,t)]^{1/2}, \tag{5.97}$$

$$g(x,t) = C^2 + \sum_{n=0}^{N} \left[\frac{A_n^2}{p_n} \exp(2y_n)\right] + 2 \sum_{m=1}^{M} \left\{\frac{B_m^2}{r_m} \exp(2z_m)[1 + \sin(a_m)\sin(2w_m + a_m)]\right\}$$
$$+ 2 \sum_{\substack{n=0 \\ n \neq n'}}^{N} \sum_{n'=0}^{N} \left\{\frac{A_n A_{n'}}{p_n + p_{n'}} \exp(y_n + y_{n'})\right\} + 4 \sum_{n=0}^{N} \sum_{m=1}^{M} \left\{\frac{A_n B_m}{Q_{nm}} \exp(y_n + z_m)\sin(w_m + a'_{n,m})\right\}$$
$$+ 4 \sum_{\substack{m=1 \\ m \neq m'}}^{M} \sum_{m'=1}^{M} \left\{\frac{B_m B_{m'}}{r_m + r_{m'}} \exp(z_m + z_{m'})\sin(a_{m,m'})\sin(w_m + w_{m'} + a_{m,m'})\right\}, \tag{5.98}$$

$$Q_{nm} = [(p_n + r_m)^2 + q_m^2]^{1/2}, \tag{5.99}$$

$$\tan(a_m) = r_m/q_m, \tag{5.100}$$

$$\tan(a'_{n,m}) = (p_n + r_m)/q_m, \tag{5.101}$$

$$\tan(a_{m,m'}) = (r_m + r_{m'})/(q_m + q_{m'}). \tag{5.102}$$

Let us discuss first of all the shape of this solution, namely its x profile for fixed (finite) t. In particular let us note that the two conditions

$$p_N > r_N, \tag{5.103}$$

$$p_0 < r_1 \quad \text{if } C = 0, \tag{5.104}$$

are necessary and sufficient in order that $u(x,t)$ be "localized," namely

$$u_x(\pm\infty,t) = 0. \tag{5.105}$$

Indeed (5.103) implies

$$u(+\infty,t) = \mathrm{sgn}(A_N)p_N^{1/2}, \tag{5.106}$$

while clearly, if $C^2 > 0$,

$$u(-\infty,t) = 0, \tag{5.107a}$$

and if $C = 0$ but (5.104) holds,

$$u(-\infty,t) = \mathrm{sgn}(A_1)p_1^{1/2}. \tag{5.107b}$$

Let us then discuss tersely the behavior of $u(x,t)$ over time, limiting our consideration to the "localized" case characterized by the conditions (5.103) and (5.104) [together of course with (5.106)]. Here we only outline the results since their detailed derivation and analysis would take too much space, and it is in any case somewhat analogous to the discussion in Appendix D.

In the remote future ($t \to +\infty$), the solution (5.97) becomes quite simple: if $C^2 > 0$, it describes a kink of the first type,

$$u_x(x,t) \approx \mathrm{sgn}(A_N)S_1(x - x_N, t; p_N), \tag{5.108}$$

$$x_N = (2p_N)^{-1}\ln[p_N C^2/(2A_N^2)]; \tag{5.109}$$

if $C = 0$, it describes a kink of the second or third type,

$$u_x(x,t) \approx \mathrm{sgn}(A_N)S_s(x - \bar{x}, t; p_0, p_N), \tag{5.110}$$

$$\bar{x} = (p_N - p_0)^{-1}\ln|A_0/A_N|, \tag{5.111}$$

$$s = \mathrm{sgn}(A_0/A_N). \tag{5.112}$$

In the remote past ($t \to -\infty$), the situation may be considerably more complicated; to describe it, let us consider separately the two cases, $C = 0$, and $C^2 > 0$.

Let us deal first with the $C = 0$ case. It is expedient, given the values of the $1 + N + 2M$ parameters p_n, r_m, and q_m, to draw as a function of V the $N + 1$ straight lines $p_n^3 + Vp_n$ (in black) and the M straight lines $r_m^3 - 3r_m q_m^2 + Vr_m$ (in red). Then focus attention on the segmented continuous line that obtains by following the *bottom* segments for each value of V. This line may have some black and some red segments; the leftmost and rightmost semi-infinite components are black, due to (5.103) and (5.104). Now move along this line from left to right and denote with W_1, W_2, etc. the values (if any) of V at which there is a change of slope from a black segment to another black segment; it is easily seen (as in Appendix D) that a necessary condition for this to happen is that the parameters, say $p_{n'_j}$ and p_{n_j}, that characterize the two contiguous black segments, be themselves contiguous [see (5.96a)], namely $n'_j = n_j - 1$, implying

$$W_j = -(p_{n_j}^2 + p_{n_j-1}^2 + p_{n_j}p_{n_j-1}), \quad j = 1,2,\ldots. \tag{5.113}$$

Note that the number of W_j's may vary between 0 and N, and of course by definition $W_j < W_{j+1} < 0$. Denote moreover with $W_1^{(-)}$, $W_2^{(-)}$, etc. the values (if any) of V at which there occurs a change from a black to a red segment, so that

$$W_j^{(-)} = -\left[p_{n_j^{(-)}}^2 + r_{m_j^{(-)}}^2 + p_{n_j^{(-)}}r_{m_j^{(-)}} + 3r_{m_j^{(-)}}q_{m_j^{(-)}}^2/(p_{n_j^{(-)}} - r_{m_j^{(-)}})\right], \quad j = 1,2,\ldots, \tag{5.114a}$$

where $p_{n_j^{(-)}}$ is the parameter that characterizes the black segment on the left and $r_{m_j^{(-)}}$, $q_{m_j^{(-)}}$ are the parameters that characterize the red segment on the right; and denote with $W_1^{(+)}$, $W_2^{(+)}$, etc. the values (if any) of V at which there occurs a change from a red to a black segment, so that

$$W_j^{(+)} = -\left[p_{n_j^{(+)}}^2 + r_{m_j^{(+)}}^2 + 3r_{m_j^{(+)}}q_{m_j^{(+)}}^2/(p_{n_j^{(+)}} - r_{m_j^{(+)}})\right], \quad j = 1,2,\ldots, \tag{5.114b}$$

where $p_{n_j^{(+)}}$ is now the parameter that characterizes the black segment on the right and $r_{m_j^{(+)}}$, $q_{m_j^{(+)}}$ are the parameters that characterize the red segment on the left. Note that there are as many $W_j^{(+)}$'s as $W_j^{(-)}$'s (possibly none), and that the inequalities

$$W_j^{(-)} < W_j^{(+)} < W_{j+1}^{(-)} < W_{j+1}^{(+)}, \quad j = 1,2,\ldots, \tag{5.114c}$$

hold. Finally let us also denote, for completeness, with $\tilde{W}_1$, $\tilde{W}_2$, etc. the values of V (if any) at which there is a change of slope from a red segment to another red segment. Note that we are, for simplicity, assuming that there occurs no "multiple point" at which more than two of the original straight lines cross simultaneously.

The behavior of $u(x,t)$ in the remote past may then be characterized as follows. Let

$$x_j(t) = \bar{x}_j + W_j t, \quad j = 1,2,\ldots, \tag{5.115}$$

with

$$\bar{x}_j = (p_{n_j} - p_{n_j-1})^{-1}\ln|A_{n_j-1}/A_{n_j}|, \quad j = 1,2,\ldots, \tag{5.116}$$

where p_{n_j-1} and p_{n_j} characterize the two black segments to the left and right of $V = W_j$ [see (5.113)]. Let moreover

$$x_j^{(\pm)}(t) = x_j^{(\pm)}(0) + W_j^{(\pm)}t, \quad j = 1,2,\ldots, \tag{5.117}$$

where

$$x_j^{(\pm)}(0) = \left[2(p_{n_j^{(\pm)}} - r_{m_j^{(\pm)}})\right]^{-1} \times \ln\left[2p_{n_j^{(\pm)}}B_{m_j^{(\pm)}}^2/(r_{m_j^{(\pm)}}A_{n_j^{(\pm)}}^2)\right], \quad j = 1,2,\ldots, \tag{5.118}$$

again with the parameters $p_{n_j^{(\pm)}}$ and $r_{m_j^{(\pm)}}$ (as well as $A_{n_j^{(\pm)}}$ and $B_{m_j^{(\pm)}}$) associated, respectively, to the black and red segments that join at $W_j^{(\pm)}$ [see (5.114a) and (5.114b)]. Then for the values of x that are well inside the intervals from

$x_j^{(-)}(t)$ to $x_j^{(+)}(t)$,

$$x_j^{(-)}(t) \ll x \ll x_j^{(+)}(t), \quad j = 1,2,..., \tag{5.119}$$

$u(x,t)$ is an oscillating wave train [possibly fairly complicated, especially if the interval $(W_j^{(-)}, W_j^{(+)})$ contains some $\tilde{W}_k$, or several equal values of r_m with different q_m's], while for the values of x that are well outside the intervals (5.119), $u(x,t)$ is constant [and therefore $u_x(x,t)$ vanishes], except in the neighborhood of the points $x_j(t)$, see (5.115), where it behaves as a kink of second or third kind, namely, away from the intervals (5.119),

$$u_x(x,t) = \sum_j [\operatorname{sgn}(A_{n_j}) S_{s_j}(x - \bar{x}_j, t; p_{n_j-1}, p_{n_j})], \tag{5.120}$$

where of course

$$s_j = \operatorname{sgn}(A_{n_j-1}/A_{n_j}). \tag{5.121}$$

It is thus seen that, in the remote past, $u(x,t)$ [see (5.97)] with $C = 0$ describes a collection of solitrons of the second or third kind (whose number may vary between 0 and N) and of separate finite wave trains (whose number may vary between 0 and M; of course there must be at least one solitron or one wave train); while, as we have seen above, in this case with $C = 0$ in the remote future it yields a single solitron of the second or third kind, see (5.110).

Let us finally discuss, quite tersely, the behavior of $u(x,t)$, see (5.97), with $C^2 > 0$, in the remote past $(t \to -\infty)$. The treatment given above remains applicable, with the addition of one more straight line to be drawn (in blue) along the V axis. Then the curve obtained from the union of the bottom lines has the rightmost semi-infinite component that is blue (and shields away part of the curve of the previous case). The previous analysis remains applicable to the part of the curve that has not been shielded away, and it may account for a number of solitrons of the second and third kind (possibly none) and of finite wave trains (possibly none). There remains to consider the contribution corresponding to the last part of the curve. There are two possibilities, depending whether the last finite segment, contiguous to the rightmost semi-infinite blue component, is red or black.

It is easily seen that, if

$$-p_0^2 > 3q_m^2 - r_m^2, \quad m = 1,2,...,M, \tag{5.122}$$

that segment is black. In this case, in addition to the contributions predicted by the preceding analysis, there is in the remote past a solitron of the first kind, namely for $x \approx x_0 - p^2 t$,

$$u_x(x,t) \approx \operatorname{sgn}(A_0) S_1(x - x_0, t; p_0), \tag{5.123}$$

with

$$x = (2p_0)^{-1} \ln[p_0 C^2/(2A^2)]. \tag{5.124}$$

If instead

$$-p_0^2 < \max_{m=1,M} (3q_m^2 - r_m^2) \equiv W^{(+)}, \tag{5.125}$$

then the rightmost (finite) segment is red; and (in contrast to the previous case) the largest of the $W_j^{(-)}$ does not now have a corresponding $W_j^{(+)}$. Let us indicate this largest $W_j^{(-)}$ (defined according to the procedure described above) as $W_J^{(-)}$. It is then easily seen that

$$W_J^{(-)} < W^{(+)} \tag{5.126}$$

and that, in addition to the contributions predicted by the previous analysis (ignoring $W_J^{(-)}$), there is in this case an additional wave train in the interval from $x_J^{(-)}(t)$ to $x^{(+)}(t)$, with $x_J^{(-)}(t)$ defined according to (5.117) and

$$x^{(+)}(t) = x^{(+)}(0) + W^{(+)}t, \tag{5.127}$$

$$x^{(+)}(0) = (2r)^{-1} \ln[rC^2/(2B^2)]. \tag{5.128}$$

The value of r appearing in the last formula coincides with the r_m which realizes the maximum in the rhs of (5.125).

It is thus seen that, in the remote past, $u(x,t)$ [see (5.97) with $C^2 > 0$] describes a collection of solitrons of second and third kind and of separate finite wave trains, and in addition, provided the inequality (5.122) holds, a single solitron of first kind; while, as we have seen above, in this case with $C^2 > 0$ in the remote future it yields a single solitron of first kind, see (5.108).

Let us end by noting that these findings suggest the following general result, applicable to any real and regular solution of the evolution PDE (1.1). Let $u(x,0)$ have finite limits as $x \to \pm\infty$, with the value at the right larger in modulus than that at the left; then, as $t \to +\infty$, $u(x,t)$ approximates a single kink. More precisely, if $u(-\infty,0) = s_1 p_1^{1/2}$, $u(+\infty,0) = s_2 p_2^{1/2}$ with $p_2 > p_1 > 0$ and $s_1 = +$ or $s_1 = -$ (likewise for s_2), then, as $t \to +\infty$,

$$u_x(x,t) \approx s_2 S_s(x - \bar{x}, t; p_1, p_2), \tag{5.129}$$

with $s = s_1 s_2$ and $\bar{x}$ some appropriate value; if $u(-\infty,0) = 0$ and $u(+\infty,0) = s p^{1/2}$, with $p > 0$ and $s = +$ or $s = -$, then, as $t \to +\infty$,

$$u_x(x,t) \approx s S_1(x - x_0, t; p), \tag{5.130}$$

for some appropriate value of x_0.

VI. FINAL COMMENTS

The results of the preceding section have displayed a remarkably explicit and complex phenomenology; of course the inelastic nature of the collisions among solitrons and antisolitrons motivates the use of this terminology (instead of "solitons" and "antisolitons"; see Ref. 3, p. 132ff).

Other explicit solutions of the evolution PDE (1.1) could be exhibited; but their display and analysis is left as an exercise for the diligent reader.

On the other hand, it should be emphasized that the solutions given above do not include the description of such elementary phenomena as the collision of two solitrons (or antisolitrons) of the first kind, or of one solitron (or antisolitron) of the first kind with a *generic* solitron (or antisolitron) of the second or third kind, or of two *generic* solitrons (or antisolitrons) of the second or third kind. The results given above suggest that such solutions do not exist. Let us note in this connection that, while the evolution character of the nonlinear PDE (1.1) implies the possibility to determine a solution $u(x,t)$ by assigning arbitrarily (within appropriate functional classes; see Sec. III) its "initial" value $u(x,t_0)$ at any chosen *finite* time t_0, this freedom of choice need not

apply without limitations in the asymptotic limit as t_0 tends, say, to negative infinity. A more detailed analysis of this problem, as well as the study of singular solutions of (1.1), will perhaps be presented in a subsequent paper.

It is well known that a large class of nonlinear evolution equations yield, after an appropriate multiscale asymptotic expansion (see, for instance, Ref. 4), the nonlinear Schrödinger equation. It is amusing to apply this procedure[4] to the nonlinear evolution equation (1.1). What one finds is that the method is indeed applicable, and one seems to get the nonlinear Schrödinger equation; but with a vanishing numerical coefficient in front of the nonlinear term! This is of course consistent with the need to use a more sophisticated method than just a change of variables (namely, the spectral transform technique) in order to solve the nonlinear Schrödinger equation.

Addendum

(i) The exceptional nature of the PDE (1.1) was previously discovered by Ibragimov and Shabat,[5] who pointed out that it belongs to the class of equations possessing an infinite Lie–Bäcklund algebra. Subsequently Kaptsov[6] noted that this equation possesses only one *local* conservation law. The linearizing transformation (3.7b) was moreover given by Sokolov and Shabat.[7]

These results have come to my attention after my paper was submitted for publication. I am not aware of any previous analysis of the detailed behavior of the solutions of the PDE (1.1).

(ii) Wiktor Eckhaus has noted that the remark at the end of Sec. VI implies the possibility of applying the limiting procedure one step further, obtaining thereby a novel nonlinear evolution equation in place of the Schrödinger equation. This remark has opened a line of research whose results serve to explain what had hitherto appeared a puzzling miracle, namely the fact that certain evolution equations turn up in many applicative contexts *and* are integrable. These findings shall be reported elsewhere.[8]

ACKNOWLEDGMENT

The research reported in this paper has been supported in part by funds provided by the Italian Ministry of Education.

APPENDIX A

In this appendix we obtain and discuss the general solution of the ODE

$$-Vg' = g''' + 3(g''g^2 + 3g'^2 g) + 3g'g^4, \tag{A1}$$

where V is a given constant.

A trivial solution of this equation is

$$g(y) = \text{arbitrary constant}. \tag{A2}$$

Hereafter this trivial solution will be ignored, as well as the trivial possibility to consider, in addition to any solution g, the solution $-g$.

The ODE (A1) can be directly integrated once, after multiplication by g. We obtain

$$-Vg^2 = 2g''g - g'^2 + 6g'g^3 + g^6 + 2B, \tag{A3}$$

where B is an arbitrary (integration) constant.

Now set

$$g(y) = f(y)/[2F(y)]^{1/2} \tag{A4}$$

with

$$F'(y) = f^2(y). \tag{A5}$$

We obtain

$$-Vf^2 = 2f''f - f'^2 + 2BF, \tag{A6}$$

and after differentiation [using (A5)] this yields the linear ODE

$$f''' + Vf' + Bf = 0. \tag{A7}$$

The general solution of this equation reads

$$f(y) = \sum_{j=1}^{3} [A_j \exp(p_j y)], \tag{A8}$$

where the three parameters p_j are the three roots of the cubic equation

$$p^3 + Vp + B = 0, \tag{A9}$$

so that they satisfy the following relations:

$$p_1 + p_2 + p_3 = 0, \tag{A10a}$$

$$p_1 p_2 + p_2 p_3 + p_3 p_1 = V, \tag{A10b}$$

$$p_1 p_2 p_3 = -B. \tag{A10c}$$

From (A5) and (A8) we moreover obtain

$$F(y) = A_0^2 + \sum_{j=1}^{3} \sum_{k=1}^{3} \left\{ \frac{A_j A_k}{p_j + p_k} \exp[(p_j + p_k)y] \right\}, \tag{A11a}$$

$$F(y) = A_0^2 - 2 \sum_{j=1}^{3} \left[\frac{A_{j+1} A_{j+2}}{p_j} \exp(-p_j y) \right] + \frac{1}{2} \sum_{j=1}^{3} \left[\frac{A_j^2}{p_j} \exp(2p_j y) \right]. \tag{A11b}$$

To write (A11b), we have used (A10a) and the cyclic convention $A_{j+3} \equiv A_j$. We are moreover assuming that none of the quantities p_j vanishes [a necessary and sufficient condition for this is that $B \neq 0$; see (A10c)]. If one or more of these quantities do vanish, the corresponding formula can be obtained by an appropriate limiting process (see below).

Insertion of this expression of $F(y)$, and of the corresponding expression (A8) of $f(y)$, in (A6), yields, using (A9), the condition

$$BA_0^2 = 0. \tag{A12}$$

There are therefore two distinct classes of solutions of (A1), those characterized by $B = 0$ and A_0 an arbitrary nonvanishing constant, and those characterized by $A_0 = 0$ and B an arbitrary constant.

In the first case ($B = 0, A_0 \neq 0$) one of the p_j's vanishes and the other two are easily computed, say

$$p_3 = 0, \quad p_1 = -p_2 = p, \tag{A13a}$$

$$p = (-V)^{1/2}. \tag{A13b}$$

Hence the solution of (A1) reads

$$g(y) = [B_1 \exp(py) + B_2 \exp(-py) + B_3]/[2F_1(y)]^{1/2} \tag{A14a}$$

with

$$F_1(y) = 1 + (B_3^2 + 2B_1B_2)y + (2p)^{-1}\{B_1^2 \exp(2py) - B_2^2 \exp(-2py) + 4B_3[B_1 \exp(py) - B_2 \exp(-py)]\}. \quad \text{(A14b)}$$

Note that this solution depends on the three arbitrary constants B_j,

$$B_j = A_j/A_0, \quad j = 1,2,3, \quad \text{(A14c)}$$

and on the parameter p related to V by (A13b).

In the second case ($A_0 = 0$) it is preferable not to solve explicitly the cubic equation (A9), and to write the solution of (A1) in the form

$$g(y) = \sum_{j=1}^{3} [A_j \exp(p_j y)] \times \left\{2 \sum_{j=1}^{3} \sum_{k=1}^{3} \frac{A_j A_k}{p_j + p_k} \exp[(p_j + p_k)y]\right\}^{-1/2}. \quad \text{(A15)}$$

Here of course the three parameters p_j are the three roots of the cubic equation (A9). Note that also this solution depends, for any given V, on three arbitrary constants, namely B [see (A9)] and the two ratios of any two of the three A_j's to the third one.

Let us now identify and analyze, for given *real* V, the solutions $g(y)$ of (A1) that are real and regular for all real values of y.

Consider first solutions of the first type, given by (A14a) with (A14b) and (A13b).

For $V = 0$, $g(y)$ reduces to the trivial solution (A2).

For *positive* V, p is imaginary [see (A13b)]; it is then clear that a necessary condition in order that $g(y)$ be real for all y is that B_3 be real and $B_1 = B_2^*$. It is moreover necessary that $F_1(y)$ be *positive* for all y, and this requires $B_3^2 = -2B_1B_2$ [see (A14b)], namely $B_3^2 = -2|B_1|^2$, which is inconsistent with the reality of B_3 [unless all the constants B_j vanish, in which case $g(y)$ becomes the ultratrivial solution $g(y) = 0$]. Hence for positive V, there is no real solution of (A1) in the first class.

For *negative* V, p is real [see (A13b)], and without loss of generality we assume it is positive,

$$p = (-V)^{1/2} > 0. \quad \text{(A16)}$$

It is then clear that a necessary condition in order that $g(y)$, see (A14a) with (A14b), be real, is that the three constants B_j be all real. It is moreover necessary that $F_1(y)$ be positive, and in order that this be true for large negative y, it is required that B_2 and B_3 both vanish [see (A14b)]. Hence the only real and regular solution of (A1) belonging to the first type reads

$$g(y) = p^{1/2} h[2p(y - \bar{y})] \quad \text{(A17a)}$$

with

$$h(z) = [1 + 2\exp(-z)]^{-1/2}. \quad \text{(A17b)}$$

Here p is related to V by (A16), and $\bar{y}$,

$$\bar{y} = -(2p)^{-1} \ln(B_1^2/p), \quad \text{(A17c)}$$

is an arbitrary real constant.

The following features of the function $h(z)$ and of its derivative are worth noting [see (5.7) and Fig. 1]:

$$0 = h(-\infty) < h(z) < h(+\infty) = 1, \quad \text{(A18a)}$$

$$h'(z) = [\exp(\tfrac{2}{3}z) + 2\exp(-\tfrac{1}{3}z)]^{-3/2} = \tfrac{1}{2}S(z/2) > 0, \quad \text{(A18b)}$$

$$h'(\pm\infty) = 0, \quad \text{(A18c)}$$

$$\operatorname*{Max}_{-\infty < z < +\infty} [h'(z)] = h'(0) = 3^{-3/2}. \quad \text{(A18d)}$$

Let us now proceed to identify and study all the solutions of the second type, see (A15), that are real and regular for all real values of y (for real V).

It is first of all plain that a necessary condition, in order that $g(y)$ be real, is that the arbitrary constant B, see (A9), also be real, so that the three parameters p_j are either all three real or one real and two complex conjugates.

Consider first the case of three real p_j's. Is is then clear that, in order that $g(y)$, see (A15), be real, all three constants A_j must also be real (up to a common arbitrary factor, that may be chosen real without loss of generality). Moreover, if p_j is not positive, the corresponding A_j must vanish, in order that $g(y)$ remain real when y becomes large and negative. But (A10a) implies that at least one of the p_j's is not positive; and if two of the three constants A_j vanish, $g(y)$ reduces to the trivial solution (A2). Hence the only case to be considered obtains when two of the p_j's are positive [and different; otherwise $g(y)$ reduces to the trivial solution (A2)], say

$$0 < p_1 < p_2 \quad \text{(A19a)}$$

and

$$A_3 = 0. \quad \text{(A19b)}$$

Note that (A19a) implies, via (A10a) and (A10b), that V is negative,

$$V = -(p_1^2 + p_2^2 + p_1p_2) < 0. \quad \text{(A19c)}$$

The corresponding solution reads

$$g(y) = p^{1/2} H_s[(1-a)p(y - \bar{y}); a], \quad \text{(A20)}$$

where we have set (for definiteness)

$$p_2 = p, p_1 = ap, 0 < a < 1, \quad \text{(A21a)}$$

so that

$$V = -p^2(1 + a + a^2) \quad \text{(A21b)}$$

and

$$H_s(z;a) = [1 + s\exp(z)] \times [a^{-1} + 4(1+a)^{-1} s \exp(z) + \exp(2z)]^{-1/2}, \quad \text{(A22a)}$$

$$\bar{y} = [(1-a)p]^{-1} \ln|A_1/A_2|, \quad \text{(A22b)}$$

$$s = \operatorname{sgn}(A_1/A_2). \quad \text{(A22c)}$$

Note that this solution depends, for a given (negative) V, on two parameters, namely $\bar{y}$ and either p or a [see (A21b)]; it depends moreover on the sign s.

The following properties of $g(y)$ are plain:

$$g(-\infty) = (ap)^{1/2} = p_1^{1/2}, \quad \text{(A23a)}$$

$$g(+\infty) = sp^{1/2} = sp_2^{1/2}, \tag{A23b}$$

$$g(\bar{y}) = 2p^{1/2}[a(1+a)/(1+6a+a^2)]^{1/2}, \quad \text{if} \quad s=+, \tag{A23c}$$

$$g(\bar{y}) = 0, \quad \text{if} \quad s=-. \tag{A23d}$$

It is also of interest to analyze the behavior of the derivative of $g(y)$,

$$g'(y) = p^{3/2}[(1-a)^2/(1+a)]Z(a^{-1}+Z) \times [a^{-1}+4(1+a)^{-1}Z+Z^2]^{-3/2}. \tag{A24a}$$

Here we have set for convenience

$$Z = s\exp[(1-a)p(y-\bar{y})]. \tag{A24b}$$

Note that, for $s=+$, Z varies from 0 to $+\infty$ as y ranges from $-\infty$ to $+\infty$, while for $s=-$, Z ranges from 0 to $-\infty$ as y ranges from $-\infty$ to $+\infty$. Accordingly, the behavior of $g'(y)$ is rather different depending on s, $s=\pm$. Let us discuss separately the two cases.

For $s=+$, $g'(y)$ vanishes as $y\to\pm\infty$, it is positive for all values of y, and it has a single local (and absolute) maximum at $y=y_3$,

$$y_3 = \bar{y} + [(1-a)p]^{-1}\ln(z_3), \tag{A25}$$

where z_3 is the (only) positive solution of the cubic equation

$$a^2(1+a)z^3 + 2az^2 - 2a^2z - (1+a) = 0. \tag{A26}$$

For instance, for $a=\frac{1}{3}$, $z_3 = \frac{1}{2}(33^{1/2}-3) \approx 1.37$, and $g'(y_3) = (2/27)p^{3/2}$.

For $s=-$, $g'(y)$ vanishes as $y\to\pm\infty$, and it also vanishes at

$$y = y_0 = \bar{y} + [(1-a)p]^{-1}\ln(1/a). \tag{A27}$$

In the interval $-\infty < y < y_0$, $g'(y)$ is negative, and it has a single local (and absolute) minimum at $y=y_2$,

$$y_2 = \bar{y} + [(1-a)p]^{-1}\ln(-z_2), \tag{A28a}$$

where z_2 is the middle solution of the cubic equation (A26) (it is easily seen that $-a^{-1} < z_2 < 0$). In the interval $y_0 < y < +\infty$, $g'(y)$ is positive, and it has a single local (and absolute) maximum at $y=y_1$,

$$y_1 = \bar{y} + [(1-a)p]^{-1}\ln(-z_1), \tag{A28b}$$

where z_1 is the smallest solution of the cubic equation (A26) (it is easily seen that $z_1 < -a^{-1}$). For instance, for $a=\frac{1}{3}$, $z_2 = -\frac{3}{2}$, and $g'(y_2) = -(\frac{4}{3})^{1/2}p^{3/2}$, while $z_3 = -\frac{1}{2}(33^{1/2}+3) \approx -4.37$ and $g'(y_3) = (2/27)p^{3/2}$.

These analytic results may be compared with the graphs displayed in Figs. 2 and 3, via the relation

$$g'(y) = p^{3/2}F_s[p(y-\bar{y}),a]. \tag{A28c}$$

Let us finally consider a solution of the second type, see (A15), with one real and two complex conjugate p_j's, say

$$p_1 = r+iq, \quad p_2 = r-iq, \quad p_3 = -2r, \tag{A29a}$$

with r and q real; note that we have already used (A10a), while (A10b) yields

$$V = q^2 - 3r^2. \tag{A29b}$$

It is then easy to ascertain that, in order that the solution $g(y)$, see (A15), be real and regular for all real values of y [and not reduce to the trivial solution (A2)], it is necessary and sufficient that r be positive, A_1 and A_2 be complex conjugate (up to a common factor), and A_3 vanish:

$$r>0, \quad A_1 = A\exp(ib), \quad A_2 = A\exp(-ib), \quad A_3 = 0, \quad b = b^*. \tag{A30}$$

Then (A15) yields the periodic solution

$$g(y) = (2r)^{1/2}\cos(qy+b) \times \{1+\sin(a)\sin[2(qy+b)+a]\}^{-1/2} \tag{A31a}$$

with

$$\tan(a) = r/q. \tag{A31b}$$

Note that in this case V may have either sign, or it may vanish; and that this solution contains, for given V, two arbitrary real constants, namely b and either r or q [see (A29b) and (A31b)].

Let us end this appendix reporting some properties of the periodic function

$$F(x,a) = \cos(x)/[1+\sin(a)\sin(2x+a)]^{1/2} \tag{A32}$$

[see (A31a)]. They read

$$F_x(x,a) = -\cos(a)\sin(x+a) \times [1+\sin(a)\sin(2x+a)]^{-3/2}, \tag{A33}$$

$$\max_{0<x<2\pi}[F(x,a)] = 1, \tag{A34a}$$

$$\min_{0<x<2\pi}[F(x,a)] = -1. \tag{A34b}$$

Graphs of this function are displayed in Fig. 4.

APPENDIX B

In this appendix we indicate how the ODE

$$2yf' + f + 6c[f''' + 3(f''f^2 + 3f'^2f) + 3f'f^4] = 0, \quad f \equiv f(y), \tag{B1}$$

can be integrated.

A first integration, after multiplication by f, can be performed directly, yielding

$$yf^2 + 3c(2f'' - f'^2 + 6f'f^3 + f^6) = B, \tag{B2}$$

where B is an integration constant.

Now set

$$f(y) = g(y)/[2G(y)]^{1/2} \tag{B3}$$

with

$$G'(y) = g^2(y). \tag{B4}$$

We obtain

$$yg^2 + 3c(2g''g - g'^2) = BG, \tag{B5}$$

and after differentiation [using (B4)] this yields the linear ODE

$$6cg''' + 2yg' + (1-B)g = 0. \tag{B6}$$

This equation can be solved by introducing the Fourier or Laplace transform of $g(y)$. The general solution shall depend on three integration constants, in addition to B; but one of these is a multiplicative constant, and therefore may be factored away when computing $f(y)$, see (B3) and (B4).

On the other hand the evaluation of $G(y)$ from $g(y)$, see (B4), yields an additional integration constant. Hence $f(y)$, when computed from (B3), (B4), and (B6), contains four integration constants (including B). But a relation among these four constants is implied by the requirement that $f(y)$ satisfy (B2) (see the analogous treatment in Appendix A).

Note that, in the special case $B = 1$,

$$g'(y) = h\,[-(3c/2)^{-1/3}y]\,, \tag{B7}$$

with $h(z)$ an Airy function satisfying the second-order linear ODE

$$h''(z) = zh(z). \tag{B8}$$

In the special case $B = 0$ it is convenient to solve directly (B5), setting

$$g(y) = \{h\,[-(6c)^{-1/3}y]\}^2 \tag{B9}$$

and getting again for $h(z)$ the Airy equation (B8).

APPENDIX C

In this appendix we outline the derivation of the results reported in Sec. IV.

Let $v(x,t)$ satisfy the linear evolution equation (3.1), namely

$$v_t(x,t) = v_{xxx}(x,t), \tag{C1}$$

and assume that $v(x,t)$ is regular for real x and vanishes asymptotically ($x \to \pm\infty$) sufficiently fast to guarantee the existence of all the integrals written below.

Now define

$$X_{n,m}(t) = (n!)^{-1}\int_{-\infty}^{+\infty} dx\, x^n [v^{(m)}(x,t)]^2, \tag{C2}$$

where

$$v^{(m)}(x,t) \equiv \frac{\partial^m v(x,t)}{\partial x^m}. \tag{C3}$$

Time differentiation of (C2) yields, using (C1) and integrating by parts,

$$\dot{X}_{n,m}(t) = 3X_{n-1,m+1}(t) - X_{n-3,m}(t) - X_{n-3,m}(t). \tag{C4}$$

This formula holds for $m = 0,1,2,\ldots$ and $n = 0,1,2,\ldots$ [with the provision that, by definition, $X_{n,m}(t) = 0$ if $n < 0$]; and it is plain to verify its consistency with (4.5), and the fact that, via (4.5), it yields (4.6). Q.E.D.

In an analogous manner it is easily seen that the moments

$$Y_n(t) = (-)^n (n!)^{-1}\int_{-\infty}^{+\infty} dx\, x^n v(x,t) \tag{C5}$$

evolve according to the formula

$$\dot{Y}_n(t) = Y_{n-3}(t), \tag{C6}$$

and that this formula implies (4.11) and (4.12). Q.E.D.

APPENDIX D

In this appendix we analyze the behavior in the remote past ($t \to -\infty$) and future ($t \to +\infty$) of the solution

$$u(x,t) = \sum_{n=0}^{N} \frac{A_n \exp(y_n)}{[g(x,t)]^{1/2}} \tag{D1}$$

of the nonlinear evolution PDE (1.1). Here

$$g(x,t) = C^2 + 2\sum_{m,n=0}^{N}\left\{\frac{A_m A_n}{p_m + p_n}\exp(y_m + y_n)\right\}, \tag{D2a}$$

$$g(x,t) = C^2 + \sum_{n=0}^{N}\left[\frac{A_n^2}{p_n}\exp(2y_n)\right] + 2\sum_{n,m=0,n\neq m}^{N}\left\{\frac{A_m A_n}{p_m + p_n}\exp(y_m + y_n)\right\}, \tag{D2b}$$

$$y_n = p_n x + p_n^3 t, \quad n = 1,2,\ldots,N. \tag{D3}$$

Our attention is limited to the case when all the parameters p_n and A_n are real; without loss of generality one can then assume

$$A_n = A_n^* \neq 0, \quad n = 0,1,\ldots,N, \tag{D4}$$

$$p_0 < p_1 < p_2 < \cdots < p_N. \tag{D5}$$

It is moreover clear that, in order that $u(x,t)$ be real and regular for all (real) values of x and t, it is necessary and sufficient that all the parameters p_n be positive, or equivalently

$$p_0 > 0, \tag{D6}$$

since clearly this condition is necessary and sufficient to guaranteee that $g(x,t)$, see (D2b), be positive definite for all real values of x and t (we are of course assuming that C be real, so that C^2 is a non-negative constant).

In order to discuss the behavior in the remote past and future it is convenient to focus attention on the derivative of $u(x,t)$ rather than on $u(x,t)$ itself,

$$u_x(x,t) = f(x,t)/[g(x,t)]^{3/2}, \tag{D7}$$

$$f(x,t) = C^2\sum_{n=0}^{N}[A_m p_m \exp(y_m)] + \sum_{l,m,n=0}^{N}\left\{\frac{A_l A_m A_n (2p_l - p_m - p_n)}{p_m + p_n} \times \exp(y_l + y_m + y_n)\right\}. \tag{D8}$$

The analysis can now be performed setting

$$x = Vt + x' \tag{D9}$$

with x' a fixed parameter, and then investigating for which values of V the function $u_x(x,t)$, see (D7), (D8), and (D2), has a nonvanishing limit as $t \to -\infty$ or $t \to +\infty$. It is moreover convenient to introduce the cubic polynomial

$$c(p) = p^3 + Vp, \tag{D10}$$

and to note that (D3) and (D9) yield

$$y_n = p_n x' + c(p_n)t. \tag{D11}$$

Note that the term with $l = m = n$ in the last sum in the rhs of (D8) is missing (i.e., it is multiplied by a vanishing factor).

Let us begin by discussing the remote future, $t \to +\infty$.

For any non-negative V, all the coefficients $c(p_n)$ are positive, and the largest of them is $c(p_N)$ [see (D10)].

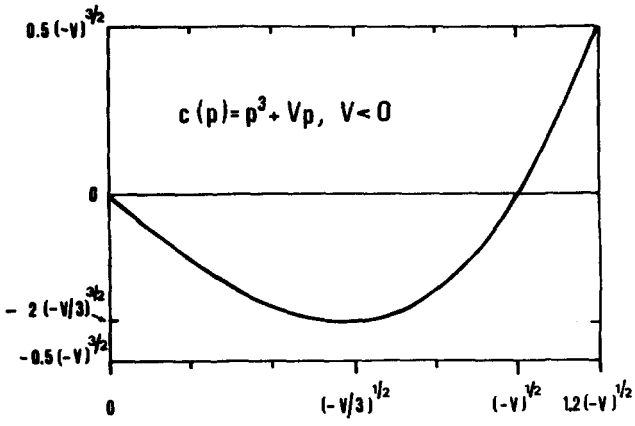


FIG. 5. Graph of $c(p)=p^3+Vp$ with $V<0$; see (D10).

Hence as $t\to+\infty$ g grows proportionally to $\exp[2c(p_N)t]$ [see (D2b) and (D11)], while f grows proportionally to $\exp\{[2c(p_N)+c(p_{N-1})]t\}$ [see (D8) and (D11)]. Hence u_x vanishes asymptotically [proportionally to $\exp\{[c(p_{N-1})-c(p_N)]t\}$; see (D7)].

To analyze the situation for negative V it is useful to refer to the graph of the function $c(p)$ [see (D10)], as displayed in Fig. 5. It is moreover expedient to consider separately two alternative cases: (i) values of V such that the $N+1$ quantities $c(p_n)$ are all different; (ii) values of V such that (at least) two of the $N+1$ quantities $c(p_m)$ coincide, say $c(p_l)=c(p_m)$ for some specific values of l and m [note that we are discussing the behavior of $u_x(x,t)$ for a given set of the $N+1$ parameters p_n, consistent with (D5) and (D6); as it is clear from Fig. 5, it is therefore excluded that three or more $c(p_n)$ coincide].

To analyze the first alternative, let us focus attention on the value of the parameter $c(p_N)$ [see (D5)]. If this parameter is positive, $c(p_N)>0$, it is clear, from the same argument given above for positive V, that $u_x\to 0$ as $t\to+\infty$ (indeed, again proportionally to $\exp\{[c(p_{N-1})-c(p_N)]t\}$; note that, if $c(p_N)>0$, then also $c(p_N)>c(p_{N-1})$, see Fig. 5). If instead the parameter $c(p_N)$ is negative, $c(p_n)<0$, then necessarily all the quantities $c(p_n)$ are also negative, $c(p_n)<0$, $n=0,1,\ldots,N$ (see Fig. 5). Hence in this case f vanishes as $t\to+\infty$ [see (D8) and (D11)] while, if $C\neq 0$, $g\to C^2$ [see (D2) and (D11)], and therefore u_x again vanishes as $t\to+\infty$ [see (D7)]. This conclusion hinges on the condition $C\neq 0$; but the same outcome obtains, by an argument analogous to that given above, if $C=0$, since in such a case f vanishes, as $t\to+\infty$, proportionally to $\exp[(2c_1+c_2)t]$, where c_1 is the least negative of the $N+1$ quantities $c(p_n)$ and c_2 is the second least negative of these $N+1$ quantities, while g vanishes proportionally to $\exp(2c_1t)$.

Let us finally assume that $c(p_N)$ vanishes, namely [see D10) and (D5)]

$$V=-p_N^2; \tag{D12}$$

note that in such a case all the other N quantities $c(p_n)$ are negative, see Fig. 5 and (D5), (D6). It is then again clear that, if $C=0$, u_x vanishes as $t\to+\infty$, since in such a case, in this limit, g tends to a finite (nonvanishing) value while f vanishes. On the other hand if $C\neq 0$ both f and g tend to a finite limit as $t\to+\infty$, and indeed one finds, as $t\to+\infty$

$$u_x(x,t)\approx \mathrm{sgn}(A_N)S_1(x-x_N,t;p_N), \tag{D13}$$

$$x_N=(2p_N)^{-1}\ln[p_NC^2/(2A_N^2)]. \tag{D14}$$

Here the function $S_1(y,t;p)$ is of course defined by (5.6) and (5.7).

Let us now complete the analysis of the behavior of u_x as $t\to+\infty$ by considering the other alternative, namely values of V such that two of the quantities $c(p_n)$ coincide, say

$$c(p_l)=c(p_m)=c. \tag{D15}$$

A glance at Fig. 5 shows that the value c is necessarily negative. On the other hand it is easily seen, by an analysis analogous to those above, that in order that u_x remain finite as $t\to+\infty$ it is necessary and sufficient that the terms with the two exponents (D15) be the dominant ones as $t\to+\infty$ both in f and g [see (D11), (D8), (D2), and (D7)]. For this to happen two conditions must hold: the constant C must vanish [see (D2)], and all the other $c(p_n)$'s [with $n\neq l$, $n\neq m$; see (D15)] must be less than the common value c, see (D15)

$$C=0, \tag{D16}$$

$$c(p_n)<c, \quad n\neq m, \quad n\neq l. \tag{D17}$$

The last condition implies (see Fig. 5) that the two parameters p_l and p_m be the extreme ones of the sequence p_n [see (D5)]; say,

$$p_l=p_0, \quad p_m=p_N. \tag{D18}$$

This condition, together with (D15) and (D10), determines the value of V,

$$V=-(p_N^2+p_0^2+p_Np_0); \tag{D19}$$

and it is then easily seen that u_x has, as $t\to+\infty$, the finite limit

$$u_x(x,t)\approx \mathrm{sgn}(A_N)S_s(x-\bar{x},t;p_0,p_N), \tag{D20}$$

with $S_s(y,t;p_0,p_N)$ defined by (5.15) and

$$\bar{x}=(p_N-p_0)^{-1}\ln|A_0/A_N|, \tag{D21}$$

$$s=\mathrm{sgn}(A_0/A_N). \tag{D22}$$

We may therefore conclude that, as $t\to+\infty$, the behavior of $u_x(x,t)$ is given by (D13) with (D14) if $C\neq 0$ and by (D20) with (D21) and (D22) if $C=0$.

Let us proceed next to consider the behavior of $u_x(x,t)$ as $t\to-\infty$.

First of all, it is easily seen that, if V is non-negative, $u_x(x,t)$ [see (D7)] vanishes as $t\to-\infty$, since all the coefficients $c(p_n)$ are positive [see (D10), (D5), and (D6)], so that, if $C\neq 0$, in the limit f vanishes and g tends to the finite value C^2 [see (D11), (D8), and (D2)], while if $C=0$ both f and g vanish in the limit, but so does u_x [see (D7)].

For negative V, it is again expedient to consider separately two alternative possibilities: (i) values of V such that the $N+1$ parameters $c(p_n)$ [see (D10)] are all different; and (ii) values of V such that there exist (at least) one pair of $c(p_n)$ that coincide [see (D15)].

In the first case, let us focus attention on the value of the parameter $c(p_0)$ [see (D15), (D5), and (D6)]. If this parameter is positive, $c(p_0)>0$, then all the other $c(p_n)$'s are

also positive, $c(p_n)>0$, for $n=1,2,..,N$ (see Fig. 5). It is then clear that, as $t\to-\infty$, f vanishes [see (D8) and (D11)]; as for g, if $C\neq 0$ it tends to the finite value C^2 [see (D2) and (D11)], while if $C=0$ it also vanishes as $t\to-\infty$; but in any case, as $t\to-\infty$, the ratio u_x, see (D7), vanishes. And it is easily seen that the same conclusion, $u_x\to 0$ as $t\to-\infty$, obtains if the parameter $c(p_0)$ is negative, $c(p_0)<0$, as a consequence of the divergence of g and f (g proportionally to $\exp(2c_M t)$, f proportionally to $\exp[(2c_M+c')t]$, where c_M is the most negative one of the $N+1$ quantities $c(p_n)$ and c' is the next to most negative one of these $N+1$ quantities). On the other hand if the parameter $c(p_0)$ vanishes, namely for

$$V=-p_1^2, \tag{D23}$$

then all the other $c(p_n)$'s are positive [see Fig. 5 and recall (D5) and (D6)]; in this case g has a finite limit as $t\to-\infty$ [see (D2) and (D11)], and so does f [see (D8) and (D11)] provided C does not vanish, $C\neq 0$. And it is easily seen that in such a case, as $t\to-\infty$,

$$u_x(x,t)\approx \mathrm{sgn}(A_0)S_1(x-x_0,t;p_0) \tag{D24}$$

with

$$x_0=(2p_0)^{-1}\ln[p_0C^2/(2A_0^2)]. \tag{D25}$$

Here of course the function $S_1(y,t;p)$ is defined by (5.6) and (5.7).

Let us finally consider the second alternative, see (D15). It is then clear that a necessary and sufficient condition in order that u_x not vanish as $t\to-\infty$ is that, in this limit, the two terms corresponding to (D15) provide the dominant (divergent) contributions both in the asymptotic behavior of f [see (D8)] and g [see (D2)]. In order for this to happen the following conditions must hold:

$$c=c(p_l)=c(p_m)<c(p_n),\quad n\neq l,\quad n\neq m. \tag{D26}$$

A glance at Fig. 5 implies that, for this condition to hold, it is necessary and sufficient that the two parameters p_l and p_m for which (D15) holds be contiguous, say

$$p_l=p_{m+1},\quad m=0,1,...,N-1; \tag{D27}$$

note that the corresponding values of V are

$$V=-(p_m^2+p_{m+1}^2+p_mp_{m+1}). \tag{D28}$$

And in such a case it is easily seen that, as $t\to-\infty$,

$$u_x(x,t)\approx \mathrm{sgn}(A_{m+1})S_{s_m}(x-\bar{x}_m,t;p_m,p_{m+1}), \tag{D29}$$

with $S_s(y,t;p_m,p_{m+1})$ defined by (5.15) and

$$\bar{x}_m=(p_{m+1}-p_m)^{-1}\ln|A_m/A_{m+1}|, \tag{D30}$$

$$s_m=\mathrm{sgn}(A_m/A_{m+1}). \tag{D31}$$

We may therefore conclude that, as $t\to-\infty$, the behavior of $u_x(x,t)$ is described by the formula

$$u_x(x,t)\approx \mathrm{sgn}(A_0)S_1(x-x_0,t;p_0)+\sum_{m=0}^{N-1}\mathrm{sgn}(A_{m+1})S_{s_m}(x-\bar{x}_m,t;p_m,p_{m+1}), \tag{D32}$$

with S_1 and S_s defined by (5.6), (5.7), and (5.15) and with x_1, $\bar{x}_m$, and s_m defined by (D25), (D30), and (D31); the first term in the rhs of (D31) is however present only if $C\neq 0$ [this is automatically guaranteed, since if C vanishes, x_0 diverges, see (D25), hence that term disappears, see (5.6) and (5.7)].

[1] J. M. Burgers, *The Nonlinear Diffusion Equation* (Reidel, Dordrecht, 1974).

[2] F. Calogero, "A solvable nonlinear wave equation," Stud. Appl. Math. **70**, 189 (1984).

[3] F. Calogero and A. Degasperis, *Spectral Transform and Solitons* (North-Holland, Amsterdam, 1982), Vol. I.

[4] V. E. Zakharov and E. A. Kuznetsov, "Multi-scale expansions in the theory of systems integrable by the inverse scattering transform," Physica D **18**, 455 (1986).

[5] N. Kh. Ibragimov and A. B. Shabat, "Infinite Lie–Bäcklund algebras," Funkcional Anal. Prilozen. **14** (4), 79 (1980).

[6] O. V. Kaptsov, "Classification of evolution equations by conservation laws," Funkcional Anal. Prilozen. **16** (1), 72 (1982).

[7] V. V. Sokolov and A. B. Shabat, "Necessary conditions on nontrivial Lie–Bäcklund algebras and existence of conservation laws," preprint of the Dept. of Physics and Mathematics of the Bashkirian Section of the Soviet Academy of Sciences, Ufa, 1982 (in Russian).

[8] F. Calogero and W. Eckhaus, "Nonlinear evolution equations, rescalings, more or less integrable model equations. I.," Inverse Problems (in press).

On the WKBJ approximation

M. El Sawi
School of Mathematical Sciences, University of Khartoum, Sudan

(Received 17 November 1983; accepted for publication 22 October 1986)

A simple approach employing properties of solutions of differential equations is adopted to derive an appropriate extension of the WKBJ method. Some of the earlier techniques that are commonly in use are unified, whereby the general approximate solution to a second-order homogeneous linear differential equation is presented in a standard form (SF) that is valid for all orders. In comparison to other methods, the present one is shown to be leading in the order of iteration, and thus possibly has the ability of accelerating the convergence of the solution.

I. INTRODUCTION

The WKBJ method has wide application in quantum mechanics where it is used to find the asymptotic form of the solution of a Sturm–Liouville equation for a large value of the eigenvalue, e.g., in Kreiger *et al.*[1] It is also widely applied in the vast field of ionospheric radio propagation where a wide list of references may be found in Ratchiffe.[2] A general survey of the WKBJ theory and some of its applications may be found in Bender and Orszag,[3] Fröman and Fröman,[4] Heading,[5,6] and Hecht and Mayer.[7]

The purpose of this work is to present a new formulation for the generalized version (SF) of the WKBJ approximation. In their pioneering work Hecht and Mayer[7] extended the WKBJ method using the Schwarzian derivative formalism to obtain solutions to the time-independent Schrödinger equation. They claim that under certain conditions, their method gives results to any degree of accuracy. Unfortunately, they use several transformations, which lead to an indirect iteration scheme, besides the fact that the calculations quickly become unwieldy. Again Fröman and Fröman,[4] using complex variable theory, obtained equations similar to the ones here (SF). They then embarked on a series of mappings and integrals in order to derive an exact formula for the general solution of the Schrödinger equation.

The present treatment gives a simple derivation for the generalized WKBJ method (SF) employing basic properties of the theory of solutions of differential equations. Also the iteration scheme adopted is simple, explicit, and a refinement of earlier ones.

It often happens that the results provided by the first-order theory are not sufficiently accurate. In such cases it becomes necessary to consider second- and higher-order corrections. For instance, Kesarwani and Varshni[8,9] used higher-order corrections to the WKBJ method to improve the results. They have shown that the inclusion of these corrections improve the accuracy of the results. This is certainly in favor of the present method (SF); since when comparing it with the normal approximative methods used, one finds that its second-order approximation is equivalent to the fourth-order approximation of these methods.

II. FUNDAMENTAL EQUATIONS

The WKBJ method is a useful tool for obtaining a global approximation to the solution of a linear differential equation whose highest derivative is multiplied by a small parameter, say ϵ. The present treatment is merely concerned with linear second-order homogeneous differential equations. Any such equation may be transformed to

$$y'' + f(x)y = 0, \quad x \in (a,b), \tag{2.1}$$

which is a form most convenient for our discussion. The function $f(x)$ is taken real, with continuous higher derivatives, and does not vanish in (a,b).

The essence of the WKBJ method is to obtain a general approximate solution to (2.1) subject to $f(x)$ being a slowly varying function. If $f(x)$ were a constant, say k^2, then one should immediately have solutions of the form

$$y^* = Ae^{ikx} + Be^{-ikx}, \tag{2.2}$$

where A and B are arbitrary constants.

In the case when $f(x)$ is no longer constant, but instead a slowly varying function, it might be reasonable to assume that the solution would not be markedly different. Therefore the normal procedure adopted is to assume a solution to (2.1) of the form

$$y = e^{i\phi(x)}, \tag{2.3}$$

thus transforming it into

$$-\phi'^2 + i\phi'' + f(x) = 0, \tag{2.4}$$

which is a Ricatti equation for ϕ'. A standard approach to find approximate solutions to it is to use an iterative method. Let us write Eq. (2.4) in the form

$$\phi'^2_{n+1} = f + i\phi''_n, \quad n = 0,1,2,..., \tag{2.5}$$

where ϕ''_n is assumed to be small in relation to the other quantities. The iterative process is started with the initial value

$$\phi''_0(x) = 0. \tag{2.6}$$

It is profitable at this stage to compute the first four terms in this iteration. These lead to, after integration,

$$\phi_4(x) = \pm T_1 + T_2 \pm T_3 + T_4, \tag{2.7}$$

where

$$T_1 = \int_a^x f^{1/2}\,dt, \quad T_2 = \frac{i}{4}\ln f,$$

$$T_3 = \int_a^x \left\{\frac{5}{32}f'^2 f^{-5/2} - \frac{1}{8}f'' f^{-3/2}\right\} dt, \tag{2.8}$$

 0022-2488/87/030556-03$02.50

$$T_4 = i\left\{\frac{5}{64}f'^2 - \frac{1}{16}f''f^{-2}\right\}.$$

For future reference, let us denote this method by (AM). In the literature however, the (AM) method is seldom used to calculate terms beyond the first order. The tendency is to use the method of formal asymptotic series expansion (FE), for example. This might be attributed to the fact that it leads to explicit results (Bender and Orszag,[3] p. 487), unlike the (AM) method where more care is required in the order of terms to be retained in the expansion involved, as well as in their signs.

III. THE WKBJ APPROXIMATION

In this section the standard form (SF) for the generalized WKBJ approximation is derived. If y_1 and y_2 are two linearly independent solutions to (2.1) then its general solution y^* is known to be

$$y^*(x) = Ay_1 + By_2, \tag{3.1}$$

where

$$y_2(x) = y_1(x)\int_a^x y_1^{-2}\,dt. \tag{3.2}$$

On choosing y_1 to be of the form

$$y_1(x) = g(x)^{-1/2}, \quad g \neq 0, \tag{3.3}$$

Eqs. (3.1) and (3.2) lead to

$$y^*(x) = Ag(x)^{-1/2} + Bg(x)^{-1/2}\int_a^x g(t)dt, \quad g \neq 0. \tag{3.4}$$

Now as y_1 is a solution to (2.1), Eqs. (2.1) and (3.3) lead to

$$f(x) = \tfrac{1}{2}g''g^{-1} - \tfrac{3}{4}g'^2g^{-2}, \quad g \neq 0. \tag{3.5}$$

This equation cannot be solved exactly for $g(x)$, except for very special choices of $f(x)$. On the other hand once $g(x)$ is prescribed, both $f(x)$ and the general solution to (2.1) are completely determined. It is this latter case that is utilized as a basis for the subsequent part of this work.

Let us take $g(x)$ to be of the form

$$g(x) = \phi' e^{2i\phi}, \tag{3.6}$$

where $\phi(x)$ is an arbitrary function of x. This is a very convenient representation for $g(x)$ as an integrand. Now substituting for it into Eq. (3.4) leads to

$$y^*(x) = \phi'^{-1/2}[Ae^{-i\phi} + Be^{i\phi}] \tag{3.7}$$

as the general solution to Eq. (2.1). Another motive for the choice of the form (3.6) for $g(x)$ is the fact that when $\phi = kx$, Eq. (3.7) reduces to (2.2) in which case it is an exact solution to (2.1), with $f(x) = k^2$ from (3.5) and (3.6).

From Eqs. (3.5) and (3.6) one may write

$$\phi'^2 - f(x) = \tfrac{3}{4}\phi''^2\phi'^{-2} - \tfrac{1}{2}\phi'''\phi'^{-1}. \tag{3.8}$$

This equation may now be solved by the iterative process where ϕ'' and ϕ''' are assumed to be small in comparison to the other quantities, which they will be for a slowly varying ϕ'. Thus as initial values of the iteration one may take

$$\phi_0'' = \phi_0''' = 0, \quad \phi_0' \neq 0, \tag{3.9}$$

so that, on retaining the positive sign only, Eq. (3.8) leads to

$$\phi'(x) = \phi_1' = f^{1/2}. \tag{3.10}$$

To pursue this iteration let us write Eq. (3.8) in the form

$$\phi_{n+1}'^2 = f(x) + \tfrac{3}{4}\phi_n''^2\phi_n'^{-2} - \tfrac{1}{2}\phi_n'''\phi_n'^{-1}, \quad n = 0,1,2,\ldots. \tag{3.11}$$

The second term in this iteration is then found to be

$$\phi_2'(x) = f^{1/2}\left[1 + \tfrac{5}{32}f'^2f^{-3} - \tfrac{1}{8}f''f^{-2}\right], \tag{3.12}$$

or on integration one gets

$$\phi_2(x) = T_1 + T_3. \tag{3.13}$$

Equations (3.7) and (3.13) lead to

$$Y^*(x) = \phi_2'^{-1/2}[Ae^{-i\phi_2} + Be^{i\phi_2}], \tag{3.14}$$

as the general solution to Eq. (2.1). This emphasizes the fact that (3.7) represents the general solution of (2.1) for all orders, apart from the order of approximation of ϕ. It thus forms a standard form (SF) for the generalized WKBJ approximation of exact approximate solutions to Eq. (2.1).

IV. COMPARISON OF THE RESULTS

Let us compare the results of the previous section, the (SF) method, with those of the (AM) and (FE) methods. To do so it is beneficial to quote the results for the formal expansion (FE) method in Bender and Orszag[3] (p. 486), rewriting them in a slightly modified form to suit the present notation. Thus the function $\phi(x)$ in Eq. (2.3) is expressed in the form

$$\phi(x) = \sum_{n=1}^{\infty} \epsilon^{n-1}S_n(x). \tag{4.1}$$

The first four terms in this expansion are

$$S_1 = \pm T_1, \quad S_2 = T_2, \quad S_3 = \pm T_3, \quad S_4 = T_4. \tag{4.2}$$

Now since ϵ is a parameter, on taking it to be unity, one finds the expressions (2.7) and (4.1), for ϕ, to be identical up to the fourth order. This shows full agreement between the results obtained by the (AM) and (FE) methods up to the order taken.

Considering next Eqs. (2.3), (4.1), and (4.2) one may write

$$Y^*(x) = e^{i(T_2+T_4)}[Ae^{-i(T_1+T_3)} + Be^{i(T_1+T_3)}]. \tag{4.3}$$

Comparison of the terms in square brackets in Eqs. (3.14) and (4.3), after substituting from (3.13), shows that they are identical. It remains to consider the other terms. On expanding the expresions below, appearing in (3.14) and (4.3), respectively, one finds

$$\phi_2'^{-1/2} = e^{i(T_2+T_4)} = f^{-1/4}\left[1 + \tfrac{1}{16}f''f^{-2} - \tfrac{5}{64}f'^2\right], \tag{4.4}$$

to the same order. Substituting from (4.4) into (3.14) and (4.3), one finds that the two results are equivalent. This proves the equivalence of the three methods, (AM), (FE), and (SF), the only difference being that the present method (SF) has the privilege of leading in the order of iteration in the sense that the second-order result of the (SF) method is the same as the fourth-order result of the (AM) and (FE) methods.

V. CONCLUDING REMARKS

The WKBJ method, despite its evident utility, suffers from lack of completeness regarding the convergence of the series solution (Kesarwani and Varshni[9]), and that these solutions fail at the turning points (Heading[5]), but still in many cases where exact solutions are not possible they are very valuable. Here one hopes that the present method (SF) might close this gap by accelerating convergence, and thus reduce calculations necessary to obtain higher-order approximations. It is also expected that the iteration scheme will converge more rapidly since it is started with a more accurate representation of the exact solution. The method has also the merit of obtaining higher approximations in a simple and direct manner and leads to explicit, linearly independent solutions unified in a single equation. It has also been demonstrated that the solutions it gives agree with those obtained by other established methods, except for the fact that it is leading in the order of iteration.

It is necessary here to throw light on some of the drawbacks of the present method. It is seen that in order to identify a more accurate representation of the solution, one must still work out the first few terms in the simpler (AM) iteration scheme. Another drawback of the method is that, since the iteration proceeds twice as fast as in the simpler (AM) method, each stage requires more differentiability on f than in the simple scheme. Despite these criticisms, the substitution (3.6) does provide a rather neat method.

Finally, one must be encouraged by the attempts that have been made more recently by Taylor[10] to assess the degree of accuracy of the WKBJ method for solutions of Eq. (2.1), where the function $f(x)$ is real and twice continuously differentiable and does not vanish.

ACKNOWLEDGMENTS

The author is indebted to Professor M. O. Taha for helpful comments and to the referee for useful remarks.

[1] J. B. Krieger, M. L. Lewis, and C. Rosezweig, "Use of the WKB method for obtaining energy eigenvalues," J. Chem. Phys. **47**, 2942 (1967).

[2] J. A. Ratchiffe, *Sun, Earth and Radio; An Introduction to the Ionosphere and Magnetosphere* (McGraw-Hill, New York, 1970).

[3] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods* (McGraw-Hill, Tokyo, 1978).

[4] N. Fröman and P. O. Fröman, *JWKB Approximation, Contributions to the Theory* (North-Holland, Amsterdam, 1965).

[5] J. Heading, *An Introduction to Phase-Integral Methods* (Methuen, London, 1962).

[6] J. Heading, "The Stokes phenomenon and generalized continuous transformations of some generalized hypergeometric functions," Proc. Cambridge Philos. Soc. **76**, 423 (1974).

[7] C. E. Hecht and J. E. Mayer, "Extension of the WKB equation," Phys. Rev. **106**, 1156 (1957).

[8] R. N. Kesarwani and Y. P. Varshni, "Third-order WKBJ eigenvalues for Lennard–Jones and Varshni V potentials," Can. J. Phys. **56**, 1488 (1978).

[9] R. N. Kesarwani and Y. P. Varshni, "Five-term WKBJ approximation," J. Math. Phys. **21**, 90 (1980).

[10] J. G. Taylor, "Improved error bounds for the Liouville–Green (or WKB) approximation," J. Math. Anal. Appl. **85**, 79 (1982).

Algebraic structures of degenerate systems and the indefinite metric

Hendrik Grundling and C. A. Hurst
Department of Math Physics, University of Adelaide, GPO Box 498, Adelaide, South Australia 5001

(Received 19 May 1986; accepted for publication 5 November 1986)

It is shown that the indefinite metric structures of degenerate systems as given by Strocchi and Wightman [F. Strocchi and A. S. Wightman, J. Math. Phys. **15**, 2198 (1974); **17**, 1930 (1976)] arise in a natural fashion from the algebraic structure of such systems, where the latter has been developed in a C^*-context by Grundling and Hurst [H. B. G. S Grundling and C. A. Hurst, Commun. Math. Phys. **98**, 369 (1985)]. Auxiliary concepts like gauge equivalence are examined, and the preceding general theory is specialized to the situation of linear boson fields with linear Hermitian constraints. Two examples of this situation are given—a one-dimensional scalar boson in a periodic universe and Landau gauge electromagnetism.

I. INTRODUCTION

The central problem that we address is the following: indefinite metric space methods are only employed in the analysis of degenerate systems. Qualitatively this is because degenerate systems contain nonphysical objects, and this creates the freedom to define nonstandard structures on such objects, if convenience dictates. Now the general algebraic structure of degenerate systems has been developed in Ref. 1, and the indefinite metric space structures necessary for gauge theories has been developed in Refs. 2–4. The question therefore arises what the connection is between these two, if any.

In early physical models it was found that nonpositive definite canonical commutation relations naturally lead to an indefinite inner product for the Fock representation. For electromagnetism, the Gupta–Bleuler approach, which is local and covariant, uses an indefinite inner product space (IIP space), while other approaches, e.g., the Coulomb gauge, represented on a Hibert space, are nonlocal and noncovariant. It would therefore appear that while the physical theory of electromagnetism can be represented on a Hilbert space, the physics is expressed in a more convenient form when represented on an IIP space.[2,5]

More recently, in the framework of the Wightman formulation of field theory, Strocchi[3] showed that all theories with local gauge transformations of the second kind (e.g., Yang–Mills field) must be represented on an IIP space for those transformations to be nontrivially represented. The fact that these gauge theories have such physical desirable properties such as confinement, infrared singularities, etc.,[4] leads one to regard the mathematical structures involved with IIP representations more seriously.

The most important gauge theories—electromagnetism, Yang–Mills, gravitation—are all degenerate theories in the sense of Dirac, i.e., the Hessian of the Lagrangian vanishes, hence constraints or supplementary conditions appear in these theories.[6] This means that there are nonphysical objects present in these theories, which is the situation in IIP representations. By a "degenerate theory," here we will mean simply a theory containing a degree of freedom that has no physical counterpart, so that in this situation the task of the physicist is to extract the physical subtheory. Such a physical subtheory should have the usual structure of a regular theory. There may be several different methods for obtaining the same physical subtheory, and our aim here is in showing the relation between two such methods, viz., the adaptation of the Strocchi–Wightman approach[2] to C^*-algebras, and the method developed in Ref. 1.

The structure of IIP theories has been extensively treated in the Wightman formalism by various authors,[2–4,7] and the algebraic aspects of these theories—still in the Wightman formalism—were considered in Ref. 8. The path integral approach was developed for IIP theories in Ref. 9. A construction similar to the Fock–Cook construction was developed by Mintchev[10] to obtain a Fock-type representation with IIP. Dadashyan and Khoruzhii[11] developed the quasilocal theory for IIP theories in the Wightman formalism. These authors also started a more general study of unbounded operator algebras on IIP spaces, a subject further developed by Jakobczyk in Ref. 12. The theory of IIP spaces is well presented in the book by Bognar.[13] To the best of our knowledge, there is only one study of IIP representations from the purely algebraic field theoretic point of view, and that is a recent publication by Jakobczyk.[14] A publication by Araki[15] considers the specific problem of group representations on an IIP space with additional structure, such as is found in the situation of Gupta–Bleuler electromagnetism, and this theory would become applicable to algebraic field theory, once the latter has been fully developed in the IIP context.

In practice, degenerate systems are always characterized by supplementary conditions that may be *ad hoc*, or be canonical constraints in the sense of Dirac,[6] or be the generators of nonphysical transformations. The imposition of the supplementary conditions is meant to select the physical theory, and one may enquire into the abstract algebraic process that results from this requirement.[1]

Quantized systems consist of an algebra of operators acting on a Hilbert space (or rigged Hilbert space), hence there are two ways of imposing supplementary conditions, i.e., first via conditions on the operators, called algebraic conditions, and second via conditions on the state vectors, called state conditions. These are written as $A = 0$ and $A|\psi\rangle = 0$, respectively. In order to avoid the complications associated with unbounded operators, we consider hence-

 0022-2488/87/030559-14$02.50

forth instead the object $U_\lambda := \exp(i\lambda A)$, and write the conditions as $U_\lambda = \mathbf{1}$ and $U_\lambda|\psi\rangle = |\psi\rangle$, respectively. If A is Hermitian, U_λ is unitary, and hence it may be possible to define abstract elements in a C^*-algebra corresponding to these, e.g., in Segal's method for the algebraic quantization of linear fields.[16] If A is non-Hermitian but A^* satisfy the same conditions as A, the unitary groups can be defined in terms of the Hermitian combinations $A + A^*$ and $i(A - A^*)$. When A is non-Hermitian and A^* does not satisfy the same conditions as A, the Hermitian product A^*A can be used to generate the unitary group U_λ, but in linear field theories[16] it may not be possible to define an element in the abstract algebra corresponding to $\exp(i\lambda A^*A)$, and so difficulties may arise. In the sections to follow, we assume that the U_λ has been defined as an element of the field algebra.

II. BASIC STRUCTURE OF DEGENERATE SYSTEMS

In this section we collect the basic algebraic structures associated with systems with state conditions, as developed in Refs. 1 and 17, which is where the interested reader can find the proofs of the statements below. As in Ref. 18 assume the following.

Assumption 2.1: All physical information of a specified system is contained in the pair $\mathcal{F}$, $\mathfrak{S}$, where the unital C^*-algebra $\mathcal{F}$ is the field algebra, and $\mathfrak{S}$ is its set of states.

Assumption 2.2: There are two specified families of one-parameter groups $\{U_i(\lambda)|\lambda\in\mathbb{R}, i\in I\}$ and $\{V_i(\lambda)|\lambda\in\mathbb{R}, i\in J\}$ in $\mathcal{F}$, called state and algebraic conditions, respectively, where the index sets I, J need not be finite. All physical information is contained in $\mathcal{F}$ and the set of *Dirac states* defined by

$$\mathfrak{S}_D := \{\omega\in\mathfrak{S}|\langle\omega|U_i(\lambda)\rangle = 1 \ \ \forall i,\lambda\}.$$

Then $\omega\in\mathfrak{S}_D$ iff $\langle\omega|AU_i(\lambda)\rangle = \langle\omega|A\rangle = \langle\omega|U_i(\lambda A\rangle \ \ \forall i, \lambda, \forall A\in\mathcal{F}$, or in terms of $L_i(\lambda) := U_i(\lambda) - 1$: $\omega\in\mathfrak{S}_D$ iff $\{L_i(\lambda)\}\subset \mathrm{Ker}\,\omega$ iff $\mathcal{F}\{L_i(\lambda)\}\cup\{L_i(\lambda)\}\mathcal{F}\subset \mathrm{Ker}\,\omega$.

Theorem 2.3: Let $\mathcal{A}(L)$ be the C^*-algebra generated by $\{L_i(\lambda)\}$. Then $\omega\in\mathfrak{S}_D$ iff $\mathcal{A}(L)\subset \mathrm{Ker}\,\omega$ iff $[\mathcal{A}(L)\mathcal{F}\cup\mathcal{F}\mathcal{A}(L)]\subset \mathrm{Ker}\,\omega$, where $[\cdot]$ denotes the closed linear space generated by its argument.

Theorem 2.4: $\mathfrak{S}_D\neq\emptyset$ iff $\mathbf{1}\notin\mathcal{A}(L)$ iff $\mathbf{1}\notin[\mathcal{A}(L)\mathcal{F}\cup\mathcal{F}\mathcal{A}(L)]$, and in this case $\mathfrak{S}_D$ contains pure states.

So our nontriviality assumption is the following.

Assumption 2.5: Henceforth assume $\mathbf{1}\notin\mathcal{A}(L)$.

For any set $\Omega\subset\mathcal{F}$, define

$$\mathcal{M}_{\mathcal{F}}(\Omega) := \{F\in\mathcal{F}|FM\in\Omega\ni MF \ \ \forall M\in\Omega\},$$

hence if Ω is a C^*-algebra, then $\mathcal{M}_{\mathcal{F}}(\Omega)$ is the largest C^*-algebra in $\mathcal{F}$ for which Ω is a two-sided ideal.

Theorem 2.6: Let $\mathcal{N} := [\mathcal{F}\mathcal{A}(L)]$, $\mathcal{D} := \mathcal{N}\cap\mathcal{N}^*$, then $\mathcal{D}$ is the largest C^*-algebra annihilated by all the Dirac states, i.e., $\mathcal{D}$ is the unique maximal C^*-algebra in $\mathcal{K} := \cap\{\mathrm{Ker}\,\omega|\omega\in\mathfrak{S}_D\}$.

Theorem 2.7: $\mathcal{O} := \{F\in\mathcal{F}| \ \ [F,H]\in\mathcal{D} \ \ \forall H\in\mathcal{D}\} = \mathcal{M}_{\mathcal{F}}(\mathcal{D})$. Then $\mathbf{1}\notin\mathcal{D}$, and $\mathcal{D}$ is a proper two-sided ideal for $\mathcal{O}$. In Ref. 6, Dirac defines his observables as "first-class variables" in an analogous way to the way that $\mathcal{O}$ is here defined. The observables in quantum theories are traditionally taken to be $\mathcal{A}(L)'$.

Define $\mathcal{S}$ to be the largest set such that $\mathcal{A}(L)\mathcal{S}\subset[\mathcal{F}\mathcal{A}(L)]$. Then $\mathbf{1}\in\mathcal{A}(L)'\subset\mathcal{S}\cap\mathcal{S}^*$.

Theorem 2.8: $\mathcal{D} = \overline{\mathcal{S}^*\mathcal{A}(L)\mathcal{S}}$ and $\mathcal{O} = \mathcal{S}\cap\mathcal{S}^*$. Hence $\mathcal{A}(L)'\subset\mathcal{O}$, and so we could choose $\mathcal{O}$ even as the set of observable quantities. $\mathcal{O}$ can be considered to be the largest C^*-algebra on which we can impose the constraints. Define the maximal C^*-algebra of physical observables as

$$\mathcal{R} := \mathcal{O}/\mathcal{D}.$$

The factoring procedure is the actual step of imposing the constraints. Now it is possible that $\mathcal{R}$ may not be simple, and this would not be acceptable for a physical algebra. So, using physical arguments, one would in practice choose a C^*-subalgebra $\mathcal{O}_c\subseteq\mathcal{O}$ containing $\mathcal{A}(L)'$ such that

$$\mathcal{R}_c := \mathcal{O}_c/(\mathcal{D}\cap\mathcal{O}_c)\subset\mathcal{R}$$

is simple, and then $\mathcal{R}_c$ is the right physical algebra. The distinction between $\mathcal{O}$ and $\mathcal{O}_c$ was not made in Ref. 1. The procedure of obtaining the objects above is called the T procedure.

Theorem 2.9: $\omega\in\mathfrak{S}_D$ iff $\pi_\omega(\mathcal{D})\Omega_\omega = 0$, where π_ω and Ω_ω are the Gel'fand–Naimark–Segal (GNS) representation of ω and its cyclic vector, respectively.

This corresponds to the heuristic $\chi|\psi\rangle = 0$ method for imposing constraints. Define

$$\Upsilon := \{\alpha\in\mathrm{Aut}\,\mathcal{F}|\mathcal{D} = \alpha[\mathcal{D}]\},$$

then since $\mathcal{O} = \mathcal{M}_{\mathcal{F}}(\mathcal{D})$, α also preserves $\mathcal{O}$ and so defines canonically an automorphism α' on $\mathcal{R}$. Define the group homomorphism T: $\Upsilon\to\mathrm{Aut}\,\mathcal{R}$ by $T(\alpha) = \alpha'$, then we expect Ker T to consist of gauge transformations.

Theorem 2.10: $\mathrm{Ker}\,T = \{\alpha\in\mathrm{Aut}\,\mathcal{F}|\langle\omega|[A]F\rangle = \langle\omega|AF\rangle \ \forall A, F\in\mathfrak{A}$ and $\forall\omega\in\mathfrak{S}_D\}\subset\Upsilon$.

Theorem 2.11: $\alpha\in\mathrm{Inn}\,\mathcal{F}\cap\Upsilon\Rightarrow\alpha'\in\mathrm{Inn}\,\mathcal{R}$. The physical admissible automorphisms of $\mathcal{F}$ denoted by Υ_c are those which are definable on $\mathcal{R}_c$, i.e., $\alpha(\mathcal{O}_c) = \mathcal{O}_c$, and $\alpha(\mathcal{D}_c)\subseteq\mathcal{D}_c := \mathcal{D}\cap\mathcal{O}_c$. Clearly, if $\alpha\in\Upsilon$, it is sufficient that it satisfies $\alpha(\mathcal{O}_c)\subseteq\mathcal{O}_c$ for it to be physically admissible. Similarly to T, we define the group homomorphism T_c: $\Upsilon_c\mapsto\mathrm{Aut}\,\mathcal{R}_c$, and then in this context the gauge transformations will be Ker T_c. The proof of Theorem 2.10 which was given before[1] easily adapts to the new situation to give the statement:

$$\alpha\in\mathrm{Ker}\,T_c\Rightarrow\langle\omega|\alpha[A]F\rangle = \langle\omega|AF\rangle \qquad \forall A,F\in\mathcal{O}_c \ \text{ and } \ \forall\omega\in\mathfrak{S}_D.$$

This will be used later.

The construction above has been shown to lead to results isomorphic to the usual Hilbert space method of imposing supplementary conditions, and moreover that it can fulfill reasonable physical requirements.[17] Moreover, Dirac electromagnetism has been developed as an example of a model which possesses the structure of the general theory above.[1,17]

Next consider the algebraic conditions $\{V_i(\lambda)\}$. Define $N_i(\lambda) := V_i(\lambda) - \mathbf{1}$. It is hard to find an abstract interpretation of the heuristic condition $N_i(\lambda) = 0$. We interpreted it previously[17] to mean either that by construction of $\mathcal{F}$ the abstract object that would have corresponded to $N_i(\lambda)$ is identically zero (cf. Ref. 19 for an example of this ap-

proach), or to mean that there is some *-homomorphism Γ: $\mathscr{P} \subset \mathscr{F} \mapsto \mathscr{R}_a$ onto, with $\{N_i(\lambda)\} \subset \mathrm{Ker}\,\Gamma$. Clearly in this case $\mathscr{R}_a = \mathscr{P}/\mathrm{Ker}\,\Gamma$. Now if Γ is not the T procedure above, there are ordering problems in systems where both types of constraints need to be imposed, and so the natural conclusion is that the two best options for dealing with algebraic conditions are (i) construct $\mathscr{F}$ in such a way that the objects in it which correspond to the heuristic constraints are identically zero, or (ii) treat all constraints on the same footing, i.e., impose them according to the T procedure.

III. SYSTEMS WITH INDEFINITE INNER PRODUCT REPRESENTATIONS

In order to set up the problem of IIP theories, one needs to decide whether the problem is abstract algebraic or representational or both. To this end, consider a typical situation in which the problem arises. In Manuceau's version[20] of Segal's method of algebraic quantization,[16] we start from a manifold which is usually a space of test functions, denoted M, and with a nondegenerate symplectic form $B(\cdot,\cdot)$ on M. Then with the method given in Ref. 20 one constructs the C^*-algebra of the canonical commutation relations (CCR's), $\overline{\Delta(M,B)}$, and this can be taken as the field algebra for the theory. In some approaches, e.g., Ref. 21, $B(\cdot,\cdot)$ is the right-hand side of the smeared CCR's, while in other approaches,[16] it is more indirectly derived from this. Thus the non-positive-definiteness of the CCR's, which is the source of the IIP, in some cases may be reflected in the algebraic structure of the theory. However, in general, it is not clear that this should be the case, for the following reasons. In the process of constructing a Fock representation for the theory, the test function space M is given an inner product so that it becomes a Hilbert space $\mathscr{H}$. The Fock–Cook construction then creates the Fock–Hilbert space $\mathscr{F}(\mathscr{H})$ from $\mathscr{H}$ as the representation space. If M is given an IIP so that $\mathscr{H}$ is just an IIP space, then by Mintchev's construction[10] $\mathscr{F}(\mathscr{H})$ is also an IIP space. Conversely, given an inner product on $\mathscr{H}$, $B(\cdot,\cdot)$ is the imaginary part of it, and this is antisymmetric. The positive-definiteness of an inner product is a property pertaining purely to its real part. However, only its imaginary part $B(\cdot,\cdot)$ enters the algebraic structure of the theory. There is a connection between $B(\cdot,\cdot)$ and the real part of the inner product $\langle\cdot,\cdot\rangle$ of $\mathscr{H}$, given by the complex structure J defined by

J is a real operator on M satisfying $J^2 = -1$;

$B(z,Jz) = 0$ iff $z = 0$;

$B(Jz,Jz') = B(z,z') \quad \forall z, z' \in M$.

Then J defines an inner product on M by $\langle z|z'\rangle := B(z,Jz') + iB(z,z')$. For quadratic Hamiltonians, this complex structure was extensively examined by Broadbridge,[22] who found that in the positive-definite case, there is a unique complex structure for each dynamic action $C(t)$ on M which renders $C(t)$ unitarily implementable in the resultant Hilbert space. He found, however, that if the complex structure induces an IIP, then its existence is highly nonunique for each $C(t)$. Hence for each $B(\cdot,\cdot)$ we can define within physical acceptability a wide variety of IIP's. (This again expresses the arbitrariness of structures that include nonphysical objects.) Thus the connection between the algebraic structure of $\overline{\Delta(M,B)}$ and the positive-definiteness of the metric becomes quite vague. Moreover some important gauge theories, e.g., the Yang–Mills field, have not as yet been cast into a C^*-algebra formulation due to the nonlinearities involved. Therefore it does not seem wise to put too many restrictions on the form of the field algebra on the basis of analogy with present theories. For these reasons, while we are aware that IIP theories may have a slightly different structure in their field algebras, we intend to examine the problem of IIP theories as a purely representational problem, i.e., the problem of the representation of some specified field algebra on a IIP space.

Algebraic field theory is based on the axiom that all physical information of a system is contained in a pair as in Assumption 2.1. If a theory contains nonphysical quantities, we need not necessarily start from such a pair $\{\mathscr{F},\mathscr{C}\}$, as long as the final theory constructed from the degenerate theory satisfies the axiom. The axiom is justified by Hilbert space quantum mechanics, as a C^*-algebra is the abstract version of closed *-algebras of bounded operators on a Hilbert space, and the states can be thought of as expectation values of the observables of the algebra. So in facing an algebra of operators on an IIP space, one may legitimately doubt whether this axiom will still be justified. If we reject the axiom for the total degenerate theory, but still adhere to it for the physical subtheory, the question will arise as to what abstract type of algebra one should take for the field algebra, and here one is faced with the fact that the theory of operator algebras on IIP space is still very rudimentary.[11,12] There are not many hints forthcoming from physics either, due to the presence of nonphysical entities in the theory. We leave the development of these algebras to the mathematicians of the future. In what follows we follow the easier alternative of accepting the structure of Sec. II and the axiom that all physical information is contained in it.

Now having decided to approach the problem as representational, i.e, some unital C^*-algebra $\mathscr{F}$, taken as the field algebra, is represented on an IIP space[12] $\mathscr{H}$, there are two possible problems to consider.

(i) Given a positive subspace $\mathscr{H}'$ as the physical subspace (e.g., selected by imposition of a supplementary condition), what algebraic structure does this imply for $\mathscr{F}$?

(ii) Given the algebraic structure of a degenerate system in $\mathscr{F}$ (as summarized in Sec. II), how do we obtain IIP representations possessing the structure set out in Ref. 2? (Cf. Definitions 3.3 and 3.6 below.)

In the following we will first consider (ii) in detail before returning to (i).

Only ordinary Hilbert space representations can arise via GNS construction from the states $\omega \in \mathscr{C}$, so in order to obtain IIP representations we assume the following.

Assumption 3.1: Given a C^*-algebra $\mathscr{F}$ as a field algebra with its set of states $\mathscr{C}$, all physical information is contained in $\{\mathscr{F},\mathscr{C}\}$. There may also exist some nonpositive functional $f \in \mathscr{F}^*$ which can contain the physical information with $\mathscr{F}$ (and may more conveniently express it). There exists a set of constraints $\{U_i(\lambda)\}$, and for all physical states ω we must have $\langle\omega|U_i(\lambda)\rangle = 1$.

The last sentence expresses the fact that $\mathcal{F}$ is a degenerate system, and from this assumption the theory presented in Sec. II will be applicable, i.e., there is the structure:

$$\{L_i(\lambda)\}\subset\mathcal{A}(L)\subset\mathcal{D}\subset\mathcal{O}\subset\mathcal{F},\ \mathfrak{S}_D,\ \{\mathcal{R},\mathfrak{S}(\mathcal{R})\},$$
$$\mathcal{R}_c := \mathcal{O}_c/\mathcal{D}_c$$

is simple. Henceforth such a structure will be called a C^*-*degenerate system*. A C^*-degenerate system may be denoted simply by $\mathcal{D}\subset\mathcal{F}$ and $\mathcal{O}_c$. Henceforth we only use the ideal structure $\mathcal{D}_c\lhd\mathcal{O}_c$, and the simplicity of $\mathcal{R}_c$, and so omit the subscript c.

Lemma 3.2: Given a linear space X with a degenerate nonnegative inner product $(\cdot,\cdot)$ on it, any bounded operator A with adjoint A^* maps $X_0 := \{x\in X\,|\,(x,x)=0\}$ into itself. Hence the definition of A mod X_0 on X/X_0 makes sense.

This is easily seen via the equation

$$|(Ax,Ax)|^2 = |(x,A^*Ax)|^2 \leqslant (x,x)(Bx,Bx) = 0$$
$$\text{for } x\in X_0,\ A\in\mathcal{B}(X),\ B := A^*A\,.$$

Now we adapt some of the structures developed by Strocchi and Wightman[2–4] in the Wightman formalism to C^*-degenerate systems. In what follows we will not address the problem of the representation of C^*-algebras as operators on IIP spaces directly; this has been done in detail in Ref. 12 for general $*$-algebras with norm.

Definition 3.3: A pre-Strocchi–Wightman structure, denoted PSW structure, consists of an IIP space $\{\mathcal{H},(\cdot,\cdot)\}$, a unital C^*-algebra (the field algebra) $\mathcal{F}_{\mathcal{H}}\subset\mathcal{L}(\mathcal{H})$ within which is specified a C^*-degenerate system $\mathcal{D}_{\mathcal{H}}\lhd\mathcal{O}_{\mathcal{H}}\subset\mathcal{F}_{\mathcal{H}}$ such that there exists a positive semidefinite subspace $\mathcal{H}'\subset\mathcal{H}$ and a cyclic vector $\Phi_0\in\mathcal{H}'$, $\overline{\mathcal{F}_{\mathcal{H}}\Phi_0}=\mathcal{H}$ (closure only if a topology is specified on $\mathcal{H}$) satisfying (i) $\mathcal{O}_{\mathcal{H}}\mathcal{H}'\subset\mathcal{H}'$, and (ii) $\mathcal{D}_{\mathcal{H}}\mathcal{H}'\subset\mathcal{H}''$, where $\mathcal{H}''$ is the neutral space of $\mathcal{H}'$ with respect to $(\cdot,\cdot)$. The physical Hilbert space is defined as $\mathcal{H}_{\text{phys}} := \overline{\mathcal{H}'/\mathcal{H}''}$. Then by Lemma 3.2 the definition of $(A\,|\mathcal{H}')\text{mod}\,\mathcal{H}''$ makes sense for all $A\in\mathcal{O}_{\mathcal{H}}$. Then the unique closure of this operator defines an operator on $\mathcal{H}_{\text{phys}}$, and the physical algebra is defined as $\mathcal{R}_{\text{phys}} := \overline{(\mathcal{O}_{\mathcal{H}}\,|\mathcal{H}')\text{mod}\,\mathcal{H}''}$.

In order of decreasing generality, $\mathcal{H}$ can be chosen to be (i) a general IIP space; (ii) a Hilbert space with an inner product $\langle\cdot|\cdot\rangle$, connected to $(\cdot,\cdot)$ by a bounded linear Hermitian operator G, called a Gram operator, such that $(A,B)=\langle A\,|GB\rangle\,\forall A,B\in\mathcal{H}$ [then $(\cdot,\cdot)$ is jointly continuous in the Hilbert space topology and $\mathcal{H}$ is decomposable[13]]; and (iii) a Krein space, i.e., G is completely invertible. Then $G^2=1$, and the components of $\mathcal{H}$ in any fundamental decomposition are intrinsically complete.

Other choices of $\mathcal{H}$ are possible, but we concentrate on these as the more interesting ones. In what follows, the prefixes general, Hilbert, and Krein will be used to indicate the nature of $\mathcal{H}$ in the PSW structures.

To keep the discussion as general as possible, consider some left $\mathcal{F}$ module X as a way of realizing $\mathcal{F}$ as operators on a space. In addition assume that X has some IIP $(\cdot,\cdot)$ such that $(A^*x,y)=(x,Ay)\ \ \forall A\in\mathcal{F},\ \forall x,y\in X$, and $\exists x_0\in X$ such that $\mathcal{F}x_0=X$.

Theorem 3.4: The collection of objects $\{X,(\cdot,\cdot),\mathcal{F},\{L_i(\lambda)\}\}$ defines a general PSW structure for each $x\in X$ which satisfies (i) $x_0\in\mathcal{O}x$, (ii) $(Ax,Ax)\geqslant 0\ \ \forall A\in\mathcal{O}$, and (iii) $(Dx,Dx)=0\,\forall D\in\mathcal{D}$. Conversely, given a PSW structure for $\mathcal{F}$, $\mathcal{H}$ is given as such a left $\mathcal{F}$ module with cyclic element Φ_0, and any $\Psi\in\mathcal{H}'$ will satisfy (i)–(iii) above. Moreover, the $\mathcal{R}_{\text{phys}}$ derived from any PSW structure is isomorphic to $\mathcal{R}$ (cf. Sec. II), and this means that the PSW structure induces a representation of $\mathcal{R}$. If the set of $x\in X$ satisfying (i)–(iii) is denoted by $\mathcal{I}$, we find that $\mathcal{I}\neq\{0\}\Rightarrow x_0\in\mathcal{I}$. The PSW structure associated with $x\in\mathcal{I}$ defines a cyclic representation of $\mathcal{R}_{\text{phys}}$ on $\mathcal{H}_{\text{phys}}$.

Proof: From $\{\mathcal{F},\{L_i(\lambda)\}\}$ obtain the chain of objects $\{L_i(\lambda)\}\subset\mathcal{A}(L)\subset\mathcal{D}\lhd\mathcal{O}\subset\mathcal{F},\mathcal{R}=\mathcal{O}/\mathcal{D}$ where $\mathcal{R}$ must be simple. Make identifications $\mathcal{H}=X$, $\Phi_0=x_0$, $\mathcal{H}'=\mathcal{O}x$, $\mathcal{H}''=X_0\cap\mathcal{O}x$. Then we find $x_0\in\mathcal{O}x$ is cyclic, $\mathcal{O}\mathcal{H}'=\mathcal{H}'$, $\mathcal{D}\mathcal{H}'=\mathcal{D}\mathcal{O}x\subset\mathcal{D}x\subset\mathcal{H}''$ by (iii), i.e., we have a PSW structure and hence $\mathcal{H}_{\text{phys}}=\mathcal{O}x/(X_0\cap\mathcal{O}x)$. The converse part follows from 3.3. Now

$$A\in\mathcal{O}\rightarrow(A\,|\mathcal{O}x)\text{mod}(X_0\cap\mathcal{O}x)$$

defines a canonical $*$-homomorphism of $\mathcal{O}$ onto $\mathcal{R}_{\text{phys}}$. From $\mathcal{D}(\mathcal{O}x)\subset X_0\cap\mathcal{O}x$ we see that $\mathcal{D}$ is in the kernel of this homomorphism. As $\mathcal{R}$ is simple, $\mathcal{D}$ is maximal in $\mathcal{O}$, hence the kernel is $\mathcal{D}$, and so $\mathcal{R}_{\text{phys}}\simeq\mathcal{R}$. The representation obtained for $\mathcal{R}$ is

$$\pi'(\xi_A) := \overline{(\pi(A)\,|\mathcal{H}')\text{mod}\mathcal{H}''}\quad\forall A\in\mathcal{O}\,.$$

This makes sense because $\pi(\mathcal{D})\mathcal{H}'\subset\mathcal{H}''$. Assume that $\mathcal{I}\neq\{0\}$. Thus $\exists x\neq 0$ such that $x_0\in\mathcal{O}x$, i.e., $\exists B\in\mathcal{O}$ such that $x_0=Bx$. Then

$$(Ax_0,Ax_0)=(ABx,ABx)\geqslant 0$$

and

$$(Dx_0,Dx_0)=(DBx,DBx)=0\ \ \forall A\in\mathcal{O},\ D\in\mathcal{D}\,,$$

since $AB\in\mathcal{O}\ \ \forall A\in\mathcal{O}$ and $DB\in\mathcal{D}\ \ \forall D\in\mathcal{D}$. By definitions it is easy to see that for an $x\in\mathcal{I}$, the equivalence class of x is a cyclic element for $\mathcal{R}_{\text{phys}}$. ■

Remark: At no stage do we require that the IIP on $\mathcal{H}'$ be nondegenerate, because by the Cauchy–Schwartz inequality, the neutral part of $\mathcal{H}'$ is contained in its degenerate part, and hence $\mathcal{H}''=\{0\}$ if $(\cdot,\cdot)$ is to be nondegenerate, and this is not desirable.

Corollary 3.5: Let X be a left $\mathcal{F}$ module with cyclic element x_0, and $\mathcal{F}$ be a C^*-degenerate system. Moreover, let X be a Hilbert space with inner product $\langle\cdot|\cdot\rangle$. Then every pair $\{G,x\}$, $G\in\mathcal{B}_h(X)$, $x\in X$ satisfying

(i) $\langle A^*y|Gz\rangle=\langle y|GAz\rangle\quad\forall y,z\in X,\forall A\in\mathcal{F}$;

(ii) $x_0\in\mathcal{O}x$;

(iii) $\langle x|GA^*Ax\rangle\geqslant 0\quad\forall A\in\mathcal{O}$;

(iv) $\langle x|GD^*Dx\rangle=0\quad\forall D\in\mathcal{D}$;

defines a Hilbert PSW structure, and if $G^2=1$, it is Krein PSW structure.

Proof: This follows from $(A,B)=\langle A\,|GB\rangle$ and Theorem 3.4. ■

Let $\mathcal{G}$ be the physical symmetry group of the field algebra, i.e., there is an action α: $\mathcal{G}\mapsto\text{Aut}\,\mathcal{F}_{\mathcal{H}}$, $\alpha_{\mathcal{G}}\subset\Upsilon_c$. [By $\alpha\in\text{Aut}\,\mathcal{F}_{\mathcal{H}}$, we mean that $\alpha(\pi(\mathcal{F}))=\pi(\tilde{\alpha}(\mathcal{F}))$, where $\tilde{\alpha}\in\text{Aut}\,\mathcal{F}$, $\mathcal{F}_{\mathcal{H}}=\pi(\mathcal{F})$, and π is the IIP representation of $\mathcal{F}$ on $\mathcal{H}$.] Then inspired by Refs. 2–4, we define the following.

Definition 3.6: A *strict Strocchi–Wightman structure*

(henceforth denoted SW structure) is a PSW structure as in 3.3, such that (i) there is a homomorphism U: $\mathcal{G}\mapsto\mathcal{L}(\mathcal{H}')$ for which $U_e = I$ and $(\Phi,A\Psi)\mapsto(U_g\Phi,\alpha_g(A)U_g\Psi)$ $\forall A\in\mathcal{F}_{\mathcal{H}'}$, $g\in\mathcal{G}$; $\Psi,\Phi\in\mathcal{H}'$ and $f(g) := (\Psi,U_g\Phi)$ is a continuous function f: $\mathcal{G}\mapsto\mathbb{C}\forall\Phi,\Psi\in\mathcal{H}'$; (ii) $U_g\mathcal{H}'\subset\mathcal{H}'$ $\forall g\in\mathcal{G}$; and (iii) Φ_0 is the only cyclic vector such that $U_g\Phi_0 = \Phi_0$ $\forall g\in\mathcal{G}$.

A weak Strocchi–Wightman structure is defined below. These correspond to what is called a "gauge" in Ref. 3. In the specific case of the Gupta–Bleuler triplet, there are additional structures and results available.[15] We verify that Definition 3.6 makes sense, i.e., that U_g preserves the PSW structure of $\{\mathcal{H}',\mathcal{F}_{\mathcal{H}'}\}$ and does not transform it into a different PSW structure. As $\mathcal{G}$ is a group, U a homomorphism, $U_gU_{g^{-1}} = U_e = I = U_{g^{-1}}U_g$, i.e., $U_g^{-1} = U_{g^{-1}}$, hence $\operatorname{Ker} U_g = \{0\}$ $\forall g\in\mathcal{G}$. Thus $U_g\mathcal{H}' = \mathcal{H}'$ $\forall g$. Now $U_g\mathcal{H}'\subset\mathcal{H}'\Rightarrow U_{g^{-1}}(U_g\mathcal{H}')\subset U_{g^{-1}}\mathcal{H}'\Rightarrow\mathcal{H}'\subset U_{g^{-1}}\mathcal{H}'$. But $g\in\mathcal{G}$ is arbitrary, hence $U_{g^{-1}}\mathcal{H}'\subset\mathcal{H}'$, and so $U_g\mathcal{H}' = \mathcal{H}'\forall g\in\mathcal{G}$. By (iii), $U_g\Phi_0 = \Phi_0\in\mathcal{H}' = U_g\mathcal{H}'$, and this is still cyclic for $\mathcal{F}_{\mathcal{H}'}$. Now $\alpha_g(\mathcal{O}) = \mathcal{O}$, and so

$$\alpha_g(\mathcal{O})U_g\mathcal{H}' = \mathcal{O}\mathcal{H}'\subset\mathcal{H}',$$

and by

$$(U_g\Phi,U_g\Psi) = (\Phi,\Psi)\ [\text{let } A = \mathbb{1} \text{ in (i)}],$$

we see that $\mathcal{H}_0 = (U_g\mathcal{H}')_0$, so

$$\mathcal{H}'' = \mathcal{H}'\cap\mathcal{H}_0 = U_g(\mathcal{H}')\cap(U_g\mathcal{H}')_0,$$

i.e., $\mathcal{H}''$ is unchanged under U_g, and so is $\mathcal{H}_{\text{phys}}$. Then

$$\alpha_g(\mathcal{D}_{\mathcal{H}'})U_g\mathcal{H}' = \mathcal{D}_{\mathcal{H}'}\mathcal{H}'\subset\mathcal{H}''.$$

This verifies the consistency of Definition 3.6. The additional structure can be easily added to Theorem 3.4 and Corollary 3.5.

Theorem 3.7: Given an SW structure as in 3.6, the automorphisms α_g and $\operatorname{Ad} U_g$ induce the same automorphism on $\mathcal{R}_{\text{phys}}$, and under the canonical isomorphism between $\mathcal{R}_{\text{phys}}$ and $\mathcal{R}$, these map into $\alpha'_g\in\operatorname{Aut}\mathcal{R}$.

Proof: $(\Phi,A\Psi) = (U_g\Phi,\alpha_g(A)U_g\Psi) = (U_g\Phi,U_gA\Psi)$ $\forall A\in\mathcal{F}_{\mathcal{H}'}$; $g\in\mathcal{G}$; $\Phi,\Psi\in\mathcal{H}'$. Thus $(U_g\Phi,(\alpha_g(A)U_g - U_gA)\Psi) = 0$, but as Φ,Ψ are arbitrary, they can be replaced by $\Phi\mapsto U_g^{-1}\Phi$, $\Psi\mapsto U_g^{-1}\Psi$, so

$$(\Phi,(\alpha_g(A) - U_gAU_g^{-1})\Psi) = 0 \quad \forall\Phi,\Psi\in\mathcal{H}',$$

and hence $(\alpha_g(A) - U_gAU_g^{-1})\Psi\in\mathcal{H}_0$ $\forall\Psi\in\mathcal{H}'$, $A\in\mathcal{F}_{\mathcal{H}'}$, $g\in\mathcal{G}$. So

$$(\alpha_g(A) - (\operatorname{Ad} U_g)A)\mathcal{H}'\subset\mathcal{H}'',$$

i.e.,

$$(\alpha_g(A)|\mathcal{H}')\bmod\mathcal{H}'' = (U_gAU_g^{-1}|\mathcal{H}')\bmod\mathcal{H}''.$$

It is easily seen that the procedure $(\alpha_g(A)|\mathcal{H}')\bmod\mathcal{H}''$ defines an automorphism on $\mathcal{R}_{\text{phys}}$ for $\alpha_g\in\Upsilon$, and so the first part of the theorem is proven. The canonical map

$$(A|\mathcal{H}')\bmod\mathcal{H}''\in\mathcal{R}_{\text{phys}}\to\{A+\mathcal{D}\} = \xi_A\in\mathcal{R}\ \forall A\in\mathcal{O}$$

takes $(\alpha_g(A)|\mathcal{H}')\bmod\mathcal{H}''$ to

$$\{\alpha_g(A)+\mathcal{D}\} = \xi_{\alpha_g(A)} = \alpha'_g(\xi_A).$$ ∎

Remark: From the above proof, one sees that in order to get the statement of the theorem to hold, it is sufficient to require that

$$(\alpha_g(A) - (\operatorname{Ad} U_g)A)\mathcal{H}'\subset\mathcal{H}'' \quad \forall A\in\mathcal{O}_{\mathcal{H}'}$$

This will be the alteration to 3.6 which defines a weak Strocchi–Wightman structure. We return to this concept after developing the theory for SW structures. The essential difference is that in SW structures, we have a quasicovariant representation of the full algebra $\mathcal{F}$ on $\mathcal{H}'$, but in weak Strocchi–Wightman structures we have only a quasicovariant representation of $\mathcal{O}$ on $\mathcal{H}'$. As before, the prefixes Krein, general, and Hilbert refer to the structure of the IIP space.

Corollary 3.8: Let $\{X,(\cdot,\cdot),\mathcal{F},\{L_i(\lambda)\}\}$ be as in 3.4, and U: $\mathcal{G}\mapsto\mathcal{L}(X)$ be a homomorphism such that $U_e = I$, $(z,Ay) = (U_gz,\alpha_g(A)U_gy)$ $\forall A\in\mathcal{F}$, $g\in\mathcal{G}$, $z,y\in X$; (z,U_gy) is a continuous function in g for fixed z,y, and x_0 is the only cyclic $U_{\mathcal{G}}$-invariant element of X. Then for each $x\in X$ which satisfies

(i) $x_0\in\mathcal{O}x$,
(ii) $(Ax,Ax)\geqslant 0$ $\forall A\in\mathcal{O}$,
(iii) $(Dx,Dx) = 0$ $\forall D\in\mathcal{D}$,
(iv) $U_g(\mathcal{O}x)\subset\mathcal{O}x$ $\forall g\in\mathcal{G}$,

we have a general SW structure, and apart from $\mathcal{R}_{\text{phys}}\simeq\mathcal{R}$ we find that $\operatorname{Ad} U_g$ maps to an $\alpha'\in\operatorname{Aut}\mathcal{R}$ under this isomorphism. Conversely, given a SW structure for $\mathcal{F}$, any $\Psi\in\mathcal{H}'$, will satisfy (i)–(iv), and moreover, the representation of $\mathcal{R}$ obtained from the SW structure (cf. 3.4) is a covariant representation, and it will be cyclic iff $\exists\Phi\in\mathcal{H}'$ such that $\pi(\mathcal{O})\Phi + \mathcal{H}'' = \mathcal{H}'$. Henceforth we assume this to be the case. Note that we allow the possibility that $\Phi\neq\Phi_0$.

Proof: This is a straightforward application of the preceding theorems and definitions. □

We also add the new structure to Corollary 3.5.

Corollary 3.9: Let X, $\mathcal{F}$, x_0, $\langle\cdot|\cdot\rangle$ be as in 3.5. Let U: $\mathcal{G}\mapsto\mathcal{B}(\mathcal{H})$ be a homomorphism such that $U_e = I$. Then every pair $\{G,x\}\in\mathcal{B}_h(X)\times X$ satisfying

(i) $\langle A^*z|Gy\rangle = \langle z|GAy\rangle$ $\forall A\in\mathcal{F}$, $\forall z,y\in X$;
(ii) $\langle Gz|Ay\rangle = \langle Gz|U_g^{-1}\alpha_g(A)U_gy\rangle$ $\forall A\in\mathcal{F}$ $\forall z,y\in X$, $g\in\mathcal{G}$;
(iii) $\langle z|GU_gy\rangle$ is continuous in g for the other quantities fixed;
(iv) $x_0\in\mathcal{O}x$;
(v) $\langle z|GA^*Ax\rangle\geqslant 0$ $\forall A\in\mathcal{O}$;
(vi) $\langle x|GD^*Dx\rangle = 0$ $\forall D\in\mathcal{D}$;
(vii) x_0 is the only cyclic $U_{\mathcal{G}}$-invariant element of X,

defines a Hilbert SW structure, and if $G^2 = I$, it is a Krein SW structure.

In what follows we wish to develop concrete examples of the left $\mathcal{F}$ module X defined above. Amongst the class of left $\mathcal{F}$ modules there are two important ones—the left ideals of $\mathcal{F}$, and the set of GNS spaces of $\mathcal{F}$. The second already has a Hilbert inner product, so that in this case one looks for a Gram operator. The left ideals of $\mathcal{F}$ are easily equipped with IIP's through the use of Hermitian functionals. By definition each principal left ideal of $\mathcal{F}$ is generated by some one element $x\in\mathcal{F}$ without a left inverse. This element will then be the required cyclic element for the left module, the principal ideal generated by it, as required. As the other left ideals of $\mathcal{F}$ do not have cyclic elements, we are not interested in non-principal ideals.

Theorem 3.10: Let there be given objects $\mathcal{F}$, $\{L_i(\lambda)\}$,

α: $\mathscr{G}\mapsto$Aut $\mathscr{F}$ as above. Let α^P denote the elements of $\mathscr{F}$ which are α invariant. Then there is a SW structure for each (f,x_0,F) in $\mathscr{F}_h^*\times\alpha^P\times\mathscr{F}$ which satisfies the following conditions:

(i) $f(x)=f(\alpha_g(x))\quad\forall x\in\mathscr{F}x_0,\quad\forall g\in\mathscr{G}$;
(ii) $x_0\in\mathscr{O}Fx_0$;
(iii) $\mathscr{O}\alpha_g(F)x_0\subset\mathscr{O}Fx_0$;
(iv) $f(x_0^*F^*\mathscr{O}_+Fx_0)\geqslant 0$;
(v) $f(x_0^*F^*\mathscr{D}_+Fx_0)=0$;
(vi) $\alpha^P\cap\mathscr{F}_lx_0=\{x_0\}$,
where $\mathscr{F}_l:=\{x\in\mathscr{F}\,|\,x_l^{-1}\exists\}$.

Conversely, if there exists a SW structure on the C^*-degenerate system above, then there is a Hermitian functional θ on $\mathscr{F}$ such that $\theta(x)=\theta(\alpha_g(x))\quad\forall x\in\mathscr{F},\forall g\in\mathscr{G},\theta(\mathscr{O}_+)\geqslant 0$, and $\theta(\mathscr{D}_+)=0$.

Proof: Let (f,x_0,F) satisfy (i)–(vi). Then make the identifications $X=\mathscr{F}x_0$, $(A,B):=f(A^*B)\quad\forall A,B\in\mathscr{F}x_0$ with the objects in 3.7, where $\mathscr{F}$ acts by left multiplication on $\mathscr{F}x_0$, (A,B) is a IIP since f is Hermitian and x_0 is the cyclic element of X. When x_0 has no left inverse, $X=\mathscr{F}x_0$ is a proper principal left ideal of $\mathscr{F}$. Let $U_g:=\alpha_g|\mathscr{F}x_0$, since $\alpha_g(\mathscr{F}x_0)=\mathscr{F}x_0$. Then $U_e=I\in\mathscr{L}(X)$,

$$\begin{aligned}(z,Ay)&=f(z^*Ay)=f(\alpha_g(z^*Ay))\\&=f(\alpha_g(z)^*\alpha_g(A)\alpha_g(y))\\&=(U_gz,\alpha_g(A)U_gy)\quad\forall A\in\mathscr{F},\end{aligned}$$

and

$$(z,U_gy)=f(z^*\alpha_g(y))$$

is continuous in g because all the operations involved in its construction are continuous. Furthermore, from (vi), x_0 is the only cyclic U-invariant element of $\mathscr{F}x_0$. Then $x=Fx_0$ will satisfy (i)–(iv) in Corollary 3.8 by (ii)–(v), hence we have a general SW structure. The converse is easily seen from the identification $\theta(x)=(\Phi_0,\pi(x)\Phi_0)$ where $x\in\mathscr{F}$, and π is the IIP representation of $\mathscr{F}$ on $\mathscr{H}$ as in Definition 3.3. ∎

Remark: Later on the functional θ will be called the class functional of the SW structure. When we want to admit spontaneous symmetry breaking, the requirement that x_0 is the only cyclic U-invariant element should be relaxed, in which case (vi) above, can be omitted.

Apart from the set of principal left ideals for the left module X, one can also consider factor spaces of left ideals. Let $\mathscr{J}$ be an $\mathscr{F}$-left ideal containing the subleft ideal $K\subset\mathscr{J}$. Then $\mathscr{J}/K$ is a left $\mathscr{F}$ module, and we need some cyclic element in it, i.e., there should exist an $x_0\in\mathscr{J}$ such that $\mathscr{F}(x_0+K)=\mathscr{J}$ or $\mathscr{F}x_0+K=\mathscr{J}$. Note that $\mathscr{J}/K\simeq\mathscr{F}x_0/(\mathscr{F}x_0\cap K)$.

Theorem 3.11: Given the objects $\mathscr{F}$, $\{L_i(\lambda)\}$, α: $\mathscr{G}\mapsto$Aut $\mathscr{F}$ as above, as well as two α-invariant left ideals $K\subset J\subset\mathscr{F}$ and an $x_0\in J$ such that $\mathscr{F}x_0+K=J$, $\alpha_{\mathscr{G}}(x_0)\subset x_0+K$, there exists a SW structure for each pair $(f,F)\in\mathscr{F}_h^*\times\mathscr{F}$ which satisfies

(i) $K\subset\{x\in J\,|\,f(x^*y)=0\quad\forall y\in J\}$;
(ii) $f(x)=f(\alpha_g(x))\quad\forall g\in\mathscr{G},\quad x\in\mathscr{F}x_0$;
(iii) $x_0\in\mathscr{O}Fx_0+K$;
(iv) $\mathscr{O}\alpha_g(F)x_0\subset\mathscr{O}Fx_0+K$;
(v) $f(x_0^*F^*\mathscr{O}_+Fx_0)\geqslant 0$;
(vi) $f(x_0^*F^*\mathscr{D}_+Fx_0)=0$;
(vii) $\{\mathscr{F}_lx_0+K\}\cap\{x+K\,|\,x\in J,\alpha(x)\in x+K\}$
$\subset\{x_0+K\}$.

Proof: Let ξ be the canonical map $J\mapsto J/K$, hence $\xi^{-1}(\xi_x)=\{x+K\}$. From (i), the IIP defined on J/K by $(\xi_x,\xi_y):=f(x^*y)$ makes sense. Then J/K is a left $\mathscr{F}$ module by $A\xi_x=\xi_{Ax}\quad\forall A\in\mathscr{F}$, which is well-defined because $A\{x+K\}=\{Ax+AK\}\subseteq\{Ax+K\}$. Clearly ξ_{x_0} is a cyclic element in J/K, and as J and K are α invariant, $\exists\alpha'$ defined on J/K by $\alpha_g'(\xi_x):=\xi_{\alpha_g(x)}$, and we define $\alpha_g':=U_g$ (cf. 3.8). Then from (ii) it is easily verified that $U_e=I$; $(\xi_z,A\xi_y)=(U_g\xi_z,\alpha_g(A)U_g\xi_y)$; and $(\xi_z,U_g\xi_y)$ is continuous in g. By (iii) we find $\xi_{x_0}\in\mathscr{O}\xi_{Fx_0}$, and this covers all possibilities because $\xi_{\mathscr{F}x_0}=J/K$. From (v) we see that $(A\xi_{Fx_0},A\xi_{Fx_0})\geqslant 0\quad\forall A\in\mathscr{O}$, and by (vi), $(D\xi_{Fx_0},D\xi_{Fx_0})=0\quad\forall D\in\mathscr{D}$. By (iv), $U_g(\mathscr{O}\xi_{Fx_0})\subset\mathscr{O}\xi_{Fx_0}$, and (vii) ensures that ξ_{x_0} is the only U-invariant cyclic vector in J/K. Hence with $X=J/K$, Corollary 3.8 ensures that there exists a general SW structure. ∎

Remark: When spontaneous symmetry breaking is required, (vii) should be omitted. When x_{0l}^{-1} exists, $\mathscr{F}x_0=\mathscr{F}$, and J/K is a factor space of $\mathscr{F}$.

From the paragraph preceding Theorem 3.11, we note that the construction $J=\mathscr{F}x_0+K$ will in fact cover all possible left ideals for which ξ_{x_0} in cyclic in J/K. It is possible to define Hilbert space structures on the principal left ideals, and then to look for Gram operators from which to obtain SW structures, but as the representations of $\mathscr{F}$ on these Hilbert spaces will be unitarily equivalent to GNS representations, we now consider the latter instead. Consider the GNS representation of $\mathscr{F}$ associated with an $\omega\in\mathfrak{S}$, i.e.,

$$\pi_\omega\colon\mathscr{F}\mapsto\mathscr{B}(\mathscr{H}_\omega),\quad\mathscr{H}_\omega:=\overline{\mathscr{F}/N_\omega},$$
$$N_\omega:=\{A\in\mathscr{F}\,|\,\omega(A^*A)=0\},\quad\xi\colon\mathscr{F}\mapsto\mathscr{F}/N_\omega$$

is the canonical map,

$$\langle\xi_A|\xi_B\rangle:=\omega(A^*B),\quad\pi_\omega(A)\xi_B:=\xi_{AB}.$$

The Gram operators can be chosen from the set $\mathscr{B}_h(\mathscr{H}_\omega)$, and these translate to $\mathscr{F}$ as

$$\{\gamma\in\mathscr{B}(\mathscr{F})\,|\,\gamma(N_\omega)=N_\omega,\ \omega(\gamma(A)^*B)=\omega(A^*\gamma(B))\quad\forall A,B\in\mathscr{F}\}.$$

There are two interesting subsets of this, i.e., where $\gamma\in$Aut $\mathscr{F}$ and where $\gamma\in\mathscr{F}$, i.e., it acts by left multiplication of an element $G\in\mathscr{F}$. The last subset is the one we concentrate on (for the first, consult Ref. 14). Define

$$\mathscr{F}_\omega^G:=\{G\in\mathscr{F}\,|\,\omega(AG^*B)=\omega(AGB)\quad\forall A,B\in\mathscr{F}\}\supset\mathscr{F}_h.$$

Theorem 3.12: Given $\mathscr{F}$, $\{L_i(\lambda)\}$, α: $\mathscr{G}\mapsto$Aut $\mathscr{F}$ as above, there is a Hilbert SW structure for each triplet $(\omega,G,E)\in\mathfrak{S}\times\mathscr{F}_\omega^G\times\mathscr{F}$ such that

(i) $\alpha_{\mathscr{G}}(N_\omega)\subset N_\omega$,
(ii) $\omega(A[G,F]B)=0\quad\forall A,B,F\in\mathscr{F}$,
(iii) $1\in\{\mathscr{O}E+N_\omega\}$,
(iv) $\omega(E^*G\mathscr{O}_+E)\geqslant 0$,
(v) $\omega(E^*G\mathscr{D}_+E)=0$.

When spontaneous symmetry breaking is not allowed, we also require

(vi) $\exists C\in\mathcal{F}\setminus\mathbf{1}$ such that $\mathcal{F}=\{\mathcal{F}C+N_\omega\}$ and $\alpha_{\mathcal{G}}(C)\subseteq\{C+N_\omega\}$.

When $G^2\in\{\mathbf{1}+N_\omega\}$, the SW structure is a Krein SW structure.

Proof: With the identifications $X=\mathcal{H}_\omega$, $x_0=\xi_{\mathbf{1}}$, $U_g\xi_A:=\xi_{\alpha_g(A)}$, which makes sense via (i), the proof is a routine verification of 3.9, but for clarity we spell out some of the details. Clearly U is a homomorphism, and $U_e=I$, and the cyclic vector $\xi_{\mathbf{1}}$ is U invariant because $U_g\xi_{\mathbf{1}}=\xi_{\alpha_g(\mathbf{1})}=\xi_{\mathbf{1}}$. Rewrite (ii): $\omega(A^*GFB)=\omega(A^*FGB)\ \ \forall A,B,F\in\mathcal{F}$ to get

$$\langle\pi_\omega(F)^*\xi_A|\pi_\omega(G)\xi_B\rangle=\langle\xi_A|\pi_\omega(G)\pi_\omega(F)\xi_B\rangle$$

which satisfies (i) of 3.9. To verify (ii) of 3.9 consider

$$\begin{aligned}&\langle\pi_\omega(G)\xi_A|U_g^{-1}\pi_\omega(\alpha_g(F))U_g\xi_B\rangle\\&=\langle\xi_{GA}|U_g^{-1}\xi_{\alpha_g(F)\alpha_g(B)}\rangle\\&=\langle\xi_{GA}|\xi_{\alpha_g^{-1}(\alpha_g(FB))}\rangle=\langle\xi_{GA}|\xi_{FB}\rangle\\&=\langle\pi_\omega(G)\xi_A|\pi_\omega(F)\xi_B\rangle\quad\forall A,B,F\in\mathcal{F},\quad\forall g\in\mathcal{G}\,.\end{aligned}$$

To verify (iii) of 3.9, note that $\langle\xi_A|\pi_\omega(G)U_g\xi_B\rangle=\omega(A^*G\alpha_g(B))$ is continuous in g if all the other quantities are fixed, because all the operations involved in this construction are continuous. For condition (iv) of Corollary 3.9, note that with the identifications $x_0\to\xi_{\mathbf{1}}$, $x\to\xi_E$, (iii) implies $\xi_{\mathbf{1}}\in\xi_{\mathcal{O}E}=\pi_\omega(\mathcal{O})\xi_E$. (iv) above can be rewritten

$$\omega(E^*GA^*AE)=\langle\xi_E|\pi_\omega(G)\pi_\omega(A^*A)\xi_E\rangle\geqslant 0\quad\forall A\in\mathcal{O}\,,$$

which verifies condition (v) of Corollary 3.9. Similarly (v) satisfies (vi) of 3.9.

Note that the first part of (vi) says that ξ_C is cyclic, and the second part that it is U_g invariant: $U_g\xi_C=\xi_{\alpha_g(C)}=\xi_C$, hence (vi) says that there are no cyclic U-invariant elements of $\mathcal{H}_\omega$ other than $\xi_{\mathbf{1}}$, and this corresponds with 3.9 (vii). For the last statement, note that $G^2\in\{\mathbf{1}+N_\omega\}$ implies that $\pi_\omega(G)^2=I$. ∎

If one can find a Dirac state which is invariant with respect to the physical transformation group, this will certainly induce a SW structure on the C^*-degenerate system, and the construction above will be an interesting alternative to the purely algebraic constructions of before. Both methods, as we saw, result in the same final physical algebra. Conversely, given a SW structure as above, from the converse part of Theorem 3.10, we see that on $\mathcal{O}$ this is the restriction of a Dirac state, covariant on $\mathcal{O}$. The interesting case is when this Dirac state on $\mathcal{O}$ cannot be extended to a covariant Dirac state on $\mathcal{F}$. This is the situation when the real usefulness of the IIP representations arise. Hence the purpose of using the IIP formalism is to obtain a cyclic covariant representation for the physical algebra $\mathcal{R}$ that may not be obtainable from covariant ordinary representations on $\mathcal{F}$.

The final concept of Ref. 3, which we adapt to the C^*-algebra context, is that of a generalized gauge transformation. This arises within the following situation. There is a C^*-degenerate system $\{\mathcal{F},\mathcal{D},\mathcal{O}_c\}$ with its transformation group α: $\mathcal{G}\mapsto\mathrm{Aut}\,\mathcal{F}$, $\alpha_{\mathcal{G}}\subset\Upsilon$, and two IIP representations $\pi_i:\mathcal{F}\mapsto\mathcal{L}(\mathcal{H}_i)$, $i=1,2$, and two SW structures

$$\{\mathcal{H}_i,\mathcal{H}_i',(\cdot,\cdot)_i,\Phi_{i0},\pi_i(\mathcal{F}),\pi_i(\mathcal{D}),U^{(i)}\},\quad i=1,2\,.$$

Definition 3.13: A *generalized gauge transformation* is a pair of SW structures as above and a bijection g: $\mathcal{H}_{1\ \mathrm{phys}}\mapsto\mathcal{H}_{2\ \mathrm{phys}}$ such that

(i) $(\Phi_1,\pi_1(A)\Psi_1)_1=(\Phi_2,\pi_2(A)\Psi_2)_2$

$\forall A\in\mathcal{O},\quad\forall\Phi_1,\Psi_1\in\mathcal{H}_1',\quad\Phi_2,\Psi_2\in\mathcal{H}_2'$,

such that $\eta_{\Psi_2}=g(\eta_{\Psi_1})$, $\eta_{\Phi_2}=g(\eta_{\Phi_1})$,

where $\eta:\mathcal{H}_2'\mapsto\mathcal{H}_i'/\mathcal{H}_i''$ is the canonical map,

(ii) $\eta_{\Phi_{20}}=g(\eta_{\Phi_{10}})$.

Clearly what this is saying, is that the two SW structures induce the same covariant cyclic representation on $\mathcal{R}$ up to unitary equivalence (cf. Corollary 3.8, converse part). A special gauge transformation is where the two SW structures are identical except for the IIP representations π_i, g is the identity, and hence (i) is

$$(\Phi,(\pi_1(A)-\pi_2(A))\Psi)=0\quad\forall A\in\mathcal{O},\quad\forall\Phi,\Psi\in\mathcal{H}'\,.$$

Call two SW structures *gauge-equivalent* if there is a generalized gauge transformation connecting them. Given a SW structure, define its *class functional* (cf. Theorem 3.10) to be $\theta(A):=(\Phi_0,\pi(A)\Phi_0)\quad\forall A\in\mathcal{F}$.

Theorem 3.14: If two SW structures are gauge equivalent, then the restrictions of their class functionals to $\mathcal{O}$ are equal. Conversely, given two SW structures with $\theta_1|\mathcal{O}=\theta_2|\mathcal{O}$, and if the $\mathcal{O}$-cyclic elements of $\mathcal{H}_i'$, $i=1,2$, denoted by Φ_i^0 (cf. 3.8), are related to Φ_{i0} by $\Phi_{i0}\in\pi_i(S)\Phi_i^0+\mathcal{H}_i''$ for some $S\in\mathcal{O}$, then these two SW structures are gauge equivalent.

Proof: From the definitions, the first part is clear. We prove the converse. Let there be given two SW structures according to the hypothesis above. The representation induced on $\mathcal{R}$ by these SW structures are

$$\pi_i'(\xi_A)=\overline{(\pi_i(A)|\mathcal{H}_i')\mathrm{mod}\,\mathcal{H}_i''}\quad\forall\in\mathcal{O},\quad i=1,2\,.$$

From the cyclic property (cf. 3.8), any $\Psi_i\in\mathcal{H}_{i\ \mathrm{phys}}$ can be written as $\Psi_i=\pi_i'(\xi_A)\eta_{\Phi_i^0}$ for some $A\in\mathcal{O}$. Note that if $\eta_{\Phi_{i0}}\neq\eta_{\Phi_i^0}$, then $\exists S_i\in\mathcal{O}$ such that $\eta_{\Phi_{i0}}=\pi_i'(\xi_{S_i})\eta_{\Phi_i^0}$. The hypothesis above merely says that $S_1=S_2=:S$. We can define a map g: $\mathcal{H}_{1\ \mathrm{phys}}\mapsto\mathcal{H}_{2\ \mathrm{phys}}$ by

$$g\left[\pi_1'(\xi_A)\eta_{\Phi_1^0}\right]:=\pi_2'(\xi_A)\eta_{\Phi_2^0}\,.$$

This map will be well defined, and a bijection if we can show that

$$\pi_1'(\xi_A)\eta_{\Phi_1^0}=0\ \text{iff}\ \pi_2'(\xi_A)\eta_{\Phi_2^0}=0\,,$$

for each $A\in\mathcal{O}$. Now

$$\begin{aligned}\phi_i(\xi_A):&=\langle\eta_{\Phi_i^0}|\pi_i'(\xi_A)\eta_{\Phi_i^0}\rangle_i=\langle\eta_{\Phi_{i0}}|\pi_i'(\xi_{S^*AS})\eta_{\Phi_{i0}}\rangle_i\\&=\theta_i(S^*AS)\quad\forall A\in\mathcal{O}\,.\end{aligned}$$

So as $S\in\mathcal{O}$, we have $\theta_1(S^*AS)=\theta_2(S^*AS)$, i.e., $\phi_1(\xi_A)=\phi_2(\xi_A)$. Then from $\pi_i'(\xi_A)\eta_{\Phi_i^0}=0$ iff $\xi_A\in N_{\phi_i}:=\{\xi_A\in\mathcal{R}|\xi_{A^*A}\in\mathrm{Ker}\,\phi_i\}$, we get that g is a bijection. To verify 3.13, note that

$$g\left[\pi_1'(\xi_S)\eta_{\Phi_1^0}\right]=g[\eta_{\Phi_{10}}]=\pi_2'(\xi_S)\eta_{\Phi_2^0}=\eta_{\Phi_{20}}\,.$$

Furthermore, $\forall\Phi_1,\Psi_1\in\mathcal{H}_1'$, $\exists B,C\in\mathcal{O}$ such that

$$\eta_{\Phi_1}=\pi_1'(\xi_B)\eta_{\Phi_1^0}$$

and

$$\eta_{\Psi_1} = \pi_1'(\xi_C)\eta_{\Phi_1^0}.$$

Then

$$(\Phi_1, \pi_1(A)\Psi_1)_1 = \langle \eta_{\Phi_1^0} | \pi_1'(\xi_{B^*AC})\eta_{\Phi_1^0}\rangle_1 = \phi_1(\xi_{B^*AC})$$
$$= \theta_1(S^*B^*ACS)$$

and

$$(\Phi_2, \pi_2(A)\Psi_2)_2 = \langle g[\eta_{\Phi_1}] | \pi_2'(\xi_A) g[\eta_{\Psi_1}]\rangle_2$$
$$= \langle \eta_{\Phi_2^0} | \pi_2'(\xi_{B^*AC})\eta_{\Phi_2^0}\rangle_2$$
$$= \phi_2(\xi_{B^*AC}) = \theta_2(S^*B^*ACS),$$

where $\eta_{\Phi_2} = g[\eta_{\Phi_1}]$, $\eta_{\Psi_2} = g[\eta_{\Psi_1}]$. This, together with $\theta_1|\mathcal{O} = \theta_2|\mathcal{O}$, verifies Definition 3.13. ■

Corollary 3.15: In the terminology of Corollary 3.8, two x_i, $i = 1,2$, satisfying 3.8 (i)–(iv), give rise to gauge equivalent SW structures if $\exists S \in \mathcal{O}$ such that

$$x_0 \in (Sx_1 + (X_0 \cap \mathcal{O}x_1)) \cap (Sx_2 + (X_0 \cap \mathcal{O}x_2)).$$

Because the later SW structures are applications of Corollary 3.8, we shall not further discuss gauge equivalence for these.

Remark: If instead of Definition 3.13 (ii) we assumed

(ii') $\eta_{\Phi_2^0} = g[\eta_{\Phi_1^0}]$,

then the results above simplify drastically, in that 3.14 becomes an if and only if statement, i.e., class functionals on $\mathcal{O}$ are equal iff their SW structures are gauge equivalent, and in 3.15 we will find that all the SW structures are gauge equivalent. We can argue for assuming (ii') instead of 3.13 (ii) (which has been taken directly from Ref. 3), on the following grounds. The bijection g is supposed to establish a connection between the two representations obtained for the physical algebra, $\mathcal{R}$, and in these representations the cyclic elements derive from Φ_i^0, and not from the Φ_{i0}'s. Now while $\eta_{\Phi_{i0}}$ is clearly invariant under the unitary transformations of $\mathcal{G}$ on $\mathcal{H}_{i\,\mathrm{phys}}$, it is not cyclic. If one requires that the specified cyclic elements of $\mathcal{H}_{i\,\mathrm{phys}}$ should be invariant under physical transformations, it would be necessary to require in addition that $U_{\mathcal{G}}^{(i)}\Phi_i^0 \in \{\Phi_i^0 + \mathcal{H}_i''\}$. Clearly, if the cyclic elements of $\mathcal{H}_{i\,\mathrm{phys}}$ do derive from Φ_{i0}, then none of these complications will arise.

As was mentioned in the remark below Theorem 3.7, in order to obtain a covariant representation of $\mathcal{R}$, it is sufficient to have the weaker condition

$$(\alpha_g(A) - (\mathrm{Ad}\, U_g)A)\mathcal{H}' \subset \mathcal{H}'' \quad \forall A \in \mathcal{O}_{\mathcal{H}}.$$

This is equivalent to the following.

Definition 3.16: A *weak Strocchi–Wightman structure* (w SW structure) is a PSW structure as in 3.3 for which (i) there is a homomorphism $U: \mathcal{G} \mapsto \mathcal{L}(\mathcal{H})$ such that

$$(\Phi, A\Psi) = (U_g\Phi, \alpha_g(A)U_g\Psi)$$
$$\forall \Phi,\Psi \in \mathcal{H}', \quad \forall A \in \mathcal{O}, \quad \forall g \in \mathcal{G},$$

and $f(g) := (\Psi, U_g\Phi)$ is a continuous function of g for the other quantities fixed;

(ii) $U_g\mathcal{H}' \subset \mathcal{H}' \forall g \in \mathcal{G}$,

(iii) Φ_0 is the only cyclic vector such that $U_g\Phi_0 = \Phi_0 \forall g \in \mathcal{G}$.

It is a straightforward matter to adapt the theorems above to this concept, but for later use we state the following theorem.

Theorem 3.17: Relax the uniqueness of the vacuum requirement. Then for each gauge transformation $\beta \in \mathrm{Ker}\, T_c$ (cf. Sec. II) and functional $f \in \mathcal{F}_h^*$ such that

(i) $f(\beta\alpha_g(A)) = f(A) \quad \forall g \in \mathcal{G}, \quad \forall A \in \mathcal{F}$,

(ii) $f(\mathcal{D}) = 0$,

(iii) $f(\mathcal{O}_+) \geqslant 0$,

we have a wSW structure.

Proof: Let $\xi: \mathcal{F} \mapsto \mathcal{F}/N_f$ be the canonical map, where

$$N_f := \{A \in \mathcal{F} \mid f(A^*B) = 0 \quad \forall B \in \mathcal{F}\}$$

is a left ideal. Then $\mathcal{F}/N_f$ is a left $\mathcal{F}$ module by $A\xi_x = \xi_{Ax} \quad \forall A,x \in \mathcal{F}$, and it has the natural IIP $f(x^*y) =: (\xi_x, \xi_y)$. The cyclic element is ξ_1. Then $\mathcal{H}' = \xi_{\mathcal{O}}$, $\mathcal{H}'' = \xi_{\mathcal{D}}$. Define $U_g(\xi_x) := \xi_{\beta\alpha_g(x)}$. Now $(\beta\alpha_g) \in \Upsilon_c$ if $\alpha_g \in \Upsilon_c$, because $\beta \in \mathrm{Ker}\, T_c \subset \Upsilon_c$, and so $U: \mathcal{G} \mapsto \mathcal{L}(\mathcal{F}/N_f)$ is a homomorphism. Now

$$\beta \in \mathrm{Ker}\, T_c \Rightarrow \langle \omega | \beta(A)x\rangle = \langle \omega | Ax\rangle \quad \forall \omega \in \mathfrak{S}_D, \forall A, \ x \in \mathcal{O}.$$

By simple manipulations

$$\beta \in \mathrm{Ker}\, T \Rightarrow \langle \omega | x\beta(A)y\rangle = \langle \omega | xAy\rangle \quad \forall x,y,A \in \mathcal{O} \quad \forall \omega \in \mathfrak{S}_D.$$

By (ii),(iii), $f|\mathcal{O} \in \mathfrak{S}_D$, and Ker T_c is a group, i.e.,

$$\beta \in \mathrm{Ker}\, T_c \Rightarrow \beta^{-1} \in \mathrm{Ker}\, T_c.$$

Hence

$$(U_g\xi_x, \alpha_g(A)U_g\xi_y)$$
$$= f(\beta\alpha_g(x^*)\alpha_g(A)\beta\alpha_g(y))$$
$$= f((\beta\alpha_g)(x^*)\beta^{-1}(\beta\alpha_g)(A)(\beta\alpha_g)(y))$$
$$= f(\beta\alpha_g(x^*Ay)) = f(x^*Ay)$$
$$= (\xi_x, A\xi_y) \quad \forall x,y, \ A \in \mathcal{O} \quad \forall g \in \mathcal{G}.$$

The function $h(g) := (\xi_x, U_g\xi_y) = f(x^*\beta\alpha_g(y))$ is continuous in g for each pair x,y. Finally,

$$U_g\mathcal{H}' = U_g\xi_{\mathcal{O}} = \xi_{\beta\alpha_g(\mathcal{O})} \subset \mathcal{H}'$$

because $(\beta\alpha_g) \in \Upsilon$. The remaining requirements in the definition of a wSW structure is verified by the same arguments as those found in the proof of 3.11. ■

The important point of this theorem is that for a strict SW structure we needed invariance of the functional under the specified automorphism group [cf. 3.11(ii)], but for a wSW structure, it suffices to have invariance of the functional up to a specific gauge transformation. This will be used in the last example of Sec. IV.

We now return to problem (i), i.e., given a cyclic unital *-algebra $\mathcal{F}$ of operators on a IIP space, and a supplementary condition $\chi \in \mathcal{F}$ such that the physical subspace $\mathcal{H}' := \{\Phi \in \mathcal{H} \mid \chi\Phi = 0\}$ is positive semidefinite and contains the cyclic vector Φ_0, what algebraic structures are implied in $\mathcal{F}$? As $\mathcal{F}$ contains nonphysical objects, it is not clear what physical topology to define on $\mathcal{F}$, although the final nondegenerate physical algebra should be a C^*-algebra.

Denote the null-space of $\mathcal{H}'$ by $\mathcal{H}''$. Then we define the algebra of observables by

$$\mathcal{O}_i := \{A \in \mathcal{F} \mid A\mathcal{H}' \subset \mathcal{H}' \supset A^*\mathcal{H}'\}$$

and the constraint algebra by

$$\mathscr{D}_i := \{D\in\mathscr{F}\,|\,D\mathscr{H}'\subset\mathscr{H}''\supset D^*\mathscr{H}'\}\ni\chi\,.$$

Note that $\{\chi\}'\subset\mathscr{O}_i$, $\mathbf{1}\in\mathscr{O}_i$, $\mathbf{1}\notin\mathscr{D}_i$, and $\mathscr{D}_i\subset\mathscr{O}_i$. The physical algebra is defined as

$$\mathscr{R}_i := (\mathscr{O}_i\,|\,\mathscr{H}')\,\mathrm{mod}\,\mathscr{H}'',$$

and this is required to be a C^*-algebra in the C^*-norm for bounded operators on the physical Hilbert space $\mathscr{H}_{\rm phys} := \overline{\mathscr{H}'/\mathscr{H}''}$. The definitions of $\mathscr{O}_i$ and $\mathscr{D}_i$ preserve sums, multiples, products, and adjoints; hence these are $*$-algebras, and we also verify that $\mathscr{D}_i$ is a two-sided ideal for $\mathscr{O}_i$. Hence $\mathscr{R}_i\simeq\mathscr{O}_i/\mathscr{D}_i$. Now using the definitions and the fact that Φ_0 is cyclic and in $\mathscr{H}'$, we note that $\mathscr{H}'=\mathscr{O}_i\Phi_0$, and $\mathscr{H}''=\mathscr{D}_i\Phi_0$. Then $\chi\mathscr{H}'=\chi\mathscr{O}_i\Phi_0=0$ will only follow from $\chi\Phi_0=0$ if $\mathscr{O}_i\subset\{A\in\mathscr{F}\,|\chi A\in[\mathscr{F}\chi]\}=:\mathscr{S}_i$. Since $\mathscr{O}_i$ is a $*$-algebra, $\mathscr{O}_i\subset\mathscr{S}_i\cap\mathscr{S}_i^*$. Conversely, given $A\in\mathscr{S}_i\cap\mathscr{S}_i^*$, we find $\chi A\mathscr{H}'\subset[\mathscr{F}\chi]\mathscr{O}_i\Phi_0\subset[\mathscr{F}\chi]\Phi_0=0$, i.e., A preserves $\mathscr{H}'$. Similarly A^* preserves $\mathscr{H}'$, and hence $\mathscr{S}_i\cap\mathscr{S}_i^*=\mathscr{O}_i$. This does look like the previous structure, except that $\mathscr{D}_i\supset\mathscr{S}_i^*\mathscr{A}(\chi)\mathscr{S}_i$, but is not equal to it. [$\mathscr{A}(\chi)$ is the $*$-algebra in $\mathscr{F}$ generated by χ.] The reason for this discrepancy is because in the usual situation we deal only with positive functionals and hence Hilbert spaces, and so there are no zero norm states such as in $\mathscr{H}''$ above. The previous structure of $\mathscr{D}$ would have been obtained if we required $\mathscr{D}_i\mathscr{H}'=0$, instead of $\mathscr{D}_i\mathscr{H}'\subset\mathscr{H}''$. Moreover, $\mathscr{S}_i\cap\mathscr{S}_i^*\neq\mathscr{M}(\mathscr{D}_i)$, though it is contained in it, because $\chi\mathscr{M}(\mathscr{D}_i)\Phi_0\neq0$, except in special circumstances. Because $\chi\mathscr{M}(\mathscr{D}_i)\Phi_0\subset\mathscr{H}''$, the generalization of the observables from χ' to $\mathscr{O}_c\subset\mathscr{M}(\mathscr{D}_i)$ also entails the generalization of $\mathscr{H}'$ to $\mathscr{O}_c\Phi_0$, and this will be reasonable only if the latter space is positive semidefinite, and if its zero norm part is $\mathscr{H}''$ exactly.

IV. EXAMPLES: LINEAR BOSON THEORIES

The discussion in this section concerns linear boson fields with linear Hermitian constraints. In Refs. 1 and 17 we considered the theory of electromagnetism as derived from Dirac's constraint theory, and this turned out to be the prototype for any linear boson theory with linear Hermitian constraints. We summarize here the structures obtained.

The field algebra $\mathscr{F}$ is chosen to be Manuceau's C^*-algebra of the CCR,[20] $\overline{\Delta(\mathscr{Q})}$, over a suitable test function space $\mathscr{Q}$ with symplectic form $B(\cdot,\cdot)$. To fix notation, we define $\overline{\Delta(\mathscr{Q})}$ and indicate its heuristic correspondence rules.

Given a canonical pair $q_i(x), p_i(x)$ on a Hilbert space $\mathscr{H}$, with some internal tensor or Lie structure indicated by the index i, and equal-time commutation relation (ETCR): $[q_i(x),p_j(x')]_{x_0=x_0'}=ig_{ij}\delta^3(\underline{x}-\underline{x}')$, smear over a suitable test function space, say $\oplus_i\mathscr{S}^{(i)}(\mathbf{R})$ to obtain the form $(\cdot,\cdot)$ and the CCR: $[q_{x_0}(F),p_{x_0}(G)]=i(F,G)$. Let $\mathscr{Q}$ be the complexification of $\oplus_i\mathscr{S}^{(i)}(\mathbf{R})$ (or equivalently, its direct sum with itself) with the usual norm. Then a symplectic form

$$B(F,G)=B(F_1+iF_2,G_1+iG_2):=(F_1,G_2)-(F_2,G_1)$$

can be defined on it. Using

$$W(F):=\exp(ip_{x_0}(F_1))\exp(iq_{x_0}(F_2))\exp[-i(F_1,F_2)/2]\,,$$

this defines a heuristic Weyl system:

$$W(F)W(F')=W(F+F')\exp[-iB(F,F')/2]\,,$$

which expresses the canonical structure. For commutation relations of the form $[A_\mu(x),A_\nu(x')]=ig_{\mu\nu}\Delta(x-x')$ we obtain a similar Weyl system.[21] Abstractly, the procedure is as follows.[20]

Definition 4.1: Given a linear topological space $\mathscr{Q}$ with a sympletic form B on it, let $\Delta(\mathscr{Q})$ be the normed $*$-algebra such that the following holds.

(i) The elements of $\Delta(\mathscr{Q})$ are complex-valued functions on $\mathscr{Q}$ with support consisting of a finite subset of $\mathscr{Q}$.

(ii) Let $\Delta(\mathscr{Q})$ have the obvious linear structure, and the multiplication law:

$$(f_1f_2)(z):=\sum_{z_1\in\mathscr{Q}}f_1(z_1)f_2(z-z_1)\exp\left[\frac{-iB(z_1,z)}{2}\right].$$

The involution is defined by $f^*(z):=\overline{f(-z)}$.

(iii) Define a norm in $\Delta(\mathscr{Q})$ by $\|f\|_1:=\Sigma_{z\in\mathscr{Q}}|f(z)|$. Denote the completion of $\Delta(\mathscr{Q})$ in this norm by $\Delta_1(\mathscr{Q})$.

The set of functions δ_z such that $\delta_z(z')=1$ if $z=z'$, and zero otherwise, forms a linear basis for $\Delta(\mathscr{Q})$. Then the C^*-algebra of the CCR, $\overline{\Delta(\mathscr{Q})}$, is defined as the enveloping C^*-algebra of $\Delta_1(\mathscr{Q})$, i.e., the closure of the latter in the following C^*-norm $\|f\|:=\sup_{\pi\in P}\|\pi(f)\|$, where P denotes the set of all nondegenerate representations of $\Delta_1(\mathscr{Q})$. Symplectic transformations T on $\mathscr{Q}$ are defined as linear transformations which satisfy $B(Tz,Tz')=B(z,z')\forall z,z'\in\mathscr{Q}$. These can all define automorphisms on $\overline{\Delta(\mathscr{Q})}$ by $\alpha[\delta_z]:=\delta_{Tz}$. Denote the group of symplectic transformations on $\mathscr{Q}$ by $S(\mathscr{Q},B)$.

The connection of $\overline{\Delta(\mathscr{Q})}$ with Weyl systems on $\{\mathscr{Q},B\}$ is furnished by the result[18] that there is a bijection between the nondegenerate representations $\pi\in P$ and the Weyl systems on $\{\mathscr{Q},B\}$, and it is realized by the relation $W_\pi(F)=\pi(\delta_F)$, $F\in\mathscr{Q}$.

Any linear Hermitian combination of p_i, q_i and their derivatives can be specified through a particular element of the complexified test function space $\mathscr{Q}$. Thus if γ is such a combination, and $C\in\mathscr{Q}$ is the element which specifies it, then we have a correspondence $\exp i\lambda\gamma\leftrightarrow\delta_{\lambda C}$, $\lambda\in\mathbf{R}$. So given a subspace $\mathscr{C}\subset\mathscr{Q}$ obtained from the heuristic constraints in this manner, the abstract constraint set (cf. 2.2) is defined as $\mathscr{U}:=\{\delta_F\,|F\in\mathscr{C}\}=\delta_{\mathscr{C}}$. The T procedure is then carried through on this $L_F(\lambda):=\delta_{\lambda F}-\mathbf{1}$, $F\in\mathscr{C}$, and $\omega\in\mathfrak{S}_D$ iff $\langle\omega|\delta_FA\rangle=\langle\omega|A\rangle=\langle\omega|A\delta_F\rangle\ \ \forall A\in\mathscr{F}$, $F\in\mathscr{C}$, etc. With

$$\mathscr{h}:=\{F\in\mathscr{Q}\,|B(F,C)=0\quad\forall C\in\mathscr{C}\}$$

we found that[1] $\delta_{\mathscr{h}}=\delta_{\mathscr{Q}}\cap\mathscr{O}$. Clearly $C^*(\delta_{\mathscr{h}})\subseteq\mathscr{A}(L)'$, and below we will show that $C^*(\delta_{\mathscr{h}})=\mathscr{A}(L)'$. In this notation, $C^*(\cdot)$ means the C^*-algebra generated by its argument in the larger C^*-algebra under consideration (here $\mathscr{F}$). There may be additional elements to these in $\mathscr{O}$, of the form $\Sigma_i\alpha_i\delta_{F_i}$ with $F_i\notin\mathscr{h}$ $\forall i$, but as it is very difficult to get our hands on these, we make the choice $\mathscr{O}_c=C^*(\delta_{\mathscr{h}})=\mathscr{A}(L)'$. Now $\mathscr{A}(L)\setminus\mathscr{O}_c$ does not affect $\mathscr{R}_c$, and hence we might as well require $\mathscr{A}(L)\subset\mathscr{O}_c$, i.e., $\mathscr{O}\subset\mathscr{h}$. In this case $\mathscr{A}(L)$ is commutative. Then $\mathscr{D}\cap\mathscr{O}_c=\mathscr{A}(L)C^*(\delta_{\mathscr{h}})$, and so the chosen physical algebra is $\mathscr{R}_c=\mathscr{A}(L)'/$

$\overline{\mathscr{A}(L)\mathscr{A}(L)'}$. In Ref. 1 we have not made this distinction between $\mathscr{O}$ and $\mathscr{O}_c$. Current work on a different problem proved $\mathscr{R}_c$ to be nontrivial, and to be simple if $\mathscr{C}$ is the degenerate part of $\mathscr{I}$ with respect to B.

First we show that $C^*(\delta_{\mathscr{I}}) = \mathscr{A}(L)'$.

Theorem 4.2: $\mathscr{A}(L)' = C^*(\delta_{\mathscr{I}})$.

Proof: It suffices to show $\mathscr{A}(L)' \subseteq C^*(\delta_{\mathscr{I}})$. Consider $A \in \mathscr{A}(L)' \cap \Delta(\mathscr{Q})$. Then $A = \Sigma_i^n \lambda_i \delta_{F_i}$, $F_i \in \mathscr{Q}$, $\lambda_i \in \mathbb{C}$, $n < \infty$. Then

$$A \in \mathscr{A}(L)' \Rightarrow \Sigma_i^n \lambda_i [\delta_{F_i}, \delta_C] = 0 \quad \forall C \in \mathscr{C}$$

$$\Rightarrow \Sigma_i \lambda_i \delta_{F_i + C} 2 \sin \tfrac{1}{2} B(F_i, C) = 0 .$$

Now as δ_F is a linear basis of $\Delta(\mathscr{Q})$, and $F_i + C = F_j + C$ iff $F_i = F_j$, we get $B(F_i, C) = 0 \quad \forall C \in \mathscr{C}$, and hence $F_i \in \mathscr{I}$. Thus $A \in C^*(\delta_{\mathscr{I}})$, i.e.,

$$\mathscr{A}(L)' \cap \Delta(\mathscr{Q}) \subset C^*(\delta_{\mathscr{I}}) \subset \mathscr{A}(L)' .$$

By Ref. 20, $\Delta(\mathscr{Q})$ is dense in $\mathscr{F} = \overline{\Delta(\mathscr{Q})}$, and hence any element in $\mathscr{F}$ can be reached as the limit of some Cauchy sequence in $\Delta(\mathscr{Q})$. So in order to prove that $C^*(\delta_{\mathscr{I}}) = \mathscr{A}(L)'$, we need to show that each element in $\mathscr{A}(L)'$ can be reached by a Cauchy sequence in $\mathscr{A}(L)' \cap \Delta(\mathscr{Q})$. Thus we wish to show that for any sequence $\{A_j\}_{j=1}^{\infty}$ converging to $H \in \mathscr{A}(L)'$, where $A_j := \Sigma_{i=1}^{n_j} \lambda_i^j \delta_{F_i} \in \Delta(\mathscr{Q})$ is such that $\|\pi(A_j) - \pi(A_k)\| \to 0$ for $j, k \to \infty \quad \forall \pi \in P$: = set of nondegenerate representations of $\Delta_1(\mathscr{Q})$, there exists a similar sequence $\{B_j\}$, where $B_j := \Sigma_{i=1}^{n_j} \gamma_i^j \delta_{F_i}$ such that all $F_i \in \mathscr{I}$, and this sequence converges to the same element $H \in \mathscr{A}(L)'$. Assume that all the F_i of all the terms of the sequences used here are united into a single set over which a single index ranges. Denote those F_i in $\mathscr{I}$ by P_i, and those F_i not in $\mathscr{I}$ by T_i. Let $\{A_j\}$ be a Cauchy sequence in $\Delta(\mathscr{Q})$ converging to an $H \in \mathscr{A}(L')$. Then we can write

$$A_j = \sum_i \lambda_i^j \delta_{F_i} = \sum_i \alpha_i^j \delta_{P_i} + \sum_i \beta_i^j \delta_{T_i} .$$

Then for $\pi \in P$: $\|\Sigma_i \beta_i^j \pi([\mathscr{A}(L), \delta_{T_i}])\| \to 0$ as $j \to \infty$, and hence for $L(C) \in \mathscr{A}(L)$, $C \in \mathscr{C}$ we have

$$\left\| \sum_i^{n_j} \beta_i^j \pi([\delta_C, \delta_{T_i}]) \right\| = \left\| 2i \sum_i^{n_j} \beta_i^j \pi(\delta_{C+T_i}) \sin \tfrac{1}{2} B(C, T_i) \right\| \to 0$$

as $j \to \infty$, $\forall \pi \in P$, and $\forall C \in \mathscr{C}$. Thus, $\forall C \in \mathscr{C}$,

$$\sum_{i,k}^{n_j} \overline{\beta}_i^j \beta_k^j \sin \frac{1}{2} B(C, T_i) \sin \frac{1}{2} B(C, T_k) \times \rho_\pi(T_k - T_i) \exp(i/2) B(T_i, T_k) \to 0$$

as $j \to \infty$ for all generating functionals ρ_π. This is seen from $\|\pi(A)\|^2 = \langle \xi_0 | \pi(A^* A) | \xi_0 \rangle$. Since only the β's depend on j, consider equality for the limits β_i. Moreover, if this equation holds for a particular $C \in \mathscr{C}$, it must hold for λC, $\lambda \in \mathbb{R}$. Let $n_j \to n$ be arbitrary large but finite for the moment. So

$$\sum_{i,k}^{n} \overline{\beta}_i \beta_k \rho_\pi(T_k - T_i) \exp \frac{i}{2} B(T_i, T_k) \times \sin \frac{\lambda}{2} B(C, T_i) \sin \frac{\lambda}{2} B(C, T_k) = 0 ,$$

$\forall \lambda \in \mathbb{R}$, $C \notin \mathscr{C}$. Since $T_i \notin \mathscr{I}$, $\exists C_i \in \mathscr{C}$ such that $B(C_i, T_i) \neq 0$. Take a suitable linear combination $C := \Sigma_i^n \alpha_i C_i$ such that $\Gamma_i := \frac{1}{2} B(C, T_i) \neq 0 \quad \forall i$. Then it may not be possible to distinguish all the T_i through such elements $C \in \mathscr{C}$, i.e., it may be that there are values of i, say l and m for which $\Gamma_l = \Gamma_m$. This is the case if for instance $T_l - T_m \in \mathscr{I}$. Moreover,

$$\Lambda_{ik} := \overline{\beta}_i \beta_k \rho_\pi(T_k - T_i) \times \exp(i/2) B(T_i, T_k) \neq 0 \quad \forall i, k \leqslant n .$$

So we wish to solve

$$\Sigma_{i,k}^n \Lambda_{ik} \sin(\lambda \Gamma_i) \sin(\lambda \Gamma_k) = 0 \quad \forall \lambda \in \mathbb{R}$$

for Λ_{ik}, where Γ_i are given nonzero numbers. Now (Λ) is a positive definite $n \times n$ matrix because

$$\tilde{\gamma} \Lambda \gamma = \sum_{i,k}^{n} \bar{\gamma}_i \Lambda_{ik} \gamma_k = \sum_{i,k}^{n} \bar{\gamma}_i \overline{\beta}_i \gamma_k \beta_k \rho_\pi(T_k - T_i) \exp \frac{i}{2} B(T_i, T_k) = \langle \xi_0 | \pi\left(\left(\sum_i^n \gamma_i \beta_i \delta_{T_i} \right)^* \left(\sum_k^n \gamma_k \beta_k \delta_{T_k} \right) \right) | \xi_0 \rangle \geqslant 0 ,$$

$\forall \gamma \in \mathbb{C}^n$. Hence Λ defines a positive sesquilinear form on $\mathbb{C}^n$ by $(\gamma, \delta) := \tilde{\gamma} \Lambda \delta$. Then an application of the Cauchy–Schwartz inequality, we find that if $\tilde{\gamma} \Lambda \gamma = 0$, then $\tilde{\delta} \Lambda \gamma = 0 \quad \forall \delta \in \mathbb{C}^n$. Hence we obtain

$$\sum_{i,k}^{n} \Lambda_{ik} \gamma_i \sin(\lambda \Gamma_k) = 0 \quad \forall \lambda \in \mathbb{R}, \quad \gamma \in \mathbb{C}^n .$$

By letting γ vary over the usual basis of $\mathbb{C}^n$, we obtain the following system of equations:

$$\Sigma_{k=1}^n \Lambda_{ik} \sin(\lambda \Gamma_k) = 0 \quad \forall \lambda \in \mathbb{R}, \quad i \leqslant n .$$

By varying λ, we can very quickly overdetermine the Λ_{ik}'s, and except for those values of i corresponding to the Γ_m which are not different from all the other Γ_i's, obtain $\Lambda_{ik} = 0$. However if X is an index set for which all Γ_i's are equal, then for those Λ_{ik} it is only possible to say $\Sigma_{i \in X} \Lambda_{ki} = 0 \quad \forall k$. Now since it is not possible to specify with the given information what the sets X are, in general we only say that $\Sigma_{ik}^n \Lambda_{ik} = 0$. This means $\forall \pi \in P$,

$$\sum_{i,k}^{n} \overline{\beta}_i \beta_k \rho_\pi(T_k - T_i) \exp \frac{i}{2} B(T_i, T_k) = \left\| \pi\left(\sum_i^n \beta_i \delta_{T_i} \right) \right\|^2 = 0$$

and so $\Sigma_i^n \beta_i \delta_{T_i} = 0$. With the necessary formal adaptations, this argument generalizes to $n = \infty$, and so $\Sigma_i^{n_j} \beta_i^j \delta_{T_i} \to 0$ as $j \to \infty$, i.e., $\{A_j\}_{j=1}^{\infty}$ converges to the same limit as

$$\{\Sigma_i^{n_j} \alpha_i^j \delta_{P_i}\}_{j=1}^{\infty} \subset C^*(\delta_{\mathscr{I}}) .$$

Therefore $\mathscr{A}(L') = C^*(\delta_{\mathscr{I}})$. ■

To conclude this general account of degenerate linear boson fields, we remark that if G is the physical transformation group (containing the dynamics), then it is generally represented by symplectic transformations on $\mathscr{Q}$, i.e., $T: G \to S(\mathscr{Q}, B)$, and we obtain the corresponding automorphisms on $\mathscr{F}$ by $\alpha_g(\delta_F) := \delta_{T_g F}$. Then for the constraints to be consistent with these automorphisms, it is sufficient to have $T_G \mathscr{C} \subset \mathscr{C}$, in which case it follows that $T_G \mathscr{I} = \mathscr{I}$.

Next we wish to demonstrate the existence of SW structures and wSW structures on the degenerate linear boson fields just described. First, consider the class of Fock representations. These representations $\pi_i: \mathcal{F} \mapsto B(\mathcal{H}_i^F)$ are associated to generating functionals ρ_i on $\mathcal{D}$ by

$$\rho_i(F) = \langle \xi_0 | \pi_i(\delta_F) | \xi_0 \rangle = \exp[-\tfrac{1}{4}(F,F)_i], \qquad (*)$$

where $(\cdot,\cdot)_i$ is an inner product on $\mathcal{D}$ such that $\mathrm{Im}(\cdot,\cdot)_i = B(\cdot,\cdot)$, and ξ_0 is the cyclic element of $\mathcal{H}_i^F$. Given a generating functional ρ_i, it will through the expression $(*)$ define a positive functional ω_i^0 on $\Delta(\mathcal{D})$. This will define a positive functional ω_i on $\mathcal{F} = \overline{\Delta(\mathcal{D})}$ only if ω_i^0 is continuous in the C^*-topology, i.e., if

$$\rho_i(F) = \exp[-\tfrac{1}{4}(F,F)_i] \leqslant C\|\delta_F\| = C \in \mathbb{R}_+ .$$

This is only satisfied if $(F,F)_i \geqslant 0 \quad \forall F \in \mathcal{D}$. In this case one finds that the GNS representation space of ω_i is isometrically equivalent to a Fock space constructed on the Hilbert space obtained from the completion of $\mathcal{D}$ in the $(\cdot,\cdot)$ topology. However, it also indicates that in the IIP situation, we should expect problems in trying to execute a similar construction from generating functionals of the form ρ_i.

Starting from a IIP space, $\mathcal{H}_i^{(1)} = \{\mathcal{D}, (\cdot,\cdot)_i\}$, Mintchev[10] constructed a Fock-type IIP space, on which he could define a Weyl system such that its vacuum expectation value is exactly ρ_i, i.e.,

$$\langle \xi_0 | W(F) | \xi_0 \rangle = \exp[-\tfrac{1}{4}(F,F)_i] = \rho_i(F) .$$

Hence we expect some structure resembling the triplet $\mathcal{H}$, $\mathcal{H}'$, $\mathcal{H}''$ of 3.3 to be present in $\mathcal{H}_i^{(1)}$ for a degenerate physical system. Now such a structure is already available in a degenerate linear boson field, in the form $\mathcal{H}_i^{(1)}, \mathcal{K}, \mathcal{C}$, and so in what follows we will demonstrate the existence of SW structures or wSW structures associated with IIP's $(\cdot,\cdot)_i$ on $\mathcal{D}$ which satisfy

$$(F,F)_i \geqslant 0 \quad \forall F \in \mathcal{K}; \quad (C,C)_i = 0 \quad \forall C \in \mathcal{C};$$

and

$$(T_g F, T_g H)_i = (F,H)_i \, \forall F,H \in \mathcal{D}, \quad g \in G.$$

Note that from the positivity on $\mathcal{K}$ we can apply the Cauchy–Schwartz inequality to get an equivalent condition for the second one: $(F,C)_i = 0 \quad \forall F \in \mathcal{K}, \; C \in \mathcal{C}$. Below we show a natural way for obtaining an indefinite functional f_i on $\mathcal{F}$ from ρ_i [defined from a $(\cdot,\cdot)_i$ as above], which will give the right SW structures. The uniqueness of the vacuum requirement will not be enforced.

Theorem 4.3: Given an IIP space $\{\mathcal{D}, (\cdot,\cdot)_i\}$ such that $(F,F)_i \geqslant 0 \quad \forall F \in \mathcal{K}$ and $(C,C)_i = 0 \quad \forall C \in \mathcal{C}$, the functional $\rho_i(F) := \exp[-\tfrac{1}{4}(F,F)_i]$ will define a state ω_i on $C^*(\delta_{\mathcal{K}})$ such that $\mathcal{D} = \overline{\mathcal{A}(L)\mathcal{A}(L)'} \subset \mathrm{Ker}\,\omega_i$. Moreover, if $(T_g F, T_g H)_i = (F,H)_i \quad \forall F,H \in \mathcal{K}, g \in G$, then ω_i is G invariant. On the other hand, if the symplectic transformation S defines a gauge transformation $\beta \in \mathrm{Ker}\, T_c$ and $(ST_g F, ST_g H)_i = (F,H)_i \; \forall F,H \in \mathcal{K}, g \in G$, then ω_i is G invariant up to this gauge transformation β.

Proof: Since ω_i is continuous on a generating set of $C^*(\delta_{\mathcal{K}})$ by

$$\omega_i(\delta_F) := \rho_i(F) = \exp[-\tfrac{1}{4}(F,F)_i] \leqslant 1 = \|\delta_F\| \quad \forall F \in \mathcal{K},$$

we get that ρ_i will define the continuous linear functional ω_i on $C^*(\delta_{\mathcal{K}})$. First, we verify that ω_i is positive on $C^*(\delta_{\mathcal{K}})$. Now if we can show that ω_i is positive on a generating set of

$$\mathcal{O}_+ = (\mathcal{A}(L)')_+ = C^*(\delta_{\mathcal{K}})_+ ,$$

then it will be positive on all convex combinations of those elements and their limits. The generating set of $\mathcal{O}_+$ is

$$\left\{ A^* A \;\middle|\; A = \sum_{i=1}^{\infty} \lambda_i \delta_{F_i}, \quad F_i \in \mathcal{K}, \; \lambda_i \in \mathbb{C}, \; \sum_i |\lambda_i| < \infty \right\} .$$

Notation; $A_n := \Sigma_{i=1}^n \lambda_i \delta_{F_i}$, $F_i \in \mathcal{K}$; $(F,F)_i = :(F)_i^2$, and the $(\cdot)_i$ is a prenorm on $\mathcal{K}$. So we wish to show that $\omega_i(A_n^* A_n) \geqslant 0 \; \forall n \in \mathbb{Z}_+$,

$$\omega_i(A_n^* A_n) = \sum_{k,j}^{n} \overline{\lambda}_k \lambda_j \exp\frac{i}{2} B(F_k,F_j) \times \exp\frac{-1}{4}(F_j - F_k)_i^2 \geqslant 0 \qquad (*)$$

$\forall \lambda_k, \forall F_k \in \mathcal{K}$. Note that this is the twisted positivity condition for ρ_i, which is usually employed to characterize generating functions of the C^*-algebra of the CCR.[18] If $(\cdot)_i$ were an ordinary norm on $\mathcal{K}$, ρ_i would be the generating function of an ordinary Fock state. The argument of the exponential in $(**)$ is

$$\begin{aligned}
&(i/2)\, B(F_k,F_j) - \tfrac{1}{4}(F_j - F_k)_i^2 \\
&\quad = (i/2)B(F_k,F_j) - \tfrac{1}{4}(F_k - F_j, F_k - F_j)_i \\
&\quad = (i/2)B(F_k,F_j) - \tfrac{1}{4}[(F_k)_i^2 + (F_j)_i^2 - 2\,\mathrm{Re}(F_k,F_j)_i] \\
&\quad = -\tfrac{1}{4}(F_k)_i^2 - \tfrac{1}{4}(F_j)_i^2 + \tfrac{1}{2}(F_k,F_j)_i ,
\end{aligned}$$

since $B(\cdot,\cdot) = \mathrm{Im}(\cdot,\cdot)_i$. Hence we only need to show that $\exp(F_k,F_j)_i$ is a positive definite kernel on $\mathcal{K} \times \mathcal{K}$ because λ_k is arbitrary, and so can be redefined as

$$\lambda_k \to \lambda_k \exp(-\tfrac{1}{4})(F_k)_i^2 .$$

Now $\exp \tfrac{1}{2}(F_k,F_j)_i$ is positive definite iff $(F_k,F_j)_i$ is (cf. Ref. 23, Theorem 2.2 p. 74), and $(\cdot,\cdot)_i$ is certainly a positive definite kernel on $\mathcal{K} \times \mathcal{K}$. Thus $\omega_i(\mathcal{O}_+) \geqslant 0$. So it follows that ω_i is a state because $\omega_i(\mathbf{1}) = \omega_i(\delta_0) = 1$. To show $\mathcal{D} \subset \mathrm{Ker}\,\omega_i$, it suffices by the general theory of Sec. II to show that $\omega_i(\delta_C) = 1 \forall C \in \mathcal{C}$, and this follows directly from $(C,C)_i = 0 \forall C \in \mathcal{C}$ which is given. The last two assertions of the theorem follow from

$$\begin{aligned}
\omega_i(\alpha_g(\delta_F)) &= \omega_i(\delta_{T_g F}) \\
&= \exp[-\tfrac{1}{4}(T_g F, T_g F)_i] \\
&= \exp[-\tfrac{1}{4}(F,F)_i] = \omega_i(\delta_F) ,
\end{aligned}$$

and likewise

$$\omega_i(\beta\alpha_g(\delta_F)) = \omega(\delta_{ST_g F}) = \omega_i(\delta_F) . \qquad \blacksquare$$

Hence on the physically relevant part of $\mathcal{F}$, $C^*(\delta_{\mathcal{K}})$, we have obtained a perfectly acceptable G-invariant Dirac state ω_i from ρ_i. The structures on the nonphysical part, $\mathcal{F} \setminus C^*(\delta_{\mathcal{K}})$, can be altered according to convenience without affecting the physics. In what follows, we discard the information from the irregular part of ρ_i, i.e., $\rho_i(F) \forall F \notin \mathcal{K}$, by extending the Dirac state ω_i to a G-invariant Hermitian functional f_i on $\mathcal{F}$. The GNS representation of f_i will define a SW structure according to the general theorems of the preceding section (e.g., 3.11). To do this we need the following lemmas.

Lemma 4.4: Given C^*-algebras $\mathscr{A}\subset\mathscr{B}$, and a set $\mathscr{L}\subset\mathscr{B}$, a continuous linear functional on $\mathscr{A}$, $f\in\mathscr{A}^*$ is extendible to a continuous linear functional $\tilde{f}$ on $\mathscr{B}$ with $\mathscr{L}\subset\mathrm{Ker}\,\tilde{f}$ iff,

(i) $[\mathscr{L}]\cap\mathscr{A}\subset\mathrm{Ker}\,f$, and

(ii) $\exists E\in\mathscr{A}\setminus\mathrm{Ker}\,f$ such that $E\notin[\mathscr{L}\cup\mathscr{K}_E]$

where

$$\mathscr{K}_E:=\{A-f(A)[E/f(E)]\,|A\in\mathscr{A}\}\,.$$

If f is Hermitian, then $\tilde{f}$ can be chosen to be Hermitian too.

Proof: Assume $\exists\tilde{f}\in\mathscr{B}^*$ such that $\tilde{f}|\mathscr{A}=:f\in\mathscr{A}^*$ is nontrivial and $\mathscr{L}\subset\mathrm{Ker}\,\tilde{f}$. Now $\mathrm{Ker}\,\tilde{f}$ is a closed linear space, and $\mathrm{Ker}\,f=\mathrm{Ker}\,\tilde{f}\cap\mathscr{A}$, hence $[\mathscr{L}]\cap\mathscr{A}\subset\mathrm{Ker}\,f$. Since f is nontrivial, $\mathscr{A}\setminus\mathrm{Ker}\,f\neq\emptyset$. Choose any $E\in\mathscr{A}\mathrm{Ker}\,f$, then $\mathscr{K}_E\subset\mathrm{Ker}\,f$ because $\forall A\in\mathscr{A}$:

$$0=f(A)-f(A)=f(A)-f(A)[f(E)/f(E)]$$
$$=f(A-f(A)[E/f(E)])\,.$$

Hence $[\mathscr{L}\cup\mathscr{K}_E]\subseteq\mathrm{Ker}\,\tilde{f}$. From $f=\tilde{f}|\mathscr{A}$ we see that $E\notin\mathrm{Ker}\,\tilde{f}$, and so $E\notin[\mathscr{L}\cup\mathscr{K}_E]$.

Conversely, assume $\exists f\in\mathscr{A}^*$, a set $\mathscr{L}\subset\mathscr{B}$ and an $E\in\mathscr{A}\setminus\mathrm{Ker}\,f$ such that $[\mathscr{L}]\cap\mathscr{A}\subset\mathrm{Ker}\,f$, and $E\notin\mathscr{L}_E:=[\mathscr{L}\cup\mathscr{K}_E]$. Then we wish to show that $\exists\tilde{f}\in\mathscr{B}^*$ such that $\mathscr{L}\subset\mathrm{Ker}\,f$ and $\tilde{f}|\mathscr{A}=f$. Now $E\notin\mathscr{L}_E$ implies that $\mathscr{L}_E\neq\mathscr{B}$, and hence the normed linear space $\mathscr{B}/\mathscr{L}_E$ is nontrival with the norm defined by $\|\xi_A\|:=\inf\{\|A+L\|\,|L\in\mathscr{L}_E\}$ where $\xi:\mathscr{B}\mapsto\mathscr{B}/\mathscr{L}_E$ is the canonical map. Then there is a bijection between the continuous linear functionals on $\mathscr{B}/\mathscr{L}_E$, and the continuous linear functionals on $\mathscr{B}$ with $\mathscr{L}_E$ in their kernels. Since $E\notin\mathscr{L}_E$, there are functionals $h\in(\mathscr{B}/\mathscr{L}_E)^*$ for which $h(\xi_E)\neq0$, and these can be normalized to get $h(\xi_E)=f(E)$. Hence there are functionals $\tilde{f}\in\mathscr{B}^*$ with $\mathscr{L}_E\subset\mathrm{Ker}\,\tilde{f}$ and $\tilde{f}(E)=f(E)$. So $\mathscr{L}\subset\mathscr{L}_E\subset\mathrm{Ker}\,\tilde{f}$, and $\mathscr{K}_E\subset\mathscr{L}_E\subset\mathrm{Ker}\,\tilde{f}$, i.e., $\forall A\in\mathscr{A}$:

$$0=\tilde{f}(A-f(A)[E/f(E)])$$
$$=\tilde{f}(A)-f(A)[\tilde{f}(E)/f(E)]=\tilde{f}(A)-f(A)\,,$$

that is, $\tilde{f}(A)=f(A)\ \ \forall A\in\mathscr{A}$, i.e., $\tilde{f}|\mathscr{A}=A$.

If f is Hermitian, we restrict our attention to real-valued functionals on the real linear space corresponding to the self-adjoint elements of the C^*-algebras. The argument above then carries over in a direct fashion. ■

Remark: If $f\in\mathscr{A}^*$ is positive, it may not be possible to get a positive extension $\tilde{f}\in\mathscr{B}^*$ with $\mathscr{L}\subset\mathrm{Ker}\,\tilde{f}$. As an example to see this, we use Theorem 2.4. That is, specify a constraint set in $\mathscr{B}\setminus\mathscr{A}$ such that the linear space generated by it has only the trivial intersection $\{0\}$ with $\mathscr{A}$, but which does not satisfy the nontriviality condition of 2.4. Then there are no states vanishing on the constraints in $\mathscr{B}$.

For a C^*-algebra $\mathscr{F}$ with a group action $\alpha: G\mapsto\mathrm{Aut}\,\mathscr{F}$ on it, denote

$$(\alpha_G-\iota)\mathscr{F}:=\{\alpha_g(F)-F\,|g\in G,F\in\mathscr{F}\}\,.$$

Then a functional $f\in\mathscr{F}^*$ is G invariant iff $[(\alpha_G-\iota)\mathscr{F}]\subset\mathrm{Ker}\,f$.

Corollary 4.5: Assume two C^*-algebras $\mathscr{A}\subset\mathscr{B}$, and an $\mathscr{A}$-preserving group action $\alpha: G\mapsto\mathrm{Aut}\,\mathscr{B}$. Then all G-invariant Hermitian functional on $\mathscr{A}$ can be extended to G-invariant Hermitian functionals on $\mathscr{B}$ if

$$[(\alpha_G-\iota)\mathscr{B}]\cap\mathscr{A}=[(\alpha_G-\iota)\mathscr{A}]\,. \qquad (***)$$

Proof: We wish to apply 4.4. Let f be a nontrivial G-invariant functional on $\mathscr{A}$. Then $\forall E\in\mathscr{A}\setminus\mathrm{Ker}\,f$ we have $E\notin[(\alpha_G-\iota)\mathscr{A}\cup\mathscr{K}_E]\subset\mathrm{Ker}\,f$, by arguments above. Assume also that the condition (***) of the theorem holds. In the notation of Lemma 4.4, take $\mathscr{L}=[(\alpha_G-\iota)\mathscr{B}]$. Then $[\mathscr{L}]\cap\mathscr{A}=[(\alpha_G-\iota)\mathscr{A}]\subset\mathrm{Ker}\,f$, and so Lemma 4.4(i) is satisfied. For Lemma 4.4(ii), note that $\mathscr{K}_E\subset\mathscr{A}\ \forall E\in\mathscr{A}\setminus\mathrm{Ker}\,f$. Since $\mathscr{A}$ is a closed linear space, we can verify that $[\mathscr{L}\cup\mathscr{K}_E]\cap\mathscr{A}=[([\mathscr{L}]\cap\mathscr{A})\cup\mathscr{K}_E]$. Hence we see from $E\notin[\mathscr{L}\cup\mathscr{K}_E]$ iff $E\notin[\mathscr{L}\cup\mathscr{K}_E]\cap\mathscr{A}$, that it is only necessary to show that $\exists E\in\mathscr{A}\setminus\mathrm{Ker}\,f$ such that $E\notin[(\alpha_G-\iota)\mathscr{A}\cup\mathscr{K}_E]$. But as we saw, all $E\in\mathscr{A}\setminus\mathrm{Ker}\,f\neq\emptyset$ will satisfy this. ■

Now for a constrained boson theory as above, we have $\mathscr{B}=\mathscr{F}=\overline{\Delta(\mathscr{Q})}$, $\mathscr{A}=C^*(\delta_{\not h})$, and $\alpha_g(\delta_F)=\delta_{T_gF}$, and $T_G\not h=\not h$. To apply the preceding corollary, we need to show that

$$[(\alpha_G-\iota)\,\overline{\Delta(\mathscr{Q})}]\cap C^*(\delta_{\not h})=[(\alpha_G-\iota)C^*(\delta_{\not h})]\,.$$

Since $(\alpha_g-\iota)$ is linear and continuous,

$$[(\alpha_G-\iota)\,\overline{\Delta(\mathscr{Q})}]=[(\alpha_G-\iota)\delta_{\mathscr{Q}}]$$

and

$$[(\alpha_G-\iota)C^*(\delta_{\not h})]=[(\alpha_G-\iota)\delta_{\not h}]\,.$$

A general element $A\in[(\alpha_G-\iota)\delta_{\mathscr{Q}}]$ is the limit of a Cauchy sequence of the form: $\{\Sigma_{i=1}^{n_j}\gamma_i^j(\delta_{T_{g_i}F_i}-\delta_{F_i})\}_{j=1}^{\infty}$. The index i here has taken into account all relevant elements of $\mathscr{Q}\times G$. Now in the proof of Theorem 4.2, we showed that given a Cauchy sequence $\{\Sigma_i^{n_j}\lambda_i^j\delta_{F_i}\}_{j=1}^{\infty}$ converging to an element in $C^*(\delta_{\not h})$, that of the two Cauchy sequences obtained from its natural decomposition with respect to $\not h$:

$$\sum_i^{n_j}\lambda_i^j\delta_{F_i}=\sum_i^{n_j}\alpha_i^j\delta_{P_i}+\sum_i^{n_j}\beta_i^j\delta_{T_i}\,,\quad P_i\in\not h\ni T_i\,,$$

the second one converges to zero: $\{\Sigma_i^{n_j}\beta_i^j\delta_{T_i}\}\to0$ as $j\to\infty$. Hence if $A\in[(\alpha_G-\iota)\delta_{\mathscr{Q}}]\cap C^*(\delta_{\not h})$, then we decompose its converging sequence in the same manner. Since $T_G\not h=\not h$, we get that $(\delta_{T_gF}-\delta_F)\in\Delta(\not h)$ iff $F\in\not h$. So with the decomposition

$$\sum_i^{n_j}\gamma_i^j(\delta_{T_{g_i}F_i}-\delta_{F_i})=\sum_i^{n_j}\alpha_i^j(\delta_{T_{g_i}P_i}-\delta_{P_i})$$
$$+\sum_i^{n_j}\beta_i^j(\delta_{T_{g_i}T_i}-\delta_{T_i})\,,$$

where $P_i\in\not h\ni T_i$, we get that the limit of $\Sigma_i^{n_j}\alpha_i^j(\delta_{T_{g_i}P_i}-\delta_{P_i})$ is also A. Hence $[(\alpha_G-\iota)\delta_{\mathscr{Q}}]\cap C^*(\delta_{\not h})=[(\alpha_G-\iota)\delta_{\not h}]$, and so all G-invariant Hermitian functionals on $\mathscr{O}_c=C^*(\delta_{\not h})$ can be extended to G-invariant Hermitian functionals on $\mathscr{F}=\overline{\Delta(\mathscr{Q})}$.

Proposition 4.6: Given a degenerate boson field as above, and IIP $(\cdot,\cdot)_i$ on $\mathscr{Q}$ which satisfies $(F,F)_i\geqslant0\ \ \forall F\in\not h$ and $(C,C)_i=0\ \ \forall C\in\mathscr{C}$, it will define an SW structure if in addition $(T_gF,T_gH)_i=(F,H)_i\ \ \forall F,H\in\not h$, $g\in G$, and it will de-

fine a wSW structure if $(ST_gF,\ ST_gH)_i = (F,H)_i$ $\forall F,H\in\mathscr{h}$, $g\in G$, where the symplectic transformation S defines a gauge automorphism on $\mathscr{F}$.

Proof: This is simply obtained by constructing ω_i by 4.3, extending it to f_i by the remarks following 4.5, and applying 3.11 to f_i for a SW structure. (Put $x_0 = 1$, $J = \mathscr{F}$, $F = 1$.) For a wSW structure, the extension lemma above adapts easily, and we obtain the result on application of 3.17. ■

It is possible to define IIP's $(\cdot,\cdot)_i$ by employing the operators $J_i\in S(\mathscr{Q},B)$ which satisfy $J_i^2 = -I$. This is done as follows: $(F,H)_i := B(F,J_iH) + iB(F,H)$. By an extension of the usual nomenclature, J_i is called a complex structure.

The first example of an application of the theory above that we present is that of a one-dimensional scalar boson constrained to live only in a periodic set of intervals. Hence we make the following choices:

$$\mathscr{Q} = \mathscr{S}^2(\mathbb{R}),\quad B(F,G) = \int (F_1G_2 - F_2G_1)\,,$$

$$\mathscr{F} = \overline{\Delta(\mathscr{Q})}\,,$$

where $F = (F_1,F_2)$. In order to define the constraint set $\mathscr{C}$, define the intervals $I_n := [2nb,(2n+1)b]\subset\mathbb{R}$, $b\in\mathbb{R}^+$ fixed, $n\in\mathbb{Z}$, $I := \cup_{n\in\mathbb{Z}_+} I_n$ and the set

$$\mathscr{M} := \{f\in\mathscr{S}(\mathbb{R})\,|\,\exists n\in\mathbb{Z} \text{ such that } \operatorname{supp} f\subset I_n\}\,.$$

Then the constraint set is defined as $\mathscr{C} = (0,\mathscr{M})$, so that one finds $\mathscr{h} = (\mathscr{M}^\perp,\mathscr{S}(\mathbb{R}))$, where

$$\mathscr{M}^\perp := \{f\in\mathscr{S}(\mathbb{R})\,|\operatorname{supp} f\notin I_n\quad \forall n\in\mathbb{Z}^+\}\,.$$

Define physical transformations as the jumps between admissible intervals, but to the same relative position in the new interval:

$$d_kF := (F_1(x+2kb), F_2(x+2kb)),\quad k\in\mathbb{Z}\,.$$

These transformations are clearly symplectic, preserve $\mathscr{C}$, and hence are physical. In order to demonstrate the existence of an SW structure for this model, we apply Proposition 4.6, i.e., we wish to find a complex structure J which satisfies the conditions:

$$B(F,JG) = B(G,JF)\quad \forall F,G\in\mathscr{Q};\quad J^2 = -1,$$

$$B(F,JC) = 0\forall F\in\mathscr{h}\quad \forall C\in\mathscr{C};\quad B(F,JF)\geqslant 0\quad \forall F\in\mathscr{h},$$

$$B(F,JF) = B(d_kF,Jd_kF)\quad \forall F\in\mathscr{Q}\,.$$

With the definition

$$P(x) := \begin{cases} 1, & \text{if } x\in I\,;\\ 0, & \text{if } x\notin I\,,\end{cases}$$

the complex structure

$$J = \begin{pmatrix} iP & P-1\\ 1-P & -iP\end{pmatrix}$$

will satisfy all the requirements. As the verification is straightforward, we leave these to the reader.

As a final example, we present electromagnetism in the Landau gauge. Rideau[24] has already constructed a wSW structure in the heuristic framework, and the algebraic part of the theory has been cast into exact C^*-algebraic language by Carey and Hurst.[19] Very little additional work is required to fit it into the present framework, apart from identifying the corresponding concepts. We present some of the details.

As in Ref. 19, the field algebra $\mathscr{F}_L$ for the Landau gauge is set up over momentum space. Let $\mathscr{S} := \{\mathbb{C}^4\text{-valued, } C^\infty\text{-functions of fast decrease on } \mathbb{R}^4\}$. The Fermi symplectic form is

$$B(f,g) := \int d^4k\,\delta^+(k^2)\,[f^\mu\bar{g}_\mu - \bar{f}^\mu g_\mu]\,,$$

where $\delta^+(k^2) := \delta(k_0 - |\underline{k}|)/2|\underline{k}|$. Let

$$C_+ := \{k\in\mathbb{R}^4\,|\,k^2 = 0,\ k_0 > 0\}$$

be the positive frequency light cone without the origin. In order to make B nondegenerate on $\mathscr{S}$, one normally factors out the off-C_+ parts of the elements of $\mathscr{S}$. The Landau symplectic form is defined as

$$B_L(f,f') := B(Zf,Zf')\,,$$

where

$$(Zf)_\mu(k) := f_\mu(k) + \frac{k_\mu}{2k_0}\left[\frac{k_\nu f^\nu(k)}{2k_0} - \frac{\partial}{\partial k_0}(k^\nu f_\nu(k))\right].$$

This corresponds to the usual heuristic expression.[24] Clearly due to the last term, B_L involves more of the Cauchy data than does B. In order to make B_L nondegenerate, define

$$\varphi\colon f\mapsto\varphi^f_\mu(\underline{k}) := (Zf)_\mu|C_+\,.$$

Then the test function space on which we choose to set up the C^*-algebra of the CCR's is $\mathscr{Q} = \mathscr{S}/\operatorname{Ker}\varphi$, and B_L is nondegenerate on $\mathscr{Q}$. Hence take $\mathscr{F}_L := \overline{\Delta(\mathscr{Q},B_L)}$ as in Manuceau.[20] As smearing is done as usual, the elements $f_\mu(k) = k_\mu f(k)\in\mathscr{S}$ will correspond to $\partial^\mu A_\mu(x)$, and so since one finds that $k_\mu f(k)\in\operatorname{Ker}\varphi\quad \forall f(k)$, the transversality condition $\partial^\mu A_\mu(x) = 0$ holds as an operator condition on $\mathscr{F}_L$. This is one example of how to treat an algebraic condition (cf. Sec. II).

In order to define a C^*-degenerate system in $\mathscr{F}_L$, note that in Ref. 2 the Maxwell equations are imposed as state conditions. Hence we would prefer our specified constraint set to contain these. The set of test functions

$$\{(k_\mu k^\nu - k^2\delta_{\mu\nu})f_\nu(k)\,|\,f\in\mathscr{S}\}$$

will correspond to $F_{\mu\nu}{}^{,\nu}(x)$ after smearing. Now

$$Z(k_\mu k^\nu - k^2\delta_{\mu\nu})f_\nu(k)|C_+ = -Zk^2f_\mu|C_+ = k_\mu k^\nu f_\nu|C_+.$$

Hence, following the literature, we take as constraints all equivalence classes of functions $f\in\mathscr{S}$ such that $Zf|C_+ = \varphi^f_\mu(\underline{k}) = k_\mu\xi(\underline{k})$, where $k_0 = |\underline{k}|$. Call this set $\mathscr{C}$. Then the observables $\mathscr{O}$ will be generated by

$$\mathscr{h} := \{f\in\mathscr{Q}\,|B_L(f,f') = 0\quad \forall f'\in\mathscr{C}\}\,,$$

$$B_L(f,f') = B(Zf,k_\mu\xi(\underline{k}))$$

$$= \operatorname{Im}\int_{C_+}\frac{d^3k}{2|\underline{k}|}\,[\varphi^f_\mu(\underline{k})k^\mu\bar{\xi}(\underline{k})] = 0\,,$$

for all functions $\xi(\underline{k})$, and this will be the case if $\varphi^f_\mu(\underline{k})k^\mu = 0$, $k_0 = |\underline{k}|$. Take the set of equivalence classes of such f's to be $\mathscr{h}$. By the definition of Z, $f\in\mathscr{h}$ iff $k_\mu f^\mu|C_+ = 0$. Clearly $\mathscr{C}\subset\mathscr{h}$. In Ref. 2 Strocchi and Wightman required that $F_{\mu\nu}\in\mathscr{O}$. Let $f_{\mu\nu}$ be an antisymmetric tensor function. Then $2k_\nu f^{\mu\nu}$ corresponds to $F^{\mu\nu}$, i.e., $F(f) = A(2k_\nu f^{\mu\nu})$. Then $(Z\,2k_\nu f^{\mu\nu})^\lambda k_\lambda|C_+ = 0$ follows

easily, which verifies $F_{\mu\nu}\in\mathscr{O}$. Through the definition of Z one also gets that if $f\in\mathscr{C}$, then $f_\mu(\underline{k}) = k_\mu\gamma(\underline{k})|C_+$ for a specific $\gamma(\underline{k})$, related to $\xi(\underline{k})$.

Now that $\mathscr{C}$ and $\not{\mathscr{h}}$ are specified, and contain the right objects, the Poincaré transformations still need to be defined. These are given on $\mathscr{D}$ as

$$(U_g f)(k) = e^{ia\cdot k}\Lambda f(\Lambda^{-1}k)\,,$$

where $g = L(\Lambda,a)$ is in the orthochronous Poincaré group. These transformations translate on the φ^f's using $(U'_g\varphi^f)_\mu(\underline{k}) = \varphi^{U_g f}_\mu(\underline{k})$ to[24]

$$U'_g\varphi^f_\mu(\underline{k}) = e^{ia\cdot k}\left\{\Lambda_\mu{}^\nu\varphi^f_\nu(\underline{\Lambda^{-1}k}) - \frac{k_\mu}{2|\underline{k}|}\right.$$
$$\left.\times\left[\Lambda_i{}^0\left(\frac{\partial\omega^f}{\partial k_i}\right)(\underline{\Lambda^{-1}k}) + ia_0\omega^f(\underline{\Lambda^{-1}k})\right]\right\},$$

where $k_0 = |\underline{k}|$, and $\omega^f(\underline{k}) := k^\mu\varphi^f_\mu(\underline{k}) = k^\mu f_\mu(\underline{k})$. One easily checks that U_g respects Ker φ, hence is defined on $\mathscr{D}$, and that it is symplectic. Moreover, if $\varphi^f_\mu(\underline{k}) = k_\mu\xi(\underline{k})$, then $\varphi^{U_g f}_\mu(\underline{k}) = k_\mu e^{ia\cdot k}\xi(\underline{\Lambda^{-1}k})$, i.e., U_g preserves $\mathscr{C}$, i.e., $\alpha_g\in\Upsilon$, where α_g is the automorphism on $\mathscr{F}_L$ defined by U_g.

Now in order to find the structures according to Proposition 4.6, it is simple just to adapt the existing structures in Ref. 24 to this context. The rigorous existence of a wSW structure for the Landau gauge will be demonstrated if we can exhibit an indefinite inner product $\langle\cdot,\cdot\rangle$ on $\mathscr{D}$ such that $\langle T_\beta U_g f, T_\beta U_g f'\rangle = \langle f,f'\rangle \quad \forall f,f'\in\mathscr{D}$, where T_β is a symplectic transformation defining a gauge transformation β, and

$$\langle f,C\rangle = 0 \quad \forall f\in\not{\mathscr{h}},\quad C\in\mathscr{C};\quad \langle f,f\rangle\geqslant 0 \quad \forall f\in\not{\mathscr{h}}\,.$$

Such an inner product is given by[24]

$$\langle f,f'\rangle = \int\frac{d^3k}{|\underline{k}|}\left[-f_\mu(\underline{k})\,\overline{f'^\mu(\underline{k})} - \frac{\omega^f(\underline{k})\,\overline{\omega^{f'}(\underline{k})}}{2|\underline{k}|^2}\right],$$

which is non-negative if $f=f'$ and $k^\mu f_\mu|C_+ = 0$ because spacelike photons are not admitted into the theory, and it is zero if $k^\mu f_\mu|C_+ = 0$ and $f'_\mu(\underline{k}) = k_\mu\xi(\underline{k})$. Now $\langle\cdot,\cdot\rangle$ can be derived from the Poincaré invariant inner product $-\int(d^3k/|\underline{k}|)f_\mu(\underline{k})\,\overline{f'^\mu(\underline{k})}$ by the gauge transformation

$$f_\mu(\underline{k})\to f_\mu(\underline{k}) + k_\mu k^\nu f_\nu(\underline{k})/4k_0^2\,.$$

As the invariant inner product is evaluated on C_+, this is not required in the gauge transformation. One easily verifies that the gauge transformation is symplectic, and its associated automorphism on $\mathscr{F}$ is in Ker T_c. Hence $\langle\cdot,\cdot\rangle$ is Poincaré invariant up to this gauge transformation. So, on considering the relevant expressions above, we find that we have shown the existence of a wSW structure.

V. CONCLUDING REMARKS

While the structure of IIP theories are modelled on the Gupta–Bleuler version of electromagnetism, the current theory does not as yet have enough machinery to deal with it. The reason for this is that the constraints χ used in Gupta–Bleuler electromagnetism are non-Hermitian, and it is therefore not possible to define elements in the linear field algebra corresponding to $\exp(i\lambda\chi)$, neither can we use the Hermitian parts of χ. On the other hand, the logical choice $\exp(i\lambda\chi^*\chi)$ for the constraint is nonlinear in its argument, and hence cannot be defined as an element in the linear field algebra. It is therefore necessary to develop a general theory of *outer* constraints, i.e., the constraints imposed are not contained in the field algebra in contrast to the situation above. This problem has already been solved separately, and will be submitted for publication soon. The theory above shifted neatly into place, and provided an acceptable C^*-field theory of the Gupta–Bleuler situation.

[1] H. B. G. S. Grundling and C. A. Hurst, Commun. Math. Phys. **98**, 369 (1985).
[2] F. Strocchi and A. S. Wightman, J. Math. Phys. **15**, 2198 (1974); **17**, 1930 (1976).
[3] F. Strocchi, Phys. Rev. D **17**, 2010 (1978).
[4] G. Morchio and F. Strocchi, Ann. Inst. H. Poincaré **33**, 251 (1980).
[5] S. N. Gupta, Proc. Phys. Soc. London Sec. A **63**, 681 (1950); K. Bleuler, Helv. Phys. Acta **23**, 567 (1950).
[6] P. A. M. Dirac, *Lectures in Quantum Mechanics* (Yeshiva U. P., New York, 1964); Can. J. Math. **2**, 129 (1950); **3**, 1 (1951), and also E. C. G. Sudarshan and N. Mukunda, *Classical Dynamics: A Modern Perspective* (Wiley, New York, 1974).
[7] J. Yngvason, Rep. Math. Phys. **12**, 57 (1977); Commun. Math. Phys. **34**, 315 (1973); F. Strocchi, *ibid.* **56**, 57 (1977).
[8] L. Jakobczyk, J. Math. Phys. **25**, 617 (1984).
[9] D. G. Boulware and D. J. Gross, Nucl. Phys. B **233**, 1 (1984).
[10] M. Mintchev, J. Phys. A **13**, 1841 (1980).
[11] K. Yu. Dadashyan and S. S. Khoruzhii, Teor. Mat. Fiz. **54**, 57 (1983).
[12] L. Jakobczyk, J. Math. Phys. **27**, 116 (1986).
[13] J. Bognar, *Indefinite Inner Product Spaces* (Springer, Berlin, 1974), see p. 89.
[14] L. Jakobczyk, Ann. Phys. **161**, 314 (1985).
[15] H. Araki, Commun. Math. Phys. **97**, 149 (1985).
[16] I. E. Segal, *Mathematical Problems of Relativistic Physics* (Am. Math. Soc., Providence, RI, 1963).
[17] H. B. G. S. Grundling, Ph.D thesis, University of Adelaide, 1986.
[18] G. G. Emch, *Algebraic Methods in Statistical Mechanics and Quantum Field Theory* (Wiley, New York, 1972).
[19] A. L. Carey and C. A. Hurst, Lett. Math. Phys. **2**, 227 (1978).
[20] J. Manuceau, Ann. Inst. H. Poincaré **8**, 139 (1968).
[21] A. L. Carey, J. M. Gaffney, and C. A. Hurst, J. Math. Phys. **18**, 629 (1977).
[22] P. Broadbridge, Ph.D. thesis, University of Adelaide, 1982; and also with C. A. Hurst, Ann. Phys. (NY) **131**, 104 (1981).
[23] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups* (Springer, New York, 1984).
[24] G. Rideau, Lett. Math. Phys. **1**, 17 (1975).

Generalized Bergman kernels and geometric quantization

G. M. Tuynman
Mathematisch Instituut, Roetersstraat 15, NL 1018 WB Amsterdam, The Netherlands

(Received 22 May 1986; accepted for publication 5 November 1986)

In geometric quantization it is well known that, if f is an observable and F a polarization on a symplectic manifold (M,ω), then the condition "X_f leaves F invariant" (where X_f denotes the Hamiltonian vector field associated to f) is sufficient to guarantee that one does not have to compute the BKS kernel explicitly in order to know the corresponding quantum operator. It is shown in this paper that this condition on f can be weakened to "X_f leaves $F+F^\dagger$ invariant"and the corresponding quantum operator is then given implicitly by formula (4.8); in particular when F is a (positive) Kähler polarization, all observables can be quantized "directly" and moreover, an "explicit" formula for the corresponding quantum operator is derived (Theorem 5.8). Applying this to the phase space $\mathbb{R}^{2n}$ one obtains a quantization prescription which resembles the normal ordering of operators in quantum field theory. When we translate this prescription to the usual position representation of quantum mechanics, the result is (among others) that the operator associated to a classical potential is multiplication by a function which is essentially the convolution of the potential function with a Gaussian function of width $\hbar$, instead of multiplication by the potential itself.

I. PRELIMINARIES

In Secs. I–III we give a brief summary of geometric quantization, mainly to fix the notation; for more details the reader is referred to Refs. 1–3.

Suppose (M,ω) is a symplectic manifold; denote by X_f the Hamiltonian vector field associated to the function f: $M\to\mathbb{R}$, defined by $i_{X_f}\omega + df = 0$. Let $L\to M$ be the prequantization line bundle over M with connection ∇ and compatible Hermitian form $(\ ,\ )$ such that $\mathrm{curv}(\nabla) = \omega/\hbar$ [we suppose that (M,ω) satisfies the quantization condition].

Let F be a (positive) polarization, i.e., F is a complex distribution of constant complex dimension $n = \frac{1}{2}\dim(M)$ satisfying

(i) $\exists k\in\mathbb{N}$, $0\leqslant k\leqslant n$: $\dim_{\mathbb{C}}(F\cap F^\dagger) = k$ († denotes complex conjugation),

(ii) $\forall m\in M$ there exists a neighborhood U such that

(a) $\exists\ z^1,\ldots,z^n$: $U\to\mathbb{C}$: $X_{z^1},\ldots,X_n$ span F on U, and $\{z^i,z^j\} = 0$ ($\{\ ,\ \}$ denotes the Poisson bracket),

(b) $\exists\ w^1,\ldots,w^k$: $U\to\mathbb{C}$: $X_{w^1},\ldots,X_{w^k}$ span $F\cap F^\dagger$ on U,

(iii) $\forall\ v\in F_m$: $i\cdot\omega(v,v^\dagger)\geqslant 0$ (positivity).

If F is a polarization then there exists by definition a real foliation D such that $D^{\mathbb{C}} = F\cap F^\dagger$; it follows that there exists another real foliation E such that $E^{\mathbb{C}} = F + F^\dagger$, $E^\perp = D$, $D^\perp = E$ (orthoplement with respect to ω). We suppose that the quotient space M/D admits a manifold structure such that the canonical projection $\pi: M\to M/D$ is a submersion. The image $\pi_* E$ in M/D is a foliation of even dimension and F induces a complex structure on the leaves of $\pi_* E$ such that the following description holds: $X_z\in F\Leftrightarrow z$ is a function on M/D, holomorphic on the leaves of $\pi_* E$.

Define R^F to be the principal $\mathrm{GL}(n,\mathbb{C})$ bundle over M of all F frames and suppose there exists a principal $\mathrm{ML}(n,\mathbb{C})$ bundle $R^{\sim F}$ over M and a 2–1 bundle covering pr of $R^{\sim F}$ over R^F, where $\mathrm{ML}(n,\mathbb{C})$ is the metalinear group with projection p: $\mathrm{ML}(n,\mathbb{C})\to\mathrm{GL}(n,\mathbb{C})$; we denote by λ: $\mathrm{ML}(n,\mathbb{C})\to\mathbb{C}$ the well-defined map which represents the "square root of det," i.e., $\lambda(g) = \pm\sqrt{\det(p(g))}$. Under the assumption that $R^{\sim F}$ exists, $B^{\sim F}$ will denote the bundle of $(-\frac{1}{2})$-F-forms, which is the $\mathbb{C}$-line bundle over M associated with the principal bundle $R^{\sim F}$ by the representation λ. The sections ν of $B^{\sim F}$ can be identified with functions ν on $R^{\sim F}$ with the transformation property

$$\nu(a\cdot g) = \nu(a)\cdot\lambda(g)^{-1} = \pm\nu(a)\cdot(\sqrt{\det(p(g))})^{-1},$$
$$\text{for } a\in R^{\sim F},\ g\in\mathrm{ML}(n,\mathbb{C}),\ p(g)\in\mathrm{GL}(n,\mathbb{C}). \tag{1.1}$$

On $B^{\sim F}$ there exists a partial (i.e., defined for vectors $v\in F$ only) flat connection ∇, and with these ingredients we define the quantum bundle (QB) with partial connection ∇ by

$$\mathrm{QB} = L\otimes B^{\sim F},\quad \text{“}\nabla = \nabla|_L + \nabla|_{B^{\sim F}}\text{.”}$$

By an *F-constant section* Ψ of QB we will mean a section which satisfies $\forall\xi\in F$: $\nabla_\xi\Psi = 0$; the same convention holds for sections of L and $B^{\sim F}$.

Since the Hilbert space constructed by geometric quantization consists of F-constant sections of QB, we will investigate these sections in more detail; in Sec. III the Hilbert space and the inner product will be defined more precisely.

By using refinements of covers one can always construct a cover $\{U_\alpha\,|\,\alpha\in A\}$ of M satisfying the following conditions.

(i) It trivializes the bundles, L, R^F, $R^{\sim F}$, $B^{\sim F}$ and QB simultaneously.

(ii) On U_α there exists a symplectic potential ϑ_α: $d\vartheta_\alpha = \omega$ and the connection ∇ on L is given by

$$\nabla_\xi s_\alpha = \xi s_\alpha - (i/\hbar)\vartheta_\alpha(\xi)s_\alpha, \tag{1.2}$$

where the function s_α represents a local section of L over U_α (this follows from the construction of L; the transition functions of L are related to the exact one-forms $\vartheta_\alpha - \vartheta_\beta$).

(iii) There exist $r^1,\ldots,r^k$: $U_\alpha\to\mathbb{R}$ and $z^{k+1},\ldots,z^n$: $U_\alpha\to\mathbb{C}$ such that $X_{r^1},\ldots,X_{r^k}$ span D,

 0022-2488/87/030573-11$02.50

$X_{r^1},...,X_{r^k},X_{z^{k+1}},...,X_{z^n} \equiv X_{r^1},...,X_{z^n}$ span F and the frame $(X_{r^1},...,X_{z^n})$ corresponds to the identity $\in GL(n,\mathbb{C})$ in the local trivialization of R^F (the existence of these functions follows from the definition of a polarization).

The elements of $R^{\sim F}$ are called metaframes, so to each F frame $(\xi_1,...\xi_n)$ correspond two metaframes called $(\xi_1,...,\xi_n)^\sim$ (remember that pr: $R^{\sim F} \to R^F$ is a 2–1 covering). Whenever we need to define exactly which metaframe we have to use, we will specify it; in particular the metaframe $(X_{r^1},...,X_{z^n})^\sim$ represents the identity in $ML(n,\mathbb{C})$ in the same way as $\mathrm{pr}((X_{r^1},...,X_{z^n})^\sim) \equiv (X_{r^1},...,X_{z^n})$ represents the identity in $GL(n,\mathbb{C})$ [see condition (iii)].

We will call a cover $\{U_\alpha | \alpha \in A\}$ a *nice cover* if it also satisfies a final condition.

(iv) For each $\alpha \in A$ there exists a local F-constant section Ψ^α of QB over U_α which is nowhere zero on U_α.

Remark: If $\{U_\alpha | \alpha \in A\}$ is a cover satisfying conditions (i)–(iii) then the local trivializations ν^α of $B^{\sim F}$, defined by

$$\nu^\alpha(\mathrm{id} \in ML(n,\mathbb{C})) = 1 \Leftrightarrow \nu^\alpha((X_{r^1},...,X_{z^n})^\sim) = 1, \qquad (1.3)$$

are F-constant sections, hence the search for F-constant local sections of QB can be reduced to the search for F-constant local sections of L.

It is well known that two F-constant sections Ψ and χ of QB differ locally by a function on M/D (better: a function constant on the leaves of D) which is holomorphic on the leaves of $\pi_* E$. Hence if F is a positive Kähler polarization (i.e., $D = \{0\} \Leftrightarrow k = 0$) then $M/D = M$, $\pi_* E = E$, and F induces a complex structure on M turning it into a Kähler manifold. The existence of a nice cover with local F-constant sections of QB tells us that QB is a holomorphic line bundle (i.e., on $U_\alpha \cap U_\beta$ the Ψ^α and Ψ^β differ by the transition function which is holomorphic and nonvanishing since Ψ^α and Ψ^β are both F constant and nonvanishing). It follows that the Hilbert space (which we will define in Sec. III) consists of holomorphic sections of QB.

In the general case (F not Kähler) one can sometimes construct a $\mathbb{C}$-line bundle over M/D which is holomorphic on the leaves of $\pi_* E$, but success is not guaranteed (e.g., the circular polarization on $\mathbb{R}^2 \setminus \{0\}$).

II. THE BKS KERNEL

In this section we will give a heuristic definition of the BKS kernel (named after Blattner, Kostant, and Sternberg); for a more thorough definition using the metaplectic bundle we refer the reader to Ref. 2.

Suppose F and F' are two polarizations for which there exist two real foliations (of constant dimension) $D^\wedge$ and $E^\wedge$ satisfying

$$\begin{aligned} &\text{(i)} \quad F^\dagger \cap F' = D^{\wedge \mathbb{C}}, \\ &\text{(ii)} \quad F^\dagger + F' = E^{\wedge \mathbb{C}}, \qquad (2.1) \\ &\text{(iii)} \quad M/D^\wedge \text{ has a manifold structure and } \\ &\qquad \pi: M \to M/D^\wedge \text{ is a submersion.} \end{aligned}$$

Then $E^{\wedge\perp} = D^\wedge$, $D^{\wedge\perp} = E^\wedge$ and one can define a pairing (i.e., a sesquilinear form) $\langle \Psi, \Psi' \rangle$ between F-constant sections Ψ of $QB = L \otimes B^{\sim F}$ and F'-constant sections Ψ' of $QB' = L \otimes B^{\sim F'}$ by

$$\langle \Psi, \Psi' \rangle = \int_{M/D^\wedge} (\Psi, \Psi'),$$

where (Ψ,Ψ') is a density (i.e., a complex measure) on $M/D^\wedge$ defined by the following process. Suppose $\Psi^{(\prime)} = s^{(\prime)} \otimes \nu^{(\prime)}$ is a local representation of $\Psi^{(\prime)}$, $s^{(\prime)}$ a section of L, ν a section of $B^{\sim F}$ and ν' a section of $B^{\sim F'}$. Choose vectors $\xi_1,...,\xi_n, \xi'_{k+1},...,\xi'_n, t_1,...,t_k \in T_m M^{\mathbb{C}}$ such that

$$\begin{aligned} &(\xi_1,...,\xi_k) \text{ span } D^\wedge, \\ &(\xi_1,...,\xi_n) \text{ span F and } (\xi_1,...,\xi_k,\xi'_{k+1},...,\xi'_n) \text{ span } F', \\ &(\xi_1,...,\xi_k,\xi^\dagger_{k+1},...,\xi^\dagger_n,\xi'_{k+1},...,\xi'_n,t_1,...,t_k) \text{ span } T_m M^{\mathbb{C}}, \end{aligned} \qquad (2.2)$$

then $((\zeta)_0) \equiv \pi_*(\xi^\dagger_{k+1},...,\xi^\dagger_n,\xi'_{k+1},...,\xi'_n,t_1,...,t_k)$ is a (complex) frame at $\pi(m) \in M/D^\wedge$ and we define the density (Ψ,Ψ') at $\pi(m)$ on the frame $((\zeta)_0)$ by

$$\begin{aligned} &(\Psi,\Psi')(\pi(m)),((\zeta)_0) \\ &\quad = (s,s')(m) \cdot \nu((\xi_1,...,\xi_n)^\sim)^\dagger \\ &\qquad \times \cdot \nu'((\xi_1,...,\xi_k,\xi'_{k+1},...,\xi'_n)^\sim) \\ &\qquad \times \sqrt{(\det((ih)^{-1}\omega(\xi^\dagger_j,\xi'_u))|_{j,u=k+1,...,n})} \\ &\qquad \times |\mathrm{Li}(\xi_1,...,t_k)|, \end{aligned} \qquad (2.3)$$

where $\mathrm{Li} \equiv (-1)^{n(n-1)/2} \cdot \omega^n/n!$ is the Liouville volume-form on M.

An equivalent definition using an arbitrary frame $((\zeta))$ at $\pi(m)$ is given by

$$\begin{aligned} &(\Psi,\Psi')(\pi(m),((\zeta))) \\ &\quad = (s,s')(m) \cdot \nu((\xi_1,...,\xi_n)^\sim)^\dagger \\ &\qquad \times \cdot \nu'((\xi_1,...,\xi_k,\xi'_{k+1},...,\xi'_n)^\sim) \\ &\qquad \times \sqrt{(\det((ih)^{-1}\omega(\xi^\dagger_j,\xi'_u))|_{j,u=k+1,...,n})} \\ &\qquad \times |\mathrm{Li}(\xi_1,...,\xi_k,\pi_*^{-1}((\zeta)))|, \end{aligned} \qquad (2.4)$$

with $\pi_*^{-1}((\zeta))$ an arbitrary lift of the frame $((\zeta))$ to $T_m M^{\mathbb{C}}$.

Using the fact that $\Psi^{(\prime)}$ is $F^{(\prime)}$ constant one can prove that the right-hand side of (2.3) and (2.4) is independent of the choice of m in the fiber above $\pi(m)$, and these formulas define a density on $M/D^\wedge$, except for two facts: (a) we have not specified which metaframe we have to choose (there always exist two possible choices which differ by a minus sign when ν_α is applied), and (b) we do not know which branch of the (complex) square root we have to use. In general one needs the metaplectic bundle to answer these questions; however, in the cases we are interested in one does not need the metaplectic bundle (see also Refs. 4 and 5).

III. THE INNER PRODUCT AND THE HILBERT SPACE

If $F = F'$ is a positive polarization then one can choose $\xi'_j = \xi_j, j = k+1,...,n$, and so $\det((ih)^{-1}\omega(\xi^\dagger_j,\xi'_u))$ is positive [use that $v \in D^{\mathbb{C}} \Leftrightarrow \omega(v^\dagger, v) = 0$] hence we can choose everywhere the positive square root without running into trouble. Furthermore, if we choose a nice cover, then each F-constant section Ψ of QB admits a local representation $\Psi = s_\alpha \otimes \nu^\alpha$ [ν^α defined by (1.3)] and the $2n-k$ vectors $X_{r^1},...,X_{r^k},X_{z^{k+1}},...,X_{z^n},X_{z^{k+1}},...,X_{z^n}$ satisfy the first two con-

ditions of (2.2), so on $\pi(U_\alpha)$ we find for two F-constant sections Ψ and Ψ' of QB, using (2.4):

$$
\begin{aligned}
&(\Psi,\Psi')(\pi(m),\{(\xi)\}) \\
&\quad = (s_\alpha,s'_\alpha)(m)\cdot\nu^\alpha\{(X_{r^j},X_{z^u})^\sim\}^\dagger\cdot\nu^\alpha\{(X_{r^j},X_{z^u})^\sim\} \\
&\qquad \times(\det\{(ih)^{-1}\omega(X_{z^{j+}},X_{z^u})\}_{j,u=k+1,\ldots,n})^{1/2} \\
&\qquad \times|\mathrm{Li}(X_{r^1},\ldots,X_{r^k},\pi_*^{-1}\{(\xi)\})| \\
&\quad = (s_\alpha,s'_\alpha)(m)\cdot(\det\{(ih)^{-1}\omega(X_{z^{j+}},X_{z^u})\}_{j,u=k+1,\ldots,n})^{1/2} \\
&\qquad \times|\mathrm{Li}(X_{r^1},\ldots,X_{r^k},\pi_*^{-1}\{(\xi)\})|. \qquad (3.1)
\end{aligned}
$$

This formula defines a density for each pair of F-constant sections Ψ,Ψ' of QB which is positive if $\Psi=\Psi'$, depends linearly on Ψ' and antilinearly on Ψ, hence we can define a Hilbert space H as the completion of the pre-Hilbert space (PH) defined by

$$
\mathrm{PH}=\left\{\Psi:\ M\to\mathrm{QB}\,|\,\Psi\text{is }F\text{ constant and }\int_{M/D}(\Psi,\Psi)<\infty\right\},
$$

$$
\text{with inner product }\langle\Psi,\Psi'\rangle=\int_{M/D}(\Psi,\Psi'). \qquad (3.2)
$$

Of course, if we wish to obtain results it remains to show in each case that $H\neq\{0\}$, a fact which sometimes leads us to consider distribution valued sections of QB, instead of C^∞ sections (e.g., on $\mathbb{R}^2\setminus\{0\}$ with the circular polarization) (see Refs. 2, 6, and 7).

IV. QUANTIZABLE OBSERVABLES

If f is an observable, i.e., $f: M\to\mathbb{R}$, then a quantization procedure should associate to f a self-adjoint operator $\mathbf{f}$ on the Hilbert space H. In geometric quantization the general procedure is given below, although it does not guarantee that the result is a self-adjoint operator. However, it turns out that in almost all interesting examples the result is a self-adjoint operator. To construct $\mathbf{f}$ one proceeds as follows: let ρ_t be the flow on M associated to the Hamiltonian vector field X_f, define the polarization $F(t)$ by $F(t)=\rho_{-t*}F$ and the bundle $\mathrm{QB}(t)$ by $\mathrm{QB}(t)=L\otimes B^{\sim F(t)}$. Then there exists an associated map on sections of L (called ρ_t^* by abuse of notation) and a map (called ρ_{t*}, also abuse of notation) ρ_{t*}: $R^{\sim F}\to R^{\sim F(t)}$ which are both defined in a "canonical" way [e.g., $\rho_{t*}: R^{\sim F}\to R^{\sim F(t)}$ is the lift to the metalinear bundles of a map which is the restriction to $R^{\sim F}$ and $R^{\sim F(t)}$, respectively, of the flow on the bundle of all n frames of M (of which $R^{\sim F}$ and $R^{\sim F(t)}$ are subbundles) associated to the flow ρ_t on M]. If s is an arbitrary section of L (with local representation s_α with respect to a nice cover) and ν an arbitrary section of $B^{\sim F}$ then

$$
\begin{aligned}
&(\rho_t^*s)_\alpha(m) \\
&\quad = s_\alpha(\rho_t m)\cdot\exp\left((i\hbar)^{-1}\int_0^t(\vartheta_\alpha(X_f)-f)(\rho_s m)ds\right), \qquad (4.1)
\end{aligned}
$$

$$
(\rho_t^*\nu)(\{\xi\}^\sim)=\nu(\{\rho_{t*}(\xi)\}^\sim),
$$

$$
\{\rho_{t*}(\xi)\}^\sim\text{ defined by continuity in }t. \qquad (4.2)
$$

The map ρ_t^*: $\mathrm{QB}\to\mathrm{QB}(t)$ defined by $\rho_t^*(s\otimes\nu)=(\rho_t^*s)\otimes(\rho_t^*\nu)$ now obviously has the property that if Ψ is a F-constant section of QB then $\rho_t^*\psi$ is a $F(t)$-constant section of $\mathrm{QB}(t)$. Finally we assume that for all $t\in(0,\epsilon)$ ($\epsilon>0$) F and $F(t)$ satisfy the conditions (2.1) for a pair of polarizations. If we denote the pairing between QB and $\mathrm{QB}(t)$ by $\langle\ ,\ \rangle_{F,F(t)}$ and the inner product in H by $\langle\ ,\ \rangle_H$ then the operator $\mathbf{f}$ is defined by the equation

$$
\langle\chi,\mathbf{f}\Psi\rangle_H=\lim_{t\downarrow0}-i\hbar\frac{d}{dt}\langle\chi,\rho_t^*\Psi\rangle_{F,F(t)},\quad \chi,\Psi\in H. \qquad (4.3)
$$

Remark: It is not at all evident that this formula defines an operator $\mathbf{f}$, and indeed there exist examples in which this "definition" does not yield a result, for instance in cases in which $\lim\langle\chi,\rho_t^*\Psi\rangle_{F,F(t)}$ is not equal to $\langle\chi,\Psi\rangle_H$.

In general one has to know the specific form of f to compute (4.3); however there exist conditions on f for which one can simplify (4.3) and in which one can obtain an explicit expression for the operator $\mathbf{f}$. Two of such conditions are well known: (a) if f is constant along F, i.e., $[X_f,F]=\{0\}$, then $\mathbf{f}$ is multiplication by f; (b) if X_f leaves F invariant, i.e., $[X_f,F]\subset F$, then $\mathbf{f}$ is represented by a first-order differential operator.

In this paper we will show that there is an even weaker condition for which we can compute $\mathbf{f}$ explicitly: the condition $[X_f,F^\dagger+F]\subset F^\dagger+F$. To compute (4.3) with this condition on f we proceed as follows: first observe that, since f is real and F a polarization, we also have $[X_f,D]\subset D$, which implies that F and $F(t)$ satisfy conditions (2.1) with $D^\wedge=D$ and $E^\wedge=E$; it follows that we may replace in (4.3) the "$\lim t\downarrow0\ d/dt$" by "$d/dt\,|_{t=0}$." Now we choose a nice cover and two arbitrary elements Ψ and Ψ' of H, and we perform the calculations on a local chart U_α where we have

$$
\Psi=s\otimes\nu^\alpha,\quad \Psi'=s'\otimes\nu^\alpha,\quad [\text{see (1.3)}],
$$

$$
\begin{aligned}
&(\Psi',\rho_t^*\Psi)(\pi(m),\{(\xi)\}) \\
&\quad = (s',\rho_t^*s)(m)\cdot\nu^\alpha\{(X_{r^j},X_{z^u})^\sim\}^\dagger \\
&\qquad \times(\rho_t^*\nu^\alpha)\{(X_{r^j},\rho_{-t*}(X_{z^u}))^\sim\} \\
&\qquad \times(\det\{(ih)^{-1}\omega(X_{z^{j+}},\rho_{-t*}X_{z^u})\}_{j,u=k+1,\ldots,n})^{1/2} \\
&\qquad \times|\mathrm{Li}(X_{r^1},\ldots,X_{r^k},\pi_*^{-1}\{(\xi)\})|,
\end{aligned}
$$

where the lift $\{X_{r^j},\rho_{-t*}(X_{z^u})\}^\sim$ to $R^{\sim F(t)}$ of the frame $\{X_{r^j},\rho_{-t*}(X_{z^u})\}$ is chosen in such a way that it depends continously on t and reduces to the well defined metaframe $(X_{r^j},X_{z^u})^\sim$ at $t=0$. Using 4.2 and the definition of ν_α (1.3) we can simplify this expression to

$$
\begin{aligned}
&(\Psi',\rho_t^*\Psi)(\pi(m),\{(\xi)\}) \\
&\quad = (s',\rho_t^*s)(m)\cdot\nu^\alpha\{(\rho_{t*}(X_{r^j}),X_{z^u})^\sim\} \\
&\qquad \times(\det\{(ih)^{-1}\omega(X_{z^{j+}},X_{z^u\cdot\rho_t})\}_{j,u=k+1,\ldots,n})^{1/2} \\
&\qquad \times|\mathrm{Li}(X_{r^1},\ldots,X_{r^k},\pi_*^{-1}\{(\xi)\})|.
\end{aligned}
$$

We now define functions $a_{j,u}$ and $b_{j,s}$ by

$$
\begin{aligned}
&[X_f,X_{r^j}]=\sum_{u=1}^{k}a_{j,u}X_{r^u},\quad 1\leqslant j\leqslant k \\
&[X_f,X_{z^j}]=\sum_{u=1}^{k}a_{j,u}X_{r^u}+\sum_{u=k+1}^{n}a_{j,u}X_{z^u} \\
&\qquad\qquad +\sum_{s=k+1}^{n}b_{j,s}X_{z^s},\quad k<j\leqslant n. \qquad (4.4)
\end{aligned}
$$

If we use the transformation property of $(-\frac{1}{2})$-F-forms (1.1), we get

$$\frac{d}{dt}\Big|_{t=0} \nu^{\alpha}((\rho_{t*}(X_{r^j}),X_{z^u})^{\sim}) = \frac{1}{2}\sum_{j=1}^{k} a_{jj},$$

$$\frac{d}{dt}\Big|_{t=0} (\det|(ih)^{-1}\omega(X_{z^j\uparrow},X_{z^u\cdot\rho_t})|_{j,u=k+1,\ldots,n})^{1/2}$$
$$= \left(\frac{1}{2}\sum_{j=k+1}^{n} a_{jj}\right)$$
$$\times(\det|(ih)^{-1}\omega(X_{z^j\uparrow},X_{z^u})|_{j,u=k+1,\ldots,n})^{1/2}.$$

Combining these results with (4.1) and (1.2) we find

$$\frac{d}{dt}\Big|_{t=0} (\Psi',\rho_t^*\Psi)(\pi(m),(\zeta)))$$
$$= \left(s',\nabla_{Xf}s + \left(\frac{i}{\hbar}f + \frac{1}{2}\sum_{j=1}^{n} a_{jj}\right)s\right)(m)$$
$$\times(\det|(ih)^{-1}\omega(X_{z^j\uparrow},X_{z^u})|_{j,u=k+1,\ldots,n})^{1/2}$$
$$\times|\mathrm{Li}(X_{r^1},\ldots,X_{r^k},\pi_*^{-1}((\zeta)))|. \tag{4.5}$$

Comparing this formula with the inner product (3.1) we might say

$$-i\hbar\frac{d}{dt}\Big|_{t=0} (\Psi',\rho_t^*\Psi)(\pi(m),(\zeta)))$$
$$= (\Psi',L_f\Psi)(\pi(m),(\zeta)))$$

or

$$-i\hbar\frac{d}{dt}\Big|_{t=0} (\Psi',\rho_t^*\Psi) = (\Psi',L_f\Psi), \tag{4.6}$$

where the section $L_f\Psi$ of QB is defined by the local expression

$$L_f(s\otimes\nu^{\alpha}) = \left(-i\hbar\nabla_{Xf}s + \left(f - \frac{1}{2}i\hbar\sum_{j=1}^{n} a_{jj}\right)s\right)\otimes\nu^{\alpha}. \tag{4.7}$$

By construction the right-hand side of (4.5) is independent of m as long as $\pi(m)$ remains fixed, so it is a well defined density at $\pi(m)\in M/D$. However the section $L_f\Psi$ of QB need not be an F-constant section of QB! So if we assume that it is allowed to change differentiation and integration in (4.3) (and we do!) then we find

$$\langle\Psi',\mathbf{f}\Psi\rangle = \int_{M/D}(\Psi',L_f\Psi) \approx \langle\Psi',L_f\Psi\rangle, \tag{4.8}$$

but nevertheless we cannot conclude that $\mathbf{f}\Psi = L_f\Psi$, because we do not know whether $L_f\Psi$ is F constant or not. What we are looking for is a way to construct the "F-constant part" out of $L_f\Psi$: $L_f\Psi$ determines a linear operator on H by (4.8) so (if it is continuous) it determines an element $\mathbf{f}\Psi$ of H. In the next section we will show how to construct this element out of $L_f\Psi$ by means of a kernel representation.

V. A GENERALIZED BERGMAN KERNEL

In this section we will assume that F is a positive Kähler polarization and that there exists a nice cover; in Sec. I we have seen that, under these assumptions, QB is a holomorphic line bundle over the complex manifold M, trivialized by the local sections Ψ^{α} associated to the nice cover. Now if χ is any section of QB, we denote by $\chi_{\alpha}: U_{\alpha}\to\mathbb{C}$ the local representation of χ with respect to the trivialization Ψ^{α}, i.e.,

$$\chi|_{U_{\alpha}}(m) = \chi_{\alpha}(m)\cdot\Psi^{\alpha}(m) \tag{5.1}$$

(N.B. the local representation carries a subscript α whereas the trivializing section Ψ^{α} carries a superscript); moreover χ is a F-constant section if and only if χ_{α} are holomorphic functions.

To each pair of F-constant section χ and χ' we associated in Sec. III a density on $M/D\equiv M$ (because F is Kähler) and in particular $(\Psi^{\alpha},\Psi^{\alpha})$ is a (local) density on M. Let ρ_{α} be a partition of unity subordinated to our nice cover, then we can define for any two sections χ and χ' of QB (holomorphic or not) a density (χ,χ') by

$$(\chi,\chi')(m) = \sum_{\alpha}\rho_{\alpha}(m)\cdot\chi_{\alpha}(m)\dagger\cdot\chi'_{\alpha}(m)\cdot(\Psi^{\alpha},\Psi^{\alpha})(m), \tag{5.2}$$

which coincides with definition (3.1) if χ and χ' are F constant because on U_{α}: $(\chi,\chi') = (\chi_{\alpha}\cdot\Psi^{\alpha},\chi'_{\alpha}\cdot\Psi^{\alpha}) = \chi_{\alpha}^{\dagger}\cdot\chi'_{\alpha}\cdot(\Psi^{\alpha},\Psi^{\alpha})$ and because $\Sigma_{\alpha}\rho_{\alpha} = 1$. It follows that we can construct a Hilbert space $L^2(M,F)$ defined as the completion of the pre-Hilbert space

$$\left\{\chi\colon M\to QB \;\middle|\; \int_M(\chi,\chi) < \infty\right\}, \tag{5.3}$$

with inner product

$$\langle\chi,\chi'\rangle = \int_M(\chi,\chi')$$

and associated norm

$$\|\chi\| = \sqrt{\langle\chi,\chi\rangle}.$$

If we introduce the measure μ_{α} on U_{α} defined by

$$d\mu_{\alpha} = \rho_{\alpha}(\Psi_{\alpha},\Psi_{\alpha}), \tag{5.4}$$

then this pre-Hilbert space is defined equivalently by

$$\left\{\chi\colon M\to QB \;\middle|\; \sum_{\alpha}\int_{U\alpha}|\chi_{\alpha}|^2\,d\mu_{\alpha} < \infty\right\}, \tag{5.5}$$

with inner product

$$\langle\chi,\chi'\rangle = \sum_{\alpha}\int_{U\alpha}\chi_{\alpha}^{\dagger}\cdot\chi'_{\alpha}\,d\mu_{\alpha}.$$

By construction the pre-Hilbert space PH defined in (3.2) is a subspace of $L^2(M,F)$, hence H as defined in Sec. III (the completion of PH) is the closure of PH in $L^2(M,F)$. What we will show first is that PH is a closed subspace (see also Ref. 7, §5.7), hence $H\simeq$ PH which we will also denote by $L^2(M,F)_{\mathrm{hol}}$, suggested by the fact that F-constant sections are just the holomorphic sections. The main ingredient to prove this claim (and others) is the following lemma.

Lemma 5.1: If χ is an F-constant section in PH then for any U_{α}, for any compact subset K of U_{α} there exists a positive constant $c = c(K)$ such that, for all $m\in K$,

$$|\chi_{\alpha}(m)| \leqslant c(K)\cdot\|\chi\|.$$

Proof: we denote by $B(m,\epsilon)$ the open ball of radius ϵ around m in a local chart contained in $\mathbb{C}^n$. Because K is compact there exists a $\delta > 0$: $\forall m\in K$: $B(m,2\delta)\subset U_{\alpha}$; since χ_{α} is holomorphic on U_{α} ("$U_{\alpha}\subset\mathbb{C}^n$") if follows by the mean val-

ue theorem and the Cauchy–Schwartz inequality that

$$|\chi_\alpha(m)|^2 \leqslant \text{const}(\delta)\cdot\int_{B(m,\delta)}|\chi_\alpha(m')|^2\cdot d\lambda^{(2n)},$$

where $\lambda^{(2n)}$ denotes the Lebesgue measure on $\mathbb{C}^n$. Now $(\Psi^\alpha,\Psi^\alpha)$ is a continuous density on U_α which is nowhere zero, hence there exists a constant c' depending on the compact subset K' defined by

$$K' = \text{closure}\Big(\bigcup_{m\in K} B(m,\delta)\Big)\subset U_\alpha,$$

such that

$$d\lambda^{(2n)} \leqslant c'\cdot(\Psi^\alpha,\Psi^\alpha) \quad \text{on } K',$$

hence

$$\int_{B(m,\delta)}|\chi_\alpha(m')|^2 d\lambda^{(2n)}$$
$$\leqslant c'\cdot\int_{B(m,\delta)}|\chi_\alpha(m')|^2\cdot(\Psi^\alpha,\Psi^\alpha)$$
$$= c'\cdot\int_{B(m,\delta)}(\chi,\chi)\leqslant c'\cdot\int_M(\chi,\chi) = c'\cdot\|\chi\|^2.$$

From these inequalities the lemma follows because const(δ) and c' are postive constants depending only on the compact subset K. Q.E.D

Using this lemma, the proof of the theorem stated below is a direct copy of the case of complex functions on a domain in $\mathbb{C}^n$: a Cauchy sequence in $L^2(M,F)_{\text{hol}}$ implies pointwise convergence, uniform on compacta.

Theorem 5.2: (1) PH is a closed subspace of $L^2(M,F)$. (2) For any $m\in U_\alpha$ the map $\text{PH}\equiv H\equiv L^2(M,F)_{\text{hol}} \to \mathbb{C}, \chi\to\chi_\alpha(m)$ is a continuous linear functional.

Corollary 5.3: For each $m\in U_\alpha$ there exists a unique F-constant section $\kappa_{(\alpha,m)}$ of QB in $L^2(M,F)_{\text{hol}}$, such that for each $\chi\in L^2(M,F)_{\text{hol}}$:

$$\chi_\alpha(m) = \langle\kappa_{(\alpha,m)},\chi\rangle.$$

Definition: The generalized Bergman kernel $K_{(\alpha,\beta)}(m,m')$ is defined by

$$K_{(\alpha,\beta)}(m,m') = \kappa_{(\alpha,m)\,\beta}(m')\dagger \quad [\text{see (5.1)}]. \qquad (5.6)$$

Corollary 5.4: For $\chi\in L^2(M,F)_{\text{hol}}$:

$$\chi_\alpha(m) = \sum_\beta\int_{U\beta}K_{(\alpha,\beta)}(m,m')\cdot\chi_\beta(m')\cdot d\mu_\beta(m').$$

Proposition 5.5: If (φ_j) is a complete orthonormal set in $L^2(M,F)_{\text{hol}}$ then

$$\sum_j\varphi_{j\alpha}(m)\varphi_{j\beta}(m')\dagger \text{ converges to } K_{(\alpha,\beta)}(m,m').$$

Proof: The first step is to prove that for fixed $m\in U_\alpha$ the series $(\varphi_{j\alpha}(m))$ is a square-summable series, i.e., $(\varphi_{j\alpha}(m))\in l^2$. Therefore let $(a_j)\in l^2$ with $\|(a_j)\|^2 = \Sigma_j|a_j|^2$, then by the Cauchy–Schwartz theorem

$$\sum_{j=1}^N|\varphi_{j\alpha}(m)|^2 = \underset{\|(a_j)\|<1}{\text{supremum}}\left|\sum_{j=1}^N a_j\varphi_{j\alpha}(m)\right|.$$

If we associate to $(a_j)\in l^2$ the element

$$\Psi = \sum_{j=1}^N a_j\varphi_j\in L^2(M,F)_{\text{hol}}$$

then $\|\Psi\|^2\leqslant\|(a_j)\|^2$, and so we have

$$\sum_{j=1}^N|\varphi_{j\alpha}(m)|^2\leqslant\underset{\|\Psi\|<1}{\text{supremum}}|\Psi_\alpha(m)|\leqslant\text{const},$$

where the last inequality follows from Lemma 5.1. So the series of partial sums is bounded from above hence $(\varphi_{j\alpha}(m))\in l^2$. Applying the Riesz–Fisher theorem it follows that $\chi_{(\alpha,m)}$, defined by

$$\chi_{(\alpha,m)} = \sum_j\varphi_{j\alpha}(m)^\dagger\varphi_j,$$

is in $L^2(M,F)_{\text{hol}}$ and we claim that $\chi_{(\alpha,m)} = \kappa_{(\alpha,m)}$. Therefore choose any $\chi\in L^2(M,F)_{\text{hol}}$, then

$$\chi = \sum_j\langle\varphi_j,\chi\rangle\varphi_j$$

[because (φ_j) is complete orthonormal]

and so [because point evaluation is continuous (Theorem 5.2)]

$$\chi_\alpha(m) = \sum_j\langle\varphi_j,\chi\rangle\varphi_{j\alpha}(m) = \sum_j\langle\varphi_{j\alpha}(m)^\dagger\varphi_j,\chi\rangle$$
$$= \left\langle\sum_j\varphi_{j\alpha}(m)^\dagger\varphi_j,\chi\right\rangle = \langle\chi_{(\alpha,m)},\chi\rangle.$$

By uniqueness of $\kappa_{(\alpha,m)}$ the equality follows, whence we have

$$K_{(\alpha,\beta)}(m,m') = \kappa_{(\alpha,m)\,\beta}(m')^\dagger$$
$$= \chi_{(\alpha,m)\,\beta}(m')^\dagger = \sum_j\varphi_{j\alpha}(m)\varphi_{j\beta}(m')^\dagger.$$

Q.E.D

Corollary 5.6: $K_{(\alpha,\beta)}(m,m')$ is holomorphic in m and $K_{(\alpha,\beta)}(m,m')^\dagger = K_{(\beta,\alpha)}(m',m)$.

Having defined the generalized Bergman kernel, we can proceed with the main story. An element $\chi\in L^2(M,F)$ defines a continuous linear functional on $L^2(M,F)_{\text{hol}}$ by means of the inner product: $\Psi\in L^2(M,F)_{\text{hol}}\to(\chi,\Psi)\in\mathbb{C}$. Since $L^2(M,F)_{\text{hol}}$ is a Hilbert space this linear functional can be represented by an element $\chi_{\text{hol}}\in L^2(M,F)_{\text{hol}}$ defined by $\langle\chi,\Psi\rangle = \langle\chi_{\text{hol}},\Psi\rangle$ where the inner product on the left is in $L^2(M,F)$ and on the right in $L^2(M,F)_{\text{hol}}$.

One can easily show that the map $\chi\to\chi_{\text{hol}}$, $L^2(M,F)\to L^2(M,F)_{\text{hol}}$ is the orthogonal projection onto the closed subspace $L^2(M,F)_{\text{hol}}$ and we will show in Proposition (5.7) that this projection can be represented by an integral formula using the kernel $K_{(\alpha,\beta)}$. The reason why we call this kernel a generalized Bergman kernel is that if one replaces $L^2(M,F)$ by the Hilbert space of square integrable functions on a domain G in $\mathbb{C}^n$ and $L^2(M,F)_{\text{hol}}$ by the subspace of holomorphic functions then the same reasoning as above applies; the associated projection admits a kernel representation which is the (usual) Bergman kernel representation.

Proposition 5.7:

$$\chi_{\text{hol}\,\alpha}(m) = \sum_\beta\int_{U\beta}K_{(\alpha,\beta)}(m,m')\cdot\chi_\beta(m')\cdot d\mu_\beta(m').$$

Proof: $\chi-\chi_{\text{hol}}$ is orthogonal to $L^2(M,F)_{\text{hol}}$, hence $\langle\kappa_{(\alpha,m)},\chi\rangle = \langle\kappa_{(\alpha,m)},\chi_{\text{hol}}\rangle = \chi_{\text{hol}\,\alpha}(m)$, where the last equality follows from Corollary 5.3. Q.E.D.

With this proposition we can solve the question posed at the end of Sec. IV concerning the construction of operators in geometric quantization.

Theorem 5.8: Let (M,ω) be a symplectic manifold, F a positive Kähler polarization and suppose there exists a nice cover (see Sec. I). If f is any observable, i.e., $f: M\to\mathbb{R}$, then the associated operator f on $H = L^2(M,F)_{\rm hol}$ is defined by

$$\mathrm{Dom}(\mathrm{f}) = \{\chi\in L^2(M,F)_{\rm hol} \,|\, L_f\chi\in L^2(M,F)\},$$
$$\chi\in\mathrm{Dom}(\mathrm{f})\Rightarrow \mathrm{f}\chi = (L_f\chi)_{\rm hol},$$

where $L_f\chi$ is defined by (4.7) and the holomorphic part $(L_f\chi)_{\rm hol}$ by Proposition 5.7.

VI. APPLICATION I

In this example we consider the symplectic manifold $M=\mathbb{R}^2$, $\omega = dp\wedge dq$ and the polarization $F=\mathbb{C}\cdot(\partial/\partial p + i\partial/\partial q) = \mathbb{C}X_{p+iq}$. In this case all bundles, L, R^F, $R^{\sim F}$, $B^{\sim F}$, and hence QB are trivial so we can identify sections of these bundles with functions on M. For L the connection ∇ is then given by

$$\nabla_\xi s = \xi s - (i/\hbar)\vartheta(\xi)s,$$

where we choose the symplectic potential $\vartheta = \frac{1}{2}(p\,dq - q\,dp)$; the compatible Hermitian form is given by $(s,s')(m) = s(m)^\dagger\cdot s'(m)$. The cover consisting of the one chart $\mathbb{R}^2$ is a nice cover because the section Ψ^0 defined by

$$\Psi^0(p,q) = \exp(-(p^2+q^2)/(4\hbar))$$

is F constant and nonvanishing. The complex structure on M induced by the positive Kähler polarization F is such that $z = p + iq$ is a holomorphic coordinate.

According to Sec. V we have to compute the density (Ψ^0,Ψ^0) in order to know the inner product in $L^2(M,F)$ and in $L^2(M,F)_{\rm hol}$; using formula (3.1) (with X_{p+iq} spanning F) we find

$$(\Psi^0,\Psi^0) = \sqrt{(2/h)}\exp(-z^\dagger z/(2\hbar))dp\,dq.$$

Consequently if we identify a section χ of QB with the function $g=\chi_0$ (i.e., $\chi = g\Psi^0$), the Hilbert space $L^2(M,F)$ is given by

$$L^2(M,F) = \left\{g:\ \mathbb{C}\to\mathbb{C}\,\middle|\,\left(\frac{2}{h}\right)^{1/2}\int_{\mathbb{C}}|g(z)|^2 \times\exp\left(\frac{-z^\dagger z}{(2\hbar)}\right)dp\,dq<\infty\right\},$$

$$\langle g,g'\rangle = \left(\frac{2}{h}\right)^{1/2}\int_{\mathbb{C}} g(z)^\dagger g'(z)\exp\left(\frac{-z^\dagger z}{(2\hbar)}\right)dp\,dq,$$

$L^2(M,F)_{\rm hol}$ is the subspace of holomorphic functions g; since $L^2(M,F)_{\rm hol} = H$ describes (according to geometric quantization) the quantum mechanical system we find here the well known Bargmann representation, mostly used for the harmonic oscillator.

The elements $g_n(z) = (2h)^{-1/4}(n!)^{-1/2}(2\hbar)^{-n/2}z^n$ form an orthonormal complete system hence [Proposition (5.5)] the Bergman kernel for this Hilbert space is given by

$$K(w,z) = \sum_n g_n(w)(g_n(z))^\dagger = (2h)^{-1/2}\exp\left(\frac{z^\dagger w}{(2\hbar)}\right),$$

so Corollary 5.4 gives the well known reproducing formula for holomorphic functions on $\mathbb{C}$:

$$g(w) = \left(\frac{2}{h}\right)^{1/2}\int k(w,z)g(z)\exp\left(\frac{-z^\dagger z}{(2\hbar)}\right)dp\,pq$$
$$= h^{-1}\int g(z)\exp\left(\frac{z^\dagger(w-z)}{(2\hbar)}\right)dp\,pq. \qquad (6.1)$$

Now let f be any observable and let $g\in H$ then

$$X_f = 2i\left(\frac{\partial f}{\partial z^\dagger}\frac{\partial}{\partial z} - \frac{\partial f}{\partial z}\frac{\partial}{\partial z^\dagger}\right),$$
$$[X_f,X_z] = 2i\left(\frac{\partial^2 f}{(\partial z^\dagger)^2}X_{z^\dagger} + \frac{\partial^2 f}{\partial z\,\partial z^\dagger}X_z\right),$$

so when we calculate formula (4.7) we get

$$L_f(g\Psi^0) = \left(2\hbar\frac{\partial f}{\partial z^\dagger}\frac{dg}{dz} + \left(f - z^\dagger\frac{\partial f}{\partial z^\dagger} + \hbar\frac{\partial^2 f}{\partial z\,\partial z^\dagger}\right)g\right)\Psi^0.$$

Applying Theorem 5.8 we find the following prescription to compute the operator f (using partial integration, omitting Ψ^0).

Prescription I: In the Bergmann representation the operator f associated to an observable f is given by

$$(\mathrm{fg})(w) = h^{-1}\int\exp\left(\frac{z^\dagger(w-z)}{(2\hbar)}\right)\left(f-\hbar\frac{\partial^2 f}{\partial z\,\partial z^\dagger}\right)g(z)dp\,dq. \qquad (6.2)$$

We now notice that for "any" holomorphic function $k(z)$ the following formula holds:

$$h^{-1}\int\exp\left(\frac{z^\dagger(w-z)}{(2\hbar)}\right)z^\dagger k(z)dp\,dq$$
$$= h^{-1}\int\exp\left(\frac{z^\dagger(w-z)}{(2\hbar)}\right)2\hbar\frac{dk}{dz}\,dp\,dq = 2\hbar\frac{dk}{dz(w)}.$$

Combining this with the previous formula we get a prescription how to obtain the quantum mechanical operators associated to polynomial observables (see also Ref. 7, §6.3.4).

Prescription II: if f is a polynomial observable, then the operator f in the Bargmann representation is given by the following process.

(a) Compute $(f-\hbar\partial^2 f/(\partial z\partial z^\dagger))$ as polynomial in z and $z^\dagger$.

(b) Write in this polynomial $z^\dagger$ *to the left of* z.

(c) Replace each $z^\dagger$ by $2\hbar\,d/dz$.

Examples:

(1) $f = \frac{1}{2}(p^2+q^2)\Rightarrow f = \frac{1}{2}z\cdot z^\dagger$

$$\Rightarrow f - \hbar\frac{\partial^2 f}{\partial z\,\partial z^\dagger} = \frac{1}{2}z^\dagger\cdot z - \frac{1}{2}\hbar,$$
$$\mathrm{f} = \frac{1}{2}\cdot\left(2\hbar\frac{d}{dz}\right)z - \frac{1}{2}\hbar = \hbar\left(z\cdot\frac{d}{dz}+\frac{1}{2}\right);$$

(2) $f = \frac{1}{2}p^2\Rightarrow f - \hbar\frac{\partial^2 f}{\partial z\,\partial z^\dagger} = \frac{z^{\dagger 2}+2z^\dagger z + z^2 - 2\hbar}{8},$

$$\mathrm{f} = \frac{1}{2}\hbar^2\frac{d^2}{dz^2} + \frac{1}{2}\hbar z\frac{d}{dz} + \frac{z^2}{8} + \frac{1}{4}\hbar;$$

(3) $f=\frac{1}{2}q^2$

$$\Rightarrow f-\hbar\frac{\partial^2 f}{\partial z\,\partial z^\dagger}=\frac{-(z^{\dagger 2}-2z^\dagger z+z^2+2\hbar)}{8},$$

$$\mathrm{f}=-\frac{1}{2}\hbar^2\frac{d^2}{dz^2}+\frac{1}{2}\hbar z\frac{d}{dz}-\frac{z^2}{8}+\frac{1}{4}\hbar.$$

Remark 1: The connection between these operators and their counterpart in the usual Schrödinger representation (H = square-integrable functions of q) will be discussed in the third application.

Remark 2: The eigenfunctions of the Hamiltonian $H=\frac{1}{2}(p^2+q^2)$ are the functions g_n defined by $g_n(z)=(2h)^{-1/4}(n!)^{-1/2}(2\hbar)^{-n/2}z^n$ with eigenvalues $(n+\frac{1}{2})\hbar$ $(n=0,1,2,\ldots)$. It follows that the operator z (multiplication by z) is a creation operator, and the operator $z^\dagger=2\hbar\,d/dz$ an annihilation operator. With this interpretation our prescription to compute f can be stated as follows: compute $f-\hbar\partial^2 f/(\partial z\,\partial z^\dagger)$ in terms of creation and annihilation observables z and $z^\dagger$ and put all annihilation operators to the left of the creation operators. Stated in these words, this prescription resembles the normal ordering used in quantum field theory, where—contrary to this case—the annihilation operators are put to the right of the creation operators.

Remark 3: In the above description of our system we have used the coordinates p (momentum) and q (position) and we have introduced a complex coordinate $z=p+iq$. However, this has physically no meaning because p and q do not have the same dimension. If we wish to obtain a physically correct description of our system we can introduce constants α and β with dimensions momentum (resp. position), and then define a new complex coordinate $z'=p/\alpha+iq/\beta$, which has no dimension. The changes in our prescription due to this change of the complex coordinate are slight:

$$H=\{g:\ \mathbb{C}\to\mathbb{C}|g \text{ holomorphic}\},$$

$$\langle g,h\rangle=\sqrt{(2\alpha\beta/h)}\int_{\mathbb{C}}g(z')^\dagger h(z')\times\exp\left(\frac{-z'^\dagger z'\alpha\beta}{2\hbar}\right)\frac{dp}{\alpha}\frac{dq}{\beta},$$

and in formula (6.2) and in the prescription $\hbar$ has to be replaced by the dimensionless constant $\hbar/(\alpha\beta)$.

VII. APPLICATION II

In this case we consider the symplectic manifold $M=S^2$ together with the symplectic form $\omega=-\lambda\epsilon$, where $\lambda\in\mathbb{R}\setminus\{0\}$ and ϵ the standard volume- (= surface-) form on S^2. For the physical interpretation of this symplectic manifold as representing the phase space of the classical spin, we refer the reader to Ref. 8. In polar coordinates (θ,ϕ) on $S^2\supset\mathbb{R}^3$, ϵ is given by $\epsilon=\sin\theta\,d\theta\wedge d\phi$; however, we prefer to use complex-holomorphic charts on $S^2=\mathbf{P}^1(\mathbf{C})$ which can be obtained by projection from the north/south pole: $U_0=\mathbb{C}=U_1$ with transition function $U_0\ni z\to 1/z=w\in U_1$, which corresponds in homogeneous coordinates $(z_0{:}z_1)$ on $\mathbb{P}^1(\mathbb{C})$ to $z=z_1/z_0\leftrightarrow(z_0{:}z_1)\leftrightarrow w=z_0/z_1$; the relation with polar coordinates is $z=\cot g(\theta/2)e^{i\phi}$. In these coordinates the volume element is given by

$$\epsilon=[-2i/(z^\dagger z+1)^2]dz\wedge dz^\dagger$$
$$=[-2i/(w^\dagger w+1)^2]dw\wedge dw^\dagger.$$

We now introduce the local symplectic potentials ϑ_j on U_j defined by

$$\vartheta_0=(i\lambda/(z^\dagger z+1))(z\,dz^\dagger-z^\dagger\,dz),$$
$$\vartheta_1=(i\lambda/(w^\dagger w+1))(w\,dw^\dagger-w^\dagger\,dw),$$

so $\vartheta_0-\vartheta_1=d(i\lambda\log(w/w^\dagger))$, from which one deduces that the gauge transformation g_{01} of the prequantum bundle L is given by

$$g_{01}(z)=\exp(i(i\lambda/\hbar)\log(w/w^\dagger))=(z/z^\dagger)^{\lambda/\hbar}.$$

Since g_{01} should be a well-defined function on $U_0\cap U_1$ this formula shows (it is not a proof, but it can be made into one) that L exists if and only if $2\lambda/\hbar\in\mathbb{Z}$, i.e.,

$$\lambda=n\cdot\hbar/2,\quad \text{for some } n\in\mathbb{Z}.$$

On (S^2,ω) we use the positive Kähler polarization $F=\mathbb{C}X_z=\mathbb{C}X_w$ (X_f as always the Hamiltonian vector field associated to the function f) so with the trivialization defined by these Hamiltonian vector fields, the gauge transformation g'_{01} of the F-frame bundle R^F is given by

$$g'_{01}(z)=-z^{-2}\quad(X_z=-z^2X_w).$$

Since $-z^{-2}$ admits a global square root the metaframe bundle $R^{\sim F}$ exists and hence the quantum bundle QB. Using formula (1.2) one can show that the local sections s^j on U_j of L defined by

$$s^0(z)=(z^\dagger z+1)^{-n/2},\quad s^1(w)=(w^\dagger w+1)^{-n/2}$$

are F constant, hence the cover $\{U_0,U_1\}$ is a nice cover. Combining the transition functions of the bundles L and $B^{\sim F}$ [the latter has transition function $\sqrt{(-z^{-2})}=i/z$] with the fact that $(w^\dagger w+1)^{-n/2}=(z^\dagger z+1)^{-n/2}\cdot(z^\dagger z)^{n/2}$, we get the result that, with respect to the trivialization of QB by the F-constant sections $\Psi^j=s^j\otimes\nu^j$ [see formula (1.3)], the transition function h_{01} of QB is given by

$$h_{01}(z)=i/z\cdot(z/z^\dagger)^{n/2}\cdot(z^\dagger z)^{n/2}=iz^{n-1}.$$

In other words, a global section χ of QB determines two holomorphic functions χ_i [see (5.1)], who are related by the equation:

$$\chi_0=h_{01}\cdot\chi_1\Leftrightarrow\chi_0(z)=iz^{n-1}\chi_1(1/z).$$

This shows that χ_i can be at most a polynomial in z of degree $n-1$, implying that if we wish to obtain something nontrivial, then n (and hence λ) should be positive, and that the resulting Hilbert space $H=L^2(M,F)_{\mathrm{hol}}$ has dimension (at most) n. At this point we mention that, had we used the polarization $F^\dagger$ instead of F, then the complex coordinate $z^\dagger$ had been the holomorphic coordinate and we had found that χ_i could be at most a polynomial of degree $-n-1$, implying that n (and hence λ) should be negative and that H has dimension (at most) $-n$. To show that this bound on the holomorphic functions χ_i is in complete agreement with the inner product, we compute the densities (Ψ^i,Ψ^i), using for-

mula (3.1):

$$(\Psi^0,\Psi^0)(z) = (z^\dagger z+1)^{-n}\cdot(\omega(X_{z^\dagger},X_z)/(ih))^{1/2}\cdot|\omega|$$
$$= (2\lambda/h)^{1/2}\cdot(z^\dagger z+1)^{-n-1}\cdot|dz\wedge dz^\dagger|$$
$$= 2(n/2\pi)^{1/2}\cdot(z^\dagger z+1)^{-n-1}\cdot dx\,dy$$
$$(\text{with } z=x+iy).$$

A similar result holds for (Ψ^1,Ψ^1) but, since $S^2\setminus U_0$ consists of one point which has measure zero, it follows that we only have to deal with the chart U_0 with its holomorphic coordinate $z=x+iy$: omitting Ψ^0 we obtain

$$H=\left\{g\colon \mathbb{C}\to\mathbb{C} \mid g \text{ a polynomial of degree at most } n-1 \text{ in } z,\right.$$
$$\left. 2\left(\frac{n}{2\pi}\right)^{1/2}\int_{\mathbb{C}}|g(z)|^2(z^\dagger z+1)^{-n-1}\,dx\,dy<\infty\right\},$$
$$\langle g,g'\rangle = 2\left(\frac{n}{2\pi}\right)^{1/2}\int_{\mathbb{C}} g(z)^\dagger g'(z)(z^\dagger z+1)^{-n-1}\,dx\,dy.$$

Careful analysis shows that the two conditions on g are equivalent, hence we can say that H consists of all polynomials of degree $n-1$ in z, and indeed, as already said, H is a Hilbert space of dimension n.

The functions $g_k(z) = (n/2\pi)^{1/4}\,(n-1 \text{ over } k)^{1/2}\,z^k$ form an orthonormal system in H and hence the generalized Bergman kernel is given by

$$K(w,z) = \left(\frac{n}{2\pi}\right)^{1/2}\sum_{k=0}^{n-1}\binom{n-1}{k}w^k z^{\dagger k}.$$

Finally let f be an observable (i.e., a real function on S^2) then on the chart U_0:

$$X_f = \frac{(z^\dagger z+1)^2}{2i\lambda}\left(\frac{\partial f}{\partial z}\frac{\partial}{\partial z^\dagger}-\frac{\partial f}{\partial z^\dagger}\frac{\partial}{\partial z}\right),$$
$$[X_f,X_z] = \frac{i}{2\lambda}\frac{\partial((z^\dagger z+1)^2\partial f/\partial z^\dagger)}{\partial z}X_z$$
$$+\frac{i}{2\lambda}\frac{\partial((z^\dagger z+1)^2\partial f/\partial z^\dagger)}{\partial z^\dagger}X_{z^\dagger}$$

hence,

$$L_f(g\Psi^0) = \left[\frac{1}{2n}\frac{\partial((z^\dagger z+1)^2\partial f/\partial z^\dagger)}{\partial z}\right.$$
$$\left.+f-z^\dagger(z^\dagger z+1)\frac{\partial f}{\partial z^\dagger}\right]g(z)\Psi^0$$
$$+\frac{1}{n}(z^\dagger z+1)^2\frac{\partial f}{\partial z^\dagger}\frac{dg}{dz}\Psi^0. \qquad (7.1)$$

Applying Theorem 5.8, using partial integration, and omitting Ψ^0 we find the following prescription for operators.

Prescription: For an observable f on S^2, the corresponding operator on polynomials g of degree $2\lambda/\hbar-1$ is given by

$$(\mathbf{fg})(w) = \frac{n}{\pi}\sum_{k=0}^{n-1}\binom{n-1}{k}w^k\int_{\mathbb{C}}\frac{z^{\dagger k}g(z)}{(z^\dagger z+1)^{n+1}}$$
$$\times\left[f(z,z^\dagger)-\frac{(z^\dagger z+1)^2}{2n}\frac{\partial^2 f}{\partial z\,\partial z^\dagger}\right]dx\,dy.$$

Examples: Let $(a,b,c)\in S^2\subset\mathbb{R}^3$, then the correspondence between the three coordinates (a,b,c) and the complex coordinate z is given by

$$a=(z^\dagger+z)\cdot(z^\dagger z+1)^{-1},\quad b=i(z^\dagger-z)\cdot(z^\dagger z+1)^{-1},$$
$$c=(z^\dagger z-1)\cdot(z^\dagger z+1)^{-1},$$

which represents the projection from the north pole onto the x-y plane. Now we introduce the spin observables $S_1=\lambda a$, $S_2=\lambda b$, and $S_3=\lambda c$; for these observables the expression in (7.1) is already holomorphic, so in this case we do not need to apply the generalized Bergman kernal to obtain the corresponding operators, which are

$$\mathbf{S}_1 = \frac{1}{2}\hbar\left[(1-z^2)\frac{d}{dz}+(n-1)z\right],$$
$$\mathbf{S}_2 = \frac{1}{2}i\hbar\left[(1+z^2)\frac{d}{dz}-(n-1)z\right],$$
$$\mathbf{S}_3 = \frac{1}{2}\hbar\left[2z\frac{d}{dz}-(n-1)\right],$$

whence

$$(\mathbf{S}_2)^2+(\mathbf{S}_2)^2+(\mathbf{S}_3)^2 = \tfrac{1}{4}\hbar^2(n^2-1)\mathbf{1}.$$

When we say that this model with a classical parameter $\lambda=(s+\frac{1}{2})\hbar$ describes a particle with spin s ($2s=0,1,2,\ldots$), then these results are in complete agreement with the usual quantum mechanical description of spin: the Hilbert space of a particle with spin s has dimension $2s+1=2\lambda/\hbar=n$ and the sum of squares of the spin operators is $s(s+1)\hbar^2\mathbf{1}=\frac{1}{4}\hbar^2(n^2-1)\mathbf{1}$. Moreover, if we express the spin operators as matrices with respect to the orthonormal basis g_k introduced above (in *descending* order!), then one recovers the usual Pauli-spin matrices; in particular in the case $n=2$ (spin-$\frac{1}{2}$) one obtains

$$\mathbf{S}_1 = \frac{1}{2}\hbar\begin{pmatrix}0&1\\1&0\end{pmatrix},\quad \mathbf{S}_2 = \frac{1}{2}\hbar\begin{pmatrix}0&-i\\i&0\end{pmatrix},$$
$$\mathbf{S}_3 = \frac{1}{2}\hbar\begin{pmatrix}1&0\\0&-1\end{pmatrix},$$

and in the case $n=3$ (spin-1):

$$\mathbf{S}_1 = \frac{1}{2}\sqrt{2}\hbar\begin{pmatrix}0&1&0\\1&0&1\\0&1&0\end{pmatrix},$$
$$\mathbf{S}_2 = \frac{1}{2}i\sqrt{2}\hbar\begin{pmatrix}0&-1&0\\1&0&-1\\0&1&0\end{pmatrix},$$
$$\mathbf{S}_3 = \hbar\begin{pmatrix}1&0&0\\0&0&0\\0&0&-1\end{pmatrix}.$$

Remark 1: If one computes in the case

$$\lambda=\hbar(\Leftrightarrow n=2\Leftrightarrow s=\tfrac{1}{2})$$

the operators associated to the observable $f=(S_j)^2$ then one finds in all cases $\mathbf{f}=[(\hbar^2)/3]\mathbf{1}$, which is in agreement with the observation that $S_1^2+S_2^2+S_3^2=\hbar^2$. There is no contradiction with the fact that the sum of squares of the spin operators is not equal to the operator associated to the sum of squares of the spin observables because we nowhere showed nor used that the operator corresponding to a product should be the product of the corresponding operators.

Remark 2: Contrary to the opinion stated in Ref. 2, §11.2, p. 205 (see also Ref. 7, §6.3.6]), it is not necessary to change the quantization method to obtain a correct description of the quantized spin. The procedure described above is quite adequate: the classical model with parameter $\lambda(=n\cdot\frac{1}{2}\hbar)$ describes, after quantization, a particle with spin $s=\lambda/\hbar-\frac{1}{2}$. In my opinion it has the definite advantage that can describe a particle *with* spin, *value* zero, i.e., in the classical model there exist spin observables, which yield, after quantization, always the value zero. The alternative is a classical model without an extra sphere in the phase space, hence *without the possibility to measure spin* (except by saying that it does not exist). At this point it should be mentioned that the value $\lambda=0$ is not allowed, because then the symplectic form reduces to zero and is no longer a symplectic form.

VIII. APPLICATION III

In this section we want to analyze the effect of the quantization prescription given in Sec. VI, when translated to the usual Schrödinger representation: $H=$ functions of the position (see also Ref. 2, §8.1 and Ref. 7, §5.11.5). We will do this in the case $M=T^*\mathbb{R}^3=\mathbb{R}^6=\mathbb{C}^3$: the phase space of a single particle in $\mathbb{R}^3$. To avoid confusion with dimensions we introduce (as in Remark 3 of Sec. VI) constants α and β and dimensionless coordinates $x_j=p_j/\alpha$ and $y_j=q_j/\beta$ on M; furthermore we define the complex coordinates $z_j=x_j+iy_j$ (the coordinate z' of Remark 3, Sec. VI), and we introduce the dimensionless constant $\gamma=\alpha\beta/\hbar$. Finally we introduce two polarizations: a "holomorphic" polarization F_h spanned by X_{z_j} and a "vertical" polarization F_v spanned by X_{y_j}.

In these coordinates $\omega=\alpha\beta\,\Sigma_j dx_j\wedge dy_j$ and as in Sec. VI we use the symplectic potential

$$\vartheta=\frac{1}{2}\alpha\beta\sum_j(x_j\,dy_j-y_j\,dx_j)\,.$$

To facilitate the notations we introduce the column vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ with entries x_j, y_j, and z_j; furthermore the symbol ^{T} applied on column vectors will denote transposition and the symbol † will denote transposition *and* complex conjugation, hence $\|\mathbf{z}\|^2=\mathbf{z}^\dagger\mathbf{z}=\mathbf{x}^T\mathbf{x}+\mathbf{y}^T\mathbf{y}=\|\mathbf{x}\|^2+\|\mathbf{y}\|^2\in\mathbb{R}$.

The bundles $L, R^F, R^{\sim F}, R^{-F}$, and hence QB are trivial for both polarizations, so we identify sections with functions on M. With the symplectic potential ϑ we compute the F-constant sections of L according to (1.2):

$$\Psi\ F_v\ \text{constant}\Leftrightarrow\Psi(\mathbf{x},\mathbf{y})=\Psi_v(\mathbf{y})\exp(-\tfrac{1}{2}i\gamma\mathbf{x}^T\mathbf{y}),$$

$$\chi\ F_h\ \text{constant}\Leftrightarrow\chi(\mathbf{x},\mathbf{y})=\chi_h(\mathbf{z})\exp(-\tfrac{1}{4}\gamma\mathbf{z}^\dagger\mathbf{z})$$

$$\text{and } \chi_h \text{ holomorphic in } \mathbf{z}.$$

Applying the theory of Sec. III we obtain two Hilbert spaces H_v and H_h given by

$$H_v=\left\{\Psi_v(\mathbf{y})\,\middle|\,\int_{\mathbb{R}^3}|\Psi_v(\mathbf{y})|^2\,d\mathbf{y}<\infty\right\},$$

$$H_h=\left\{\chi_h(\mathbf{z})\,\middle|\,(\gamma/\pi)^{3/2}\int_{\mathbb{C}^3}|\chi_h(z)|^2\times\exp\left(-\frac{1}{2}\gamma\mathbf{z}^\dagger\mathbf{z}\right)d\mathbf{x}\,d\mathbf{y}<\infty\right\}.$$

Since the polarizations F_v and F_h satisfy conditions (2.1) with $D^\wedge=\{0\}$ and $E^\wedge=T^*M$, there exists a pairing between H_v and H_h given by [using formula (2.4)]:

$$\langle\Psi_v,\chi_h\rangle=\left(\frac{1}{2}\frac{i\gamma}{\pi}\right)^{3/2}\int_{\mathbb{C}^3}\left(\Psi_v(\mathbf{y})\exp\left(-\frac{1}{2}i\gamma\mathbf{x}^T\mathbf{y}\right)\right)^\dagger\times\chi_h(z)\exp\left(-\frac{1}{4}\gamma\mathbf{z}^\dagger\mathbf{z}\right)d\mathbf{x}\,d\mathbf{y}.$$

This pairing defines unitary maps $U_{vh}: H_h\to H_v$ and $U_{hv}: H_v\to H_h$ by

$$\langle\Psi_v,U_{vh}\chi_h\rangle_{Hv}=\langle\Psi_v,\chi_h\rangle=\langle U_{hv}\Psi_v,\chi_h\rangle_{Hh},\qquad(8.1)$$

which are given explicitly by the formulas

$$(U_{vh}\chi_h)(\mathbf{y})=\left[e^{\pi i/4}\left(\frac{\frac{1}{2}\gamma}{\pi}\right)^{1/2}\right]^3\times\int\chi_h(\mathbf{z})\exp\left(-\frac{1}{4}\gamma[\mathbf{z}^\dagger\mathbf{z}-2i\mathbf{x}^T\mathbf{y}]\right)d\mathbf{x},$$

$$(U_{hv}\Psi_v)(\mathbf{z})=\left[e^{-\pi i/4}\left(\frac{\frac{1}{2}\gamma}{\pi}\right)^{1/2}\right]^3\int\Psi_v(s)\times\exp\left(-\frac{1}{4}\gamma[-\mathbf{z}^T\mathbf{z}+4i\mathbf{s}^T\mathbf{z}+2\mathbf{s}^T\mathbf{s}]\right)d\mathbf{s}.$$

That these formulas are indeed given by (8.1), that they are unitary, and that they are inverse to each other can be verified by using the reproducing formula (6.1), the Fourier integral $\int\exp(ixy)dx=2\pi\delta(y)$ and the Gaussian integral $\int\exp(-\pi x^2/a)dx=\sqrt{a}$ (for $a>0$).

With these ingredients we can translate the prescription of Sec. VI as given by formula (6.2) to the usual Schrödinger quantization: let $f=f(\mathbf{x},\mathbf{y})$ be an arbitrary observable and let $\Psi_v\in H_v$, then

$$\mathbf{f}\Psi_v=U_{vh}(\mathbf{f}(U_{hv}\Psi_v))=U_{vh}((L_f(U_{hv}\Psi_v))_{\text{hol}})$$

or, more explicitly,

$$(\mathbf{f}\Psi_v)(\mathbf{y})=\left[\frac{1}{2}\left(\frac{\gamma}{\pi}\right)^{3/2}\right]^3\iiint\exp(i\gamma\mathbf{r}^T(\mathbf{y}-\mathbf{t}))\times\exp(-\tfrac{1}{2}\gamma[\|\mathbf{t}-\mathbf{s}\|^2+\|\mathbf{y}-\mathbf{s}\|^2]\times(f(\mathbf{r},\mathbf{s}-(4\gamma)^{-1}[\Delta_r+\Delta_s]f)\times\Psi_v(\mathbf{t})d\mathbf{r}\,d\mathbf{s}\,d\mathbf{t}.\qquad(8.2)$$

Example I: If the observable f does not depend on $\mathbf{x}$, i.e., it depends only on the position coordinates, then the integration over $\mathbf{r}$ and $\mathbf{t}$ in (8.2) can be performed and one obtains

$$(\mathbf{f}\Psi_v)(\mathbf{y})=\left[\left(\frac{\gamma}{\pi}\right)^{3/2}\int(f-(4\gamma)^{-1}\Delta f)(\mathbf{s})\times\exp(-\gamma\|\mathbf{y}-\mathbf{s}\|^2)d\mathbf{s}\right]\cdot\Psi_v(\mathbf{y}),$$

in other words $\mathbf{f}$ is multiplication with a function $f^0(\mathbf{y})$ defined by

$$f^0(\mathbf{y})=\left(\frac{\gamma}{\pi}\right)^{3/2}\int(f-(4\gamma)^{-1}\Delta_s f)(\mathbf{s})\times\exp(-\gamma\|\mathbf{y}-\mathbf{s}\|^2)d\mathbf{s}$$
$$=(1-(4\gamma)^{-1}\Delta_y)\times\left[\left(\frac{\gamma}{\pi}\right)^{3/2}\int f(\mathbf{s})\exp(-\gamma\|\mathbf{y}-\mathbf{s}\|^2)d\mathbf{s}\right].\qquad(8.3)$$

We now notice that the constant $1/\gamma$ is usually very small ($\alpha = 1$ m, $\beta = 1$ kg·m/sec $\Rightarrow \gamma \approx 10^{34}$), so if we neglect for the moment the additional term $(4\gamma)^{-1}\Delta$ in (8.3) then f^0 is the convolution of f with a Gaussian curve of width $\approx 1/\gamma$, i.e., roughly speaking f^0 is the average of the potential f over a region of dimension $1/\gamma$ around the point $\mathbf{y}$. This prescription of replacing the potential by such an average is sometimes used in quantum mechanics to explain certain correction terms (e.g., the Darwin term in the Hamiltonian of the hydrogen atom, see Ref. 9 with the explanation that the electron is not a mathematical point). Let us investigate in more detail the integral (8.3), to see its effect upon different potentials:

$f = 1$	y_1	$y_2 y_3$	$(y_1)^2$	$(y_2)^3$	$(y_2)^4$
$f^0 = 1$	y_1	$y_2 y_3$	$(y_1)^2$	$(y_2)^3$	$(y_2)^4 - 3(2\gamma)^{-2}$.

Another interesting potential is the Coulomb potential $f(\mathbf{y}) = \|\mathbf{y}\|^{-1}$ for which one obtains

$$f^0(\mathbf{y}) = \|\mathbf{y}\|^{-1}\left(\frac{\gamma}{\pi}\right)^{1/2}\int_{-\|\mathbf{y}\|}^{\|\mathbf{y}\|}\exp(-\gamma\tau^2)d\tau + \left(\frac{\gamma}{\pi}\right)^{1/2}\exp(-\gamma\|\mathbf{y}\|^2)$$
$$= \|\mathbf{y}\|^{-1}\cdot\operatorname{erf}(\sqrt{\gamma}\|\mathbf{y}\|) + (\gamma/\pi)^{1/2}\exp(-\gamma\|\mathbf{y}\|^2),$$

and we see that indeed f^0 differs from f only in a region of dimension $1/\gamma$.

Example 2: In this example we consider observables which are linear in the momentum (i.e., linear in $\mathbf{x}$), so suppose $f(\mathbf{x},\mathbf{y}) = x_j \cdot g(\mathbf{y})$, then one finds after partial integration in (8.2) with respect to the variable $\mathbf{t}$:

$$(\mathbf{f}\Psi_v)(\mathbf{y}) = -\frac{i}{\gamma}\frac{\partial\Psi_v}{\partial y_j}(\mathbf{y})\left[\left(\frac{\gamma}{\pi}\right)^{3/2}\int (g-(4\gamma)^{-1}\Delta g)(\mathbf{s})\exp(-\gamma\|\mathbf{y}-\mathbf{s}\|^2)d\mathbf{s}\right]$$
$$+\Psi_v(\mathbf{y})\cdot\left[\frac{-\frac{1}{2}i}{\gamma}\cdot\left(\frac{\gamma}{\pi}\right)^{3/2}\int\frac{\partial(g-(4\gamma)^{-1}\Delta g)}{\partial\mathbf{s}_j}\exp(-\gamma\|\mathbf{y}-\mathbf{s}\|^2)d\mathbf{s}\right],$$

or without the wave function Ψ_v:

$$\mathbf{f} = \left[\left(\frac{\gamma}{\pi}\right)^{3/2}\int (g-(4\gamma)^{-1}\Delta g)(\mathbf{s})\exp(-\gamma\|\mathbf{y}-\mathbf{s}\|^2)d\mathbf{s}\right]\frac{-i}{\gamma}\frac{\partial}{\partial y_j}$$
$$+\left[\frac{-\frac{1}{2}i}{\gamma}\cdot\left(\frac{\gamma}{\pi}\right)^{3/2}\int\frac{\partial(g-(4\gamma)^{-1}\Delta g)}{\partial s_j}\exp(-\gamma\|\mathbf{y}-\mathbf{s}\|^2)d\mathbf{s}\right]. \tag{8.4}$$

There are several interesting possibilities, of which we will study only two: the case $f = x_j$ (i.e., the linear momentum) and the case $f = y_j x_k - y_k x_j$ (i.e., the angular momentum). Evaluating the integrals in (8.4) we see that the second term vanishes in both cases and (after some calculations) we obtain

$$f = x_j \qquad \Rightarrow \mathbf{f} = \frac{-i}{\gamma}\frac{\partial}{\partial y_j},$$
$$f = y_j x_k - y_k x_j \qquad \Rightarrow \mathbf{f} = \frac{-i}{\gamma}\left(y_j\frac{\partial}{\partial y_k} - y_k\frac{\partial}{\partial y_j}\right)$$

or, reintroducing $p_j = \alpha x_j$, $q_j = \beta\gamma_j$ and $\gamma = \alpha\beta/\hbar$:

$$f = p_j \qquad \Rightarrow \mathbf{f} = -i\hbar\frac{\partial}{\partial q_j},$$
$$f = q_j p_k - q_k p_j \qquad \Rightarrow \mathbf{f} = -i\hbar\left(q_j\frac{\partial}{\partial q_k} - q_k\frac{\partial}{\partial q_j}\right).$$

Example 3: We could go on with higher powers of $\mathbf{x}$, but the calculations become more and more complicated. However, the observable $f = \frac{1}{2}\|\mathbf{x}\|^2$ is interesting enough to calculate; after twice integrating by parts in (8.2) one finally obtains

$$\mathbf{f} = -(2\gamma^2)^{-1}\Delta_{\mathbf{y}},$$

or, in other words, the kinetic energy $\frac{1}{2}\|\mathbf{p}\|^2$ is represented by $-\frac{1}{2}\hbar^2\Delta_q$.

IX. FINAL REMARKS

Remark 1: It should be noted that the theory of generalized Bergman kernels can be applied as well to geometric quantization using $(-\frac{1}{2})$-F-densities instead of $(-\frac{1}{2})$-F-forms, because in that case too, two F-constant sections of QB differ by a function which is holomorphic on the leaves of $\pi_* E$.

Remark 2: The difference between the use of $(-\frac{1}{2})$-F-densities and $(-\frac{1}{2})$-F-forms in the first application is that in prescriptions I and II the expression $f - \hbar\,\partial^2 f/\partial z\,\partial z^\dagger$ should be replaced by $f - 2\hbar\,\partial^2 f/\partial z\,\partial z^\dagger$.

In the second application the difference is that (a) in the description of the Hilbert space H, the "parameter" n should be replaced by $n+1$ (so the dimension of H becomes $n+1$ instead of n) and (b) in the prescription for $\mathbf{f}$ the expression $f - (z^\dagger z+1)^2/2n\,\partial^2 f/\partial z\,\partial z^\dagger$ should be replaced by $f - (z^\dagger z+1)^2/n\,\partial^2 f/\partial z\,\partial z^\dagger$ (N.B. here n should *not* be replaced by $n+1$).

If one now calculates the operators $\mathbf{S}_j$, one obtains the same Pauli-spin matrices, except for a different value of n: $n=1$ now represents $s=\frac{1}{2}$, $n=2$ represents $s=1$, etc., in accordance with $(\mathbf{S}_1)^2 + (\mathbf{S}_2)^2 + (\mathbf{S}_3)^2 = \frac{1}{4}\hbar^2\, n(n+1)\mathbf{1}$. Moreover, for $n=1$ ($\Leftrightarrow s=\frac{1}{2}\Leftrightarrow\lambda=\frac{1}{2}\hbar$) the operators associated to S_j^2 are all equal to $(\hbar^2/12)\mathbf{1}$.

The differences in the third application are all due to the differences as described for application I; the connection between H_v and H_h remains the same.

ACKNOWLEDGMENTS

I would like to thank J. Wiegerinick for listening to my problems and telling me there existed things like Bergman kernels. I am also indebted to R. Brummelhuis and Professor E. Thomas for their valuable remarks.

[1]D. J. Simms and N. M. J. Woodhouse, *Lectures on Geometric Quantization, Lecture Notes in Physics*, Vol. 53 (Springer, Berlin, 1977).
[2]J. Sniatycki, *Geometric Quantization and Quantum Mechanics, Applied Mathematical Sciences*, Vol. 30 (Springer, Berlin, 1980).
[3]G. M. Tuynman, *Proceedings of the Seminar 1983–1985 Mathematical Structures in Field Theories, Vol. I: Geometric Quantization* (CWI, Amsterdam, 1985), CWI-syllabus 8.
[4]J. H. Rawnsley, "On the pairing of polarizations," Comm. Math. Phys. **58**, 1 (1978).
[5]J. H. Rawnsley, "Half-forms," preprint CNRS-Luminy (Marseille) 80/PE.1248, 1980.
[6]V. Guillemin and S.Sternberg, *Geometric Asymptotics, Mathematical Surveys*, Vol. 14 (American Mathematical Society, Providence, RI, 1977).
[7]N. Woodhouse, *Geometric Quantization* (Clarendon, Oxford, 1980).
[8]J. -M. Souriau, *Structures des systèmes dynamiques* (Dunod, Paris, 1969).
[9]C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Mécanique quantique II* (Hermann, Paris, 1973). [Translated as *Quantum Mechanics II* (Wiley, New York, 1977).]

Loop gauge theory and group cohomology

A. Vourdas[a)]
DAMTP, University of Liverpool, P. O. Box 147, Liverpool L67 3BX, United Kingdom

(Received 9 July 1985; accepted for publication 29 October 1986)

A generalized fiber bundle model in which the fibers are Hilbert spaces is studied. Unitary transformations are used to define a unitary isomorphism ("parallelism") among them. The Weyl group is first used to "connect" the projective Hilbert spaces (ray spaces) and to introduce a one-form connection that defines which coset at a point y is parallel to a given coset at another point x. Then, the central extension of the Weyl group by U(1)′ is studied in order to introduce the most general mapping between the elements of these cosets. This leads to a two-form connection and makes the model a good candidate for a fiber bundle approach to string theories.

I. INTRODUCTION

Gauge theory in physics is similar to fiber bundle theory in geometry. At each point of the base manifold (space-time) we have a fiber isomorphic to a group G. The connection is a one-form which defines an isomorphism ("parallelism") between the G at a point x and the G at a different point y. The one-form connection and the corresponding two-form curvature are interpreted in physics as the potential and the gauge field. In this paper we study a more general fiber bundle type of model. We consider Hilbert spaces as fibers and introduce unitary transformations in order to define a unitary isomorphism ("parallelism") among them.

Our model is the generalized version of scalar electrodynamics (e.g., Ref. 1) which includes not only the $\pm e$ charges but also all the higher charges $\pm 2e$, $\pm 3e$, etc. The wave function is $\phi(x,\theta)$, where θ is a coordinate for the U(1) gauge group. The electric charge is treated in a quantum mechanical way (like the momentum). The charge operator is $\hat{q} = ie\,\partial_\theta$ and the e plays the role of Planck's constant for the θ dimension. Periodicity in θ $(\phi(x,\theta + 2\pi) = \phi(x,\theta))$ gives a discrete charge spectrum. We regard the $\phi(x,\theta)$ as a collection of wave functions $\phi_x(\theta)$ at the various points x of four-dimensional space-time. At each point x we introduce a Hilbert space

$$H_x = \{\text{complex periodic functions of } \theta\}.$$

The unitary transformations will be used to "connect" (i.e., to define a unitary isomorphism between) the Hilbert spaces at the various points x. The group of unitary transformations is[2]

$$G = \{\exp(i\alpha\hat{q})\exp(iN\hat{\theta})\exp(i\gamma)\}. \tag{1}$$

We start with the projective Hilbert spaces PH_x (or ray spaces in Weyl's terminology). In PH the coset $\{\exp(i\gamma) \times|\theta\rangle|\text{arbitrary } \gamma, \text{ fixed } \theta\}$ represents one element. Kets with different quantum phase represent the same element. The group of unitary transformations for this space is $W = G\,|\mathrm{U}(1)'(\mathrm{U}(1)' = \{\exp(i\gamma)\})$ and is known as Weyl group of quantum canonical transformations.[2] The U(1)′ represents the "quantum phase" and is the center of the group G. We call it U(1)′ in order to distinguish it from the original U(1) group, which is in our model the θ dimension.

Local Weyl transformations will lead to the potential operator $A_\mu\,\hat{q}$. The coset $I_1 = \{\exp(i\gamma)|\theta\rangle|\text{arbitrary } \gamma\}$ at x, is now parallel to the coset $I_2 = \{\exp(i\gamma)|\theta + \int_{x(c)}^{y} A_\mu\,\delta x_\mu\rangle|$ arbitrary $\gamma\}$ at y. This expresses the well-known fact that parallel transport along a curve C changes the U(1) phase by the path-dependent quantity $\int_{x(c)}^{y} A_\mu\,\delta x_\mu$. Up to this point our model simply contains the structure of a standard gauge theory (a one-form connection). Since, however, the I_1 and I_2 are cosets we can introduce a mapping between the elements of I_1 and the elements of I_2. If we do that in the most general way, we are led to a new two-form connection (and the corresponding three-form curvature).

In order to introduce this mapping in the most general way, we explore all the groups G that can be constructed from a given W and U(1)′. The only constraints are that $W = G\,|\mathrm{U}(1)'$ and that U(1)′ is the center of G. The subject of central extensions[3,4] explores this problem. In general the group G will have the elements given in (1) with the multiplication rule

$$\begin{aligned} g_1 g_2 &= \{\exp(i\alpha_1 q)\exp(iN_1\theta)\exp(i\gamma_1)\} \\ &\quad\times\{\exp(i\alpha_2 q)\exp(iN_2\theta)\exp(i\gamma_2)\} \\ &= \exp\{i(\alpha_1+\alpha_2)q\}\exp\{i(N_1+N_2)\theta\} \\ &\quad\times\exp\{i(\gamma_i+\gamma_2+\sigma(\alpha_1,N_1;\alpha_2,N_2))\}. \end{aligned} \tag{2}$$

The $\sigma(\alpha_1,N_1;\alpha_2,N_2)$ is called a factor set and is restricted by the associativity requirement.

Our basic assumption is that there is no preference to a particular extension and that we should try to construct a theory, covariant under transformations with any of the multiplication rules given in (2). The fundamental, for fiber bundle theory, concept of parallelism has been broadened in our model; at each point of space-time we introduce a whole class of groups G, all with the same elements but different multiplication rules. We then show that in order to define a mapping ("parallelism") between these multiplication rules at the various points of space-time, a two-form "connection" is required. If $g(x)$ is an element of the group G, then $g^{-1}(x)\partial_\mu g(x)$ depends on the multiplication rule, i.e., on the extension. For a given $g(x)$ and under a transformation from one extension into another

$$g^{-1}\,\partial_\mu\, g \to g^{-1}\,\partial_\mu\, g + i\Lambda_\mu.$$

The Λ_μ is a quantity that we will calculate. This transforma-

[a)] Present address: Fachbereich Physik, Universität Marburg, Mainzer Gasse 33, D-3550 Marburg, West Germany.

 0022-2488/87/030584-08$02.50

tion of the extension introduces loop gauge transformations[5] and consequently a two-form potential $A_{\mu\nu}$ and a three-form gauge field $f_{\mu\nu\lambda}$. In this sense we have a particular type of a string model. Our strings are simply tubes of magnetic flux. This is the original but a very particular interpretation and application of string theory. Another interpretation (following Ref. 6) uses the lower modes of the infinite spectrum, which necessarily exists in every string theory, to describe all the existing physical particles and therefore unify all the physical theories. Our model contains of course an infinite spectrum but it is an infinite spectrum of electric and magnetic charges.

The standard fiber bundle theory (i.e., the standard local gauge theory) is too restrictive and cannot accommodate quantized monopoles and quantized magnetic strings. The Bianchi identity $\partial_\mu {}^*f_{\mu\nu}=0$ does not allow magnetic sources. Magnetic monopoles[7] are introduced as line singularities in space-time. They create a nontrivial topology which is able to accommodate many cohomology classes of $f_{\mu\nu}$. With this argument we get the standard Dirac[7] currents

$$J_{\mu\nu}=\int \delta(x-y)\{y^\mu(\tau,\sigma),y^\nu(\tau,\sigma)\}d\tau\, d\sigma, \tag{3}$$

$$\partial_\mu J_{\mu\nu}=J'_\nu=\int \delta(x-y)\frac{dy^\mu(\tau)}{d\tau}d\tau, \tag{4}$$

which are semiclassical. The string follows one particular world surface $y^\mu(\tau,\sigma)$ and not all the surfaces as quantum physics would require. The $J_{\mu\nu}$ is given in (3) in terms of this particular surface $y^\mu(\tau,\sigma)$ and not in terms of a wave function. The magnetic monopole is also semiclassical and it follows only one world line ($y^\mu(\tau)\equiv y^\mu(\tau,\sigma=0)$) in space-time and not all the world lines as quantum physics would require. The magnetic current of Eq. (4) is given in terms of a δ function of the position and not in terms of a monopole wave function.

The quantization of string theory is currently under development and is discussed from various points of view. One approach is the Polyakov[8] functional integration over surfaces. Another more recent approach[9] is based on BRST (Becchi–Rouet–Stora–Tyutin) invariance. The connection of this approach with Connes' noncommutative geometry has been discussed in Ref. 10. A fiber bundle type of approach to string theory is desirable for theoretical and practical reasons. It will state clearly and explicitly its symmetry content; this should be the starting point of the theory, but historically it was developed in a different way and its symmetry content is rather unclear. It will also show how from fundamental assumptions we are led to the physics of string theory, in analogy with gauge theories and gravity. For mathematicians it will open new directions to explore. Here we study a generalization of fiber bundles which uses Hilbert spaces as fibers and which leads to the two-form connection and the two-form current that characterize string theories. We show that the currents $J_{\mu\nu}$ and J'_ν have not the semiclassical form of Eqs. (3) and (4) but that they are expressed in terms of the field $\phi(x,\theta)$. Therefore we claim that our model is a good candidate for a fiber bundle approach to string theories. It is quite clear that the aspects of quantized string that we discuss here are quite different from those in Ref. 8 or those in Ref. 9. At this stage it is not explicitly clear the connection among them; this is a more difficult task.

We conclude this section by sketching the relation between our work and Ref. 11 which also played an important role in the development of the subject. The later is on SU(N) gauge theories but it is known[12] that there is a relation between string theories and the SU(N) ($N\to\infty$) gauge theory. Let us discretize our θ position space, by taking N points uniformly distributed in the θ dimension. We have in this case quantum mechanics on a discrete position space of N points (Ref. 2, Schwinger, p. 63, Weyl, § IV.14). The momentum space is discrete and contains N momenta, i.e., we have N charges, $1e,\ldots, Ne$ (defined mod N). The quantum phase is in this case an element of Z_N. The wave function becomes $\phi(x,K)$, $K=1,\ldots,N$, and the Hilbert space H_x is finite (N) dimensional. The group of unitary transformations is U(N) and the Weyl group U(N)$|Z_N$. From a mathematical point of view this model is similar to the U(N) gauge theory. In Ref. 11, the quantum mechanical algebra that is appropriate for this case,

$$V^lU^k=U^kV^l\exp\{i(kl)/N\},\quad k,l \text{ integers},$$
$$U^N=V^N=1,$$

has been used, to introduce general pseudoperiodic boundary conditions in a finite box. A finite number of different classes have been found. Our model is in the $N\to\infty$ limit (Ref. 2, Schwinger, p. 259, Weyl, § IV.15). We use the standard quantum mechanical algebra for a circle position space. Our momentum space is discrete and infinite and our quantum mechanical phase is an element of the group U(1)'. The extensions of W by U(1)' introduce a similar effect with the nontrivial boundary conditions in Ref. 11. Indeed, we study $H^2(W,\mathrm{U}(1)')$ in Sec. II and we find an infinite number of classes.

We finally mention Ref. 13 where obstruction to group extension (which is related[3] to H^3) has been used in the study of magnetic monopoles. The associativity and consequently the Jacobi and Bianchi constraints are violated. In our paper we respect associativity and we explore $H^2=Z^2|B^2$.

II. EXTENSIONS AND COHOMOLOGY OF THE WEYL GROUP

The Weyl group for a circle position space is[2] $W=G|\mathrm{U}(1)'$,

$$G=\{\exp(i\alpha\hat{q})\exp(iN\hat{\theta})\exp(i\gamma)\},$$
$$\mathrm{U}(1)'=\{\exp(i\gamma)\},\quad \hat{q}=ie\,\partial_\theta,\quad N \text{ integer}. \tag{5}$$

Here, W is an Abelian group with elements the cosets

$$w_1=w(N_1,\alpha_1)=\{\exp(i\alpha_1\hat{q})\exp(iN_1\hat{\theta})\times\exp(i\gamma)|\gamma \text{ arbitrary}\}, \tag{6}$$
$$w(N_1,\alpha_1)w(N_2,\alpha_2)=w(N_1+N_2,\alpha_1+\alpha_2).$$

We are going to search for all the nontrivial ways of reconstructing G from a certain $W=G|\mathrm{U}(1)'$ and U(1)'. This problem is known in the mathematical literature as group extension.[3,4] Here we are only interested in extensions where the U(1)' is the center of G(central extensions). We

use the notation $g_1 = w_1 \exp(i\gamma_1)$ for an element of G, where w_1 is an element of W and $\exp(i\gamma_1)$ an element of $U(1)'$. We also use the notation $g_1 = \exp(i\alpha_1\hat{q})\exp(iN_1\hat{\theta})\exp(i\gamma_1)$ keeping in mind that the $\exp(i\alpha_1\hat{q})\exp(iN_1\hat{\theta})$ is the coset of Eq. (6). Another notation used in the literature is $(w_1, \exp i\gamma_1)$ where it is clear that the $w_1 \exp(i\gamma_1)$ is not a product in the ordinary sense. The most general way to define multiplication between two elements of G is

$$g_1g_2 = (w_1 \exp i\gamma_1)(w_2 \exp i\gamma_2)$$
$$= w_1w_2 \exp i(\gamma_1 + \gamma_2 + \sigma(w_1,w_2)), \tag{7}$$

which we have already presented in Eq. (2). The $\sigma(w_1,w_2)$ is called a factor set and is an arbitrary real function restricted (i) by the associativity rule, which implies

$$\sigma(w_1,w_2) + \sigma(w_1w_2,w_3) = \sigma(w_1,w_2w_3) + \sigma(w_2,w_3); \tag{8}$$

and (ii) by $\sigma(1,w) = \sigma(w,1) = 0$, where 1 is the unit element of the group W, i.e., the coset $w(\alpha = 0, N = 0)$. We can have even more general multiplication rules, but here we only study the limited case of central extensions. The function $\sigma(w_1,w_2)$ is a two-cocycle. The definition of two-cocycle is[3]

$$\delta\sigma = \sigma(w_1,w_2) + \sigma(w_1w_2,w_3) - \sigma(w_1,w_2w_3)$$
$$- \sigma(w_2,w_3) = 0, \tag{9}$$

and is precisely the associativity requirement (8); δ is the coboundary operator and $\delta^2 = 0$. We call $Z^2(W,U(1)')$ the group of two-cocycles.

Let $\tau(w)$ be an arbitrary function with $\tau(w = 1) = \tau(\alpha = 0, N = 0) = 0$. The $\sigma(w_1,w_2) = \tau(w_1) + \tau(w_2) - \tau(w_1w_2)$ obeys the requirements (i) and (ii) and is a special case of a two-cocycle. In fact it is by definition a two-coboundary

$$\delta\tau = \tau(w_1) + \tau(w_2) - \tau(w_1w_2). \tag{10}$$

We call $B^2(W,U(1)')$ the group of two-coboundaries. The two-cohomology group is $H^2(W,U(1)') = Z^2(W,U(1)')/B^2(W,U(1)')$. Each factor set $\sigma(w_1,w_2)$ [defined up to $\tau(w_1) + \tau(w_2) - \tau(w_1w_2)$] characterizes a two-cohomology class.

The commutator of two elements of the group G is now

$$[g_1,g_2] \equiv g_1^{-1}g_2^{-1}g_1g_2 = \exp[iA(w_1,w_2)], \tag{11}$$
$$A(w_1,w_2) = -A(w_2,w_1) = \sigma(w_1,w_2) - \sigma(w_2,w_1).$$

For central extensions[3]

$$[g_1g_2,g_3] = [g_1,g_3]\,[g_2,g_3],$$

therefore

$$A(w_1w_2,w_3) = A(w_1,w_3) + A(w_2,w_3), \tag{12}$$

which we rewrite as

$$A(N_1 + N_2,\alpha_1 + \alpha_2;N_3,\alpha_3)$$
$$= A(N_1,\alpha_1;N_3,\alpha_3) + A(N_2,\alpha_2;N_3,\alpha_3).$$

From this and the relation $A(w,1) = A(1,w) = 0$ we conclude that $A(w_1,w_2)$ is a multiple of $(N_1\alpha_2 - N_2\alpha_1)$,

$$A(w_1,w_2) = me(N_1\alpha_2 - N_2\alpha_1). \tag{13}$$

We see that we get the "expected" result $e(N_1\alpha_2 - N_2\alpha_1)$ with an extra factor m. For a noncompact dimension θ, the m can be any real number. This result is known in the literature (e.g., Ref. 4). In our case θ is a compact dimension and we require that for $\alpha_1 = 2\pi$, $N_1 = 0$, $\alpha_2 = 0$, $N_2 = 1$, the $[g_1,g_2] = 1 \to A(w_1,w_2) = 2\pi M$ (M integer). We get the relation $em = M$. This is the Dirac quantization condition in our model.

We have proved that the commutator of two elements of the group G is

$$[\exp(i\alpha_1\hat{q})\exp(iN_1\hat{\theta})\exp(i\gamma_1),$$
$$\exp(i\alpha_2\hat{q})\exp(iN_2\hat{\theta})\exp(i\gamma_2)]$$
$$= \exp[i(em)(\alpha_2N_1 - \alpha_1N_2)], \quad em = M. \tag{14}$$

Taking into account that a two-coboundary is symmetric in w_1, w_2 [Eq. (10)] we conclude that each m characterizes a cohomology class. This class contains all the extensions with factor sets

$$\sigma(w_1,w_2) = me(N_1\alpha_2 - N_2\alpha_1) + \tau(\alpha_1,N_1) + \tau(\alpha_2,N_2)$$
$$- \tau(\alpha_1 + \alpha_2,N_1 + N_2). \tag{15}$$

Therefore the $H^2(W,U(1)')$ is at least equal to the set of integers Z. It remains to be explored if there are symmetric factor sets $\sigma(w_1,w_2) = \sigma(w_2,w_1)$ that are not two-coboundaries.

Using Eq. (14) we can easily prove

$$\exp(-iN\hat{\theta})\hat{q}\exp(iN\hat{\theta}) = \hat{q} - (em)N. \tag{16}$$

The unit element $m = 0$ of $H^2(W,U(1)')$ contains all the extensions with factor sets the two-coboundaries [elements of $B^2(W,U(1)')$]. In this case the commutator [Eq. (14)] is equal to 1.

We should point out that the problem of central extension of the group W by the $U(1)'$ is equivalent to the problem of projective representations of the group W. A projective representation is defined as the ordinary representation with an extra phase factor

$$P(w_1)P(w_2) = \exp[i\sigma(w_1,w_2)]P(w_1w_2).$$

Assume now that the elements of the group G depend on x and that they are independent of θ. In other words, the $\alpha(x)$, $\gamma(x)$ are functions of x and independent of θ. The N is of course an integer independent of x, θ. Assume also that the factor set $\sigma(x,w_1(x), w_2(x))$ is a two-coboundary that depends explicitly on x. We expect that the derivative of $g(x)$ will depend on the multiplication rule that we choose. We calculate the

$$g^{-1}(x)\partial_\mu g(x) = \lim_{\Delta x_\mu \to 0} \frac{g^{-1}(x)g(x + \Delta x) - 1}{\Delta x_\mu}, \tag{17}$$

where

$$g(x) = \exp(i\alpha\hat{q})\exp(iN\hat{\theta})\exp(i\gamma),$$
$$g^{-1}(x) = \exp(-i\alpha\hat{q})\exp(-iN\hat{\theta})$$
$$\times\exp\{-i[\gamma + \sigma(-\alpha,-N;\alpha,N)]\},$$

$$g(x+\Delta x)=\exp[i(\alpha+\delta\alpha)\hat{q}]\times\exp(iN\theta)\exp[i(\gamma+\delta\gamma)],$$
$$g^{-1}(x)g(x+\Delta x)=\exp(i\delta\alpha\hat{q})\times\exp i[\delta\gamma-\sigma(-\alpha,-N;\alpha,N)+\sigma(-\alpha,-N;\alpha+\delta\alpha,N)]. \tag{18}$$

The σ is a two-coboundary and, according to (10),

$$\sigma(-\alpha,-N;\alpha,N)=\tau(-\alpha,-N)+\tau(\alpha,N),$$
$$\sigma(-\alpha,-N;\alpha+\delta\alpha,N)=\tau(-\alpha,-N)+\tau(\alpha+\delta\alpha,N)-\tau(\delta\alpha,0)=\tau(-\alpha,-N)+[\tau(\alpha,N)+\delta\alpha\,\partial_\alpha\tau(\alpha,N)+\cdots]-[\delta\alpha\,\partial_\alpha\,\tau(0,0)+\cdots]. \tag{19}$$

The $\tau(\alpha=0,N=0)=0$ but the $\partial_\alpha\,\tau(\alpha,N)$ at the point $\alpha=0, N=0$ is not necessarily zero. Therefore

$$g^{-1}(x)g(x+\Delta x)=\exp(i\delta\alpha\hat{q})\times\exp i[\delta\gamma+\delta\alpha(\partial_\alpha\,\tau(\alpha,N)-\partial_\alpha\,\tau(0,0))+\cdots] \tag{20}$$

and from Eq. (17) we conclude

$$g^{-1}(x)\partial_\mu\,g(x)=i\,\partial_\mu\,\alpha\hat{q}+i(\partial_\mu\,\gamma+\beta\,\partial_\mu\,\alpha), \tag{21}$$
$$\beta=\partial_\alpha\,\tau(\alpha,N)-\partial_\alpha\,\tau(0,0). \tag{22}$$

The terms $(\partial_\mu\,\alpha)\hat{q}$, $\partial_\mu\,\gamma$ are familiar but we also have an "extra" term $\beta\,\partial_\mu\,\alpha$ due to the nontrivial multiplication rule. If $\tau(\alpha(x),N)$ does not depend explicitly on x the $\beta\,\partial_\mu\,\alpha$ is an exact one-form, i.e.,

$$\beta\,\partial_\mu\,\alpha=\partial_\mu\,\tau(\alpha(x),N)-\partial_\mu\,\tau(\alpha(x)=0,N=0).$$

The interesting case is when $\beta\,\partial_\mu\,\alpha$ is not an exact form and the exterior differentiation gives a nonzero result. From (21) we can easily see that

$$\partial_\mu(g^{-1}\,\partial_\nu\,g)-\partial_\nu(g^{-1}\,\partial_\mu\,g)=\partial_\mu(i\beta\,\partial_\nu\,\alpha)-\partial_\nu(i\beta\,\partial_\mu\,\alpha)=i[\partial_\mu\,\beta\,\partial_\nu\,\alpha-\partial_\nu\,\beta\,\partial_\mu\,\alpha]\equiv i\frac{\partial(\beta,\alpha)}{\partial(x^\mu,x^\nu)}. \tag{23}$$

This is by definition an exact two-form and therefore is closed, i.e.,

$$\sum_{\text{cycl}}\partial_\rho\frac{\partial(\beta,\alpha)}{\partial(x^\mu,x^\nu)}=0. \tag{23'}$$

The g, g^{-1} are not independent and we expect that $(\partial_\mu\,g^{-1})(\partial_\nu\,g)-(\partial_\nu\,g^{-1})(\partial_\mu\,g)=0$. We can prove this explicitly, using (21) for $g^{-1}\,\partial_\mu\,g$ and a similar equation for $(\partial_\mu\,g^{-1})g$. Therefore (23) gives

$$g^{-1}[\partial_\mu,\partial_\nu]g\equiv g^{-1}(\partial_\mu\,\partial_\nu-\partial_\nu\,\partial_\mu)g=i\frac{\partial(\beta,\alpha)}{\partial(x^\mu,x^\nu)}. \tag{24}$$

This result is not zero if τ and therefore β depend explicitly on x. As a result of the x dependence of the nontrivial multiplication rule we cannot interchange the order of differentiation. It is known (e.g., Ref. 14) that when the $[\partial_\mu,\partial_\nu]$ acts on path-dependent quantities, we get a nonzero result which is the curvature. Here we see that a similar thing happens for path-independent quantities, if we use a nontrivial multiplication rule. The $\beta\,\partial_\mu\,\alpha$, $\partial(\beta,\alpha)/\partial(x^\mu,x^\nu)$ can be understood as "connection" and "curvature" hidden in the nontrivial multiplication rule. We explain now in what sense the terms "connection" and "curvature" are used here. At a point x of space-time we have a group G, which has been constructed from W and U(1)'. However, G is not the trivial $W\times$ U(1)' but a "twisted" one in the sense of Eq. (7). The "connection" $\beta\,\partial_\mu\,\alpha$ is a "correction" to the $g^{-1}\,\partial_\mu\,g$ due to the fact that the multiplication rule is x dependent.

For given $\alpha(x)$, $\gamma(x)$, N and for a transformation from an extension with factor set $\sigma(w_1,w_2)=\tau(w_1)+\tau(w_2)-\tau(w_1w_2)$ into another extension with factor set $\sigma+\sigma_1=\sigma+\tau_1(w_1)+\tau_1(w_2)-\tau_1(w_1w_2)$,

$$g^{-1}\,\partial_\mu\,g\to g^{-1}\,\partial_\mu\,g+i\delta\beta\,\partial_\mu\,\alpha,$$
$$g^{-1}[\partial_\mu,\partial_\nu]g\to g^{-1}[\partial_\mu,\partial_\nu]g+i\frac{\partial(\delta\beta,\alpha)}{\partial(x^\mu,x^\nu)}, \tag{25}$$
$$\delta\beta=\partial_\alpha\,\tau_1(\alpha,N)-\partial_\alpha\,\tau_1(0,0).$$

We call this extension transformation or loop gauge transformation. The second terminology will become clear if we reexpress these results using the path-dependent

$$G(x,C)=g(x)\exp\left[-i\int_{(c)}^{x}\beta\,\partial_\mu\,\alpha\,\delta x_\mu\right]=\exp[i\alpha\hat{q}]\exp[iN\hat{\theta}]\exp i\left[\gamma-\int_{(c)}^{x}\beta\,\partial_\mu\,\alpha\,\delta x_\mu\right], \tag{26}$$

where C is a path in the four-dimensional space-time and the integration is taken along C, from a reference point 0 up to x. The $G(x,C)$ is an extension-dependent quantity because β, which is given in (22), is an extension-dependent quantity.

For given $\alpha(x)$, $\gamma(x)$, N and for a transformation from an extension σ into another $\sigma+\sigma_1$,

$$G(x,C)\to G(x,C)\exp\left[-i\int_{c}^{x}\delta\beta\,\partial_\mu\,\alpha\,\delta x_\mu\right], \tag{27}$$
$$\delta\beta=\partial_\alpha\,\tau_1(\alpha,N)-\partial_\alpha\,\tau_1(0,0).$$

We call (27) an extension transformation or loop gauge transformation. Now,

$$G^{-1}(x,C)\partial_\mu\,G(x,C)=i(\partial_\mu\,\alpha)\hat{q}+i(\partial_\mu\,\gamma), \tag{28}$$
$$G^{-1}(x,C)\,[\partial_\mu,\partial_\nu]\,G(x,C)=0, \tag{29}$$

i.e., the $G^{-1}(x,C)\partial_\mu\,G(x,C)$ is invariant under extension transformations. The path-dependent term $\exp[-i\int_{(c)}^{x}\beta\,\partial_\mu\,\alpha\,\delta x_\mu]$ has canceled the effect of the nontrivial multiplication rule.

The definition of area differentiation for path-dependent quantities is well known (e.g., Ref. 14). We apply it to Eq. (26) and we get

$$G^{-1}(x,C)\frac{\delta}{\delta\sigma_{\mu\nu}}G(x,C)=i\frac{\partial(\beta,\alpha)}{\partial(x^\mu,x^\nu)}. \tag{30}$$

This is a different way of expressing the result of Eq. (24).

We consider now general factor sets in $Z^2(W,\text{U}(1)')$. Equation (18) is still valid, but σ is now given by Eq. (15)

and therefore Eq. (19) has an extra term $-N(em)\delta\alpha$ in the right-hand side. Consequently Eq. (21) will have an extra term $-iN(em)\partial_\mu\,\alpha$ on the right-hand side. This term can be understood as a result of the noncommutativity between $(\partial_\mu\,\alpha)\hat{q}$ and $\exp(iN\hat{\theta})$ [Eq. (16)]. In the extensions with factor sets in $B^2(W,\mathrm{U}(1)')$ that we studied before, the $m=0$ and this term is absent. For given $\alpha(x)$, $\gamma(x)$, N and for a transformation from an extension in a class m_1,

$$\sigma = em_1(N_1\alpha_2 - N_2\alpha_1) + \tau(N_1,\alpha_1) + \tau(N_2,\alpha_2) - \tau(N_1+N_2,\alpha_1+\alpha_2)$$

into the corresponding extension (i.e., with the same function τ) in the class m_2

$$\sigma = em_2(N_1\alpha_2 - N_2\alpha_1) + \tau(N_1,\alpha_1) + \tau(N_2,\alpha_2) - \tau(N_1+N_2,\alpha_1+\alpha_2),$$

the $g^{-1}\,\partial_\mu\,g \to g^{-1}\,\partial_\mu\,g - iNe(m_1-m_2)\partial_\mu\,\alpha$, where $e(m_1-m_2)$ is an integer. We see here the special role the noncommutativity between $\hat{q}$ and $\hat{\theta}$ plays in our model. The value of m defines the strength of this noncommutativity [Eq. (14)] and by going from a class m_1 into a different class m_2 we have a U(1)′ gauge transformation.

III. LOOP GAUGE THEORY AND EXTENSIONS WITH FACTOR SET IN $B^2(W,\mathrm{U}(1)')$

We introduce[1] the complex wave function $\phi(x,\theta)$, where θ is a coordinate for the U(1) gauge group. We need of course[1] a length scale L for U(1) and we put $L=1$. The electric charge operator is $\hat{q} = ie\,\partial_\theta$.

Consider now the Weyl transformations W acting on the wave function $\phi(x,\theta)$ defined up to a phase factor. We are working, at the moment, with the cosets

$$\{\exp(i\gamma)\phi(x,\theta)\,|\,\text{arbitrary }\gamma\}.$$

Reference 4 explains in detail how we define continuity and derivatives in this case. We work with representatives of the cosets and at the end we multiply the result with an arbitrary phase factor $\exp(i\gamma)$.

We introduce local Weyl transformations, by taking $\alpha(x)$ to be a function of x and N an integer independent of x. The W is an Abelian group [Eq. (6)] and the $\exp(iN\hat{\theta})$ commutes with the $\exp(i\alpha\hat{q})$ in the sense that the commutator is the unit element of W, i.e., the coset $\{\exp(i\gamma)\,|\,\text{arbitrary }\gamma\}$. Therefore the $\exp(iN\hat{\theta})$ plays no role at this stage.

We define now the covariant derivative

$$D_\mu = \partial_\mu - iA_\mu\hat{q}, \tag{31}$$

which transforms like

$$\exp(-i\alpha\hat{q})D_\mu\exp(i\alpha\hat{q}) \to D_\mu, \qquad A_\mu \to A_\mu + \partial_\mu\,\alpha. \tag{32}$$

Here the $\exp(i\alpha\hat{q})$ denotes the coset of Eq. (6) with an arbitrary integer N. Covariant derivative of this type has been used before (e.g., Ref. 1). Note that the potential (connection) is an operator $A_\mu\hat{q}$. The $A_\mu(x)$ and $\alpha(x)$ depend only on x and are independent of θ. This is a known restriction in the five-dimensional theories associated with the requirement that translations in the internal dimension should not change the length of the projection of a curve in the four-dimensional space-time (Chap. 17 of Ref. 15).

We also introduce the gauge field (curvature) operator

$$[D_\mu,D_\nu] = -if_{\mu\nu}\hat{q}, \quad f_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \tag{33}$$

which obeys the Bianchi identity

$$\sum_{\text{cycl}}[D_\mu,[D_\nu,D_\rho]] = \left\{\sum_{\text{cycl}}\partial_\mu f_{\nu\rho}\right\}\hat{q} = 0. \tag{34}$$

Let $C = X^\mu(\tau)$ be a path in $\mathbb{R}^4$. We introduce the wave function (defined up to a phase factor)

$$\exp\left[-i\int_{(c)}^{x} A_\mu\,dx_\mu\hat{q}\right]\phi(x,\theta) = \phi\left(x,\theta+\int_{(c)}^{x}A_\mu\,dx_\mu\right) = \sum_N \phi_N(x)\exp\left[iN\left(\theta+e\int_{(c)}^{x}A_\mu\,dx_\mu\right)\right]. \tag{35}$$

The integration is taken along C from a reference point O up to the point x. This is the path-dependent gauge-independent Mandelstam wave function[14] for our model. The operator $\exp[-i\int_{(c)}^{x}A_\mu\,dx_\mu\,\hat{q}]$ is the loop operator for our scheme. Particularly interesting is the case of closed loops C_{xx} which have x as origin and as end point. The wave function is $\phi(x,\theta+e\int_{C_{xx}}A_\mu\,dx_\mu)$ and depends on the point x and on the loop C_{xx}.

So far we have used the group W of transformations and the wave function was defined up to a phase factor [element of U(1)′]. The connection $A_\mu\hat{q}$ defines "parallelism" between cosets. For example, the coset $\{\exp(i\gamma)\,|\,\theta\rangle_x\,|\,\text{arbitrary }\gamma\}$ at the point x, is "parallel" to the coset

$$\left\{\exp(i\gamma)\,\middle|\,\theta+\int_{x(c)}^{y}A_\mu\,\delta x_\mu\rangle_y\,\middle|\,\text{arbitrary }\gamma\right\}$$

at the point y. We will now go further and define "parallelism" between the elements of the coset at x and the elements of the coset at y. The extensions of W by U(1)′ are needed here. We will introduce a theory covariant under transformations in any of these extensions.

It is easy to introduce covariant derivative for transformations in the trivial ($\sigma=0$) extension only. In this case the extension is simply $W\times\mathrm{U}(1)'$. The $g^{-1}\,\partial_\mu\,g = i(\partial_\mu\,\alpha)\hat{q} + i\,\partial_\mu\,\gamma$ and the covariant derivative is

$$D_\mu = \partial_\mu - iA_\mu\hat{q} - i\Gamma_\mu, \qquad A_\mu \to A_\mu + \partial_\mu\,\alpha, \qquad \Gamma_\mu \to \Gamma_\mu + \partial_\mu\,\gamma. \tag{36}$$

For a given $g(x)$, consider now a transformation from the trivial extension $\sigma=0$ into an extension σ with the factor set a two-coboundary [element of $B^2(W,\mathrm{U}(1)')$]. Note that in this case $m=0$ and [Eq. (16)] $\exp(-iN\hat{\theta})\hat{q}\exp(iN\hat{\theta}) = \hat{q}$. We now use Eq. (25) to get

$$g^{-1}D_\mu\,g \to g^{-1}D_\mu\,g + i\beta\,\partial_\mu\,\alpha, \tag{37}$$

$$g^{-1}[D_\mu,D_\nu]g \to g^{-1}[D_\mu,D_\nu]g + i\frac{\partial(\beta,\alpha)}{\partial(x^\mu,x^\nu)}. \tag{38}$$

The D_μ is given in Eq. (36) and can no longer be called covariant derivative. The value of the quantity $g^{-1}D_\mu\,g$ depends on the multiplication rule, i.e., on the extension. Simi-

larly the $g^{-1}[D_\mu, D_\nu]g$ also depends on the extension. We call (37) and (38) loop gauge transformations or extension transformations.

We introduce now a "potential in loop space" $A_{\mu\nu}(x)$ which for given $g(x)$ and under transformation from the extension σ into $\sigma + \sigma_1$ transforms like

$$A_{\mu\nu} \to A_{\mu\nu} + \frac{\partial(\delta\beta,\alpha)}{\partial(x^\mu,x^\nu)}, \tag{39}$$

where $\delta\beta$ is given in Eq. (25). We now define the

$$\begin{aligned}\hat{\Omega}_{\mu\nu} &= [D_\mu,D_\nu] - iA_{\mu\nu} = [\partial_\mu,\partial_\nu] - iF_{\mu\nu}\hat{q} \\ &\quad - i[(\partial_\mu \Gamma_\nu - \partial_\nu \Gamma_\mu) + A_{\mu\nu}] \\ &= D_{\mu\nu} - iF_{\mu\nu}\hat{q} - i(\partial_\mu \Gamma_\nu - \partial_\nu \Gamma_\mu)\end{aligned} \tag{40}$$

$$D_{\mu\nu} = [\partial_\mu,\partial_\nu] - iA_{\mu\nu}. \tag{41}$$

The $g^{-1}(x)\hat{\Omega}_{\mu\nu}g(x)$ is covariant under transformations of the extension (loop gauge transformations).

The gauge field (in loop space) is[5]

$$F_{\mu\nu\lambda} = \sum_{\text{cycl}} \partial_\mu A_{\nu\lambda} \tag{42}$$

and is invariant under loop gauge transformations [Eq. (23)']. Its dual is

$$\Sigma_\mu = \epsilon_{\mu\nu\rho\sigma} F_{\nu\rho\sigma} = \epsilon_{\mu\nu\rho\sigma} \partial_\nu A_{\rho\sigma} = \partial^*_\nu A_{\mu\nu}. \tag{43}$$

Equation (42) has important geometrical content and is called Cartan structural equation in the loop space.

They obey the Bianchi identity (in loop space)

$$\epsilon_{\mu\nu\rho\sigma} \partial_\mu F_{\nu\rho\sigma} = 0 \text{ or } \partial_\mu \Sigma_\mu = 0. \tag{44}$$

Note also that

$$\begin{aligned}\epsilon_{\rho\sigma\mu\nu}(\partial_\mu \Sigma_\nu - \partial_\nu \Sigma_\mu) &= \partial^2_\lambda A_{\rho\sigma} + \partial_\lambda \partial_\rho A_{\sigma\lambda} \\ &\quad + \partial_\lambda \partial_\sigma A_{\lambda\rho}.\end{aligned} \tag{45}$$

We can rewrite (43) and (44) as

$$\partial_\mu {}^*\Omega_{\mu\nu} = \epsilon_{\mu\nu\rho\sigma} \partial_\mu \Omega_{\rho\sigma} = \Sigma_\nu, \tag{46}$$

$$\partial_\nu(\partial_\mu {}^*\Omega_{\mu\nu}) = \partial_\nu \Sigma_\nu = 0. \tag{47}$$

In the quantity $\Omega_{\mu\nu}$ we have the gauge field $F_{\mu\nu}$ and the potential in loop space $A_{\mu\nu}$. Consequently Eq. (46) is a combination of the Bianchi identity (34) and the structural Eqs. (42) and (43). Equation (47) expresses the Bianchi identity in loop space.

We can introduce the potential in loop space $A_{\mu\nu}$, through Eq. (30). In Eq. (27) we have seen how the $G(x,C)$ transforms under an extension transformation (loop gauge transformation). From Eq. (30) it is clear that a covariant area differentiation is

$$\frac{\delta}{\delta\sigma_{\mu\nu}} - iA_{\mu\nu}, \quad A_{\mu\nu} \to A_{\mu\nu} + \frac{\partial(\delta\beta,\alpha)}{\partial(x^\mu,x^\nu)}, \tag{48}$$

$\delta\beta$ is given in (27).

We now introduce a surface dependent but loop-gauge independent wave function. For a closed loop C_{xx} with origin and end point at x,

$$\begin{aligned}&\exp\left[i\int_{S_{C_{xx}}} A_{\mu\nu}\, d\sigma_{\mu\nu}\right]\phi\left(x,\theta + e\int_{C_{xx}} A_\mu\, dx_\mu\right) \\ &\quad = \exp\left[i\int_{S_{C_{xx}}} A_{\mu\nu}\, d\sigma_{\mu\nu} + i\int_{C_{xx}} A_\mu\, dx_\mu\, \hat{q}\right]\phi(x,\theta) \\ &\quad = \exp\left[\int_{S_{C_{xx}}} i(A_{\mu\nu} + f_{\mu\nu}\,\hat{q})d\sigma_{\mu\nu}\right]\phi(x,\theta),\end{aligned} \tag{49}$$

where $S_{C_{xx}}$ is a surface with boundary C_{xx}. This wave function depends on the surface $S_{C_{xx}}$ with boundary C_{xx}.

For a closed path C_{xx}, the Weyl group that we used at the beginning of this section led to the result that the coset $\{e^{i\gamma}|\theta\rangle|\text{arbitrary } \gamma\}$ is parallel to the coset $\{e^{i\gamma}|\theta + \int_{C_{xx}} A_\mu\, \delta x_\mu\rangle|\text{arbitrary } \gamma\}$. Now we have defined a mapping between the elements of these cosets and the ket $\epsilon^{i\gamma}|\theta\rangle$ is parallel to the

$$\exp\left[i\left(\gamma + \int_{S_{C_{xx}}} A_{\mu\nu}\, \delta\sigma_{\mu\nu}\right)\right]\left|\theta + \int_{C_{xx}} A_\mu\, \delta x_\mu\right\rangle.$$

For open paths, we can write a similar result using a surface between our curve and a reference curve, usually taken at infinity.

IV. THE ACTION

We study now an example of a Lagrangian in which the above ideas can be applied. We separate the action in five parts. The terms S_1 and S_5 are not invariant under loop gauge transformations and the loop current $J_{\mu\nu}$ is not conserved. The strings are open and the end points, which are described with the current $J'_\mu = \partial_\mu J_{\nu\mu}$, are magnetic monopoles. We will see that the magnetic current J'_μ corresponds to the group U(1)' and is coupled to the potential Γ_μ. The Γ_μ cannot be absorbed by the $A_{\mu\nu}$, because our action is not invariant under loop gauge transformations.

Another consequence of this noninvariance of the action is that we need to specify a particular loop gauge, for which our Lagrangian is written. If we want to change loop gauge we need to add extra terms in the Lagrangian. We fix the loop gauge by using the multiplication rule with factor set $\sigma = N_1\alpha_2 - N_2\alpha_1$ [Eq. (15) with $em = 1, \tau = 0$]. In this particular loop gauge the multiplication rule is very simple.

The first part of the action is

$$\begin{aligned}S_1 &= \frac{1}{2}\int d\theta\, d^4x |D_\mu \phi|^2, \\ D_\mu &= \partial_\mu - iA_\mu\hat{q} - i\Gamma_\mu,\end{aligned} \tag{50}$$

θ is considered as a dimensionless variable. A length L is required in the θ dimension[1] and it has been taken equal to 1.

The second part of the action describes the coupling between $A_{\mu\nu}$ and the source $J_{\mu\nu}$. Equation (41) suggests that at least one possibility is the term

$$\begin{aligned}S_2 &= \frac{k_1}{2}\int d\theta\, d^4x (D_{\mu\nu}\phi)^*(D_{\mu\nu}\phi), \\ D_{\mu\nu} &= [\partial_\mu,\partial_\nu] - iA_{\mu\nu},\end{aligned} \tag{51}$$

which in the loop gauge that we have chosen becomes

$$S_2 = \frac{k_1}{2}\int d\theta\, d^4x\, A^2{}_{\mu\nu}|\phi|^2. \tag{51'}$$

We introduce here a term with higher-order derivatives to describe the source of the $A_{\mu\nu}$. Higher-order derivatives are usually undesirable in renormalizable models. However, Lagrangians like ours are usually considered as effective Lagrangians, for a deeper model where the Higgs mechanism and Nielsen–Olesen[16] strings occur. In this context, we suggest that (51) is a possibility for a source term.

The third part is the action for the $F_{\mu\nu}$ field

$$S_3 = -\frac{1}{4}\int F^2{}_{\mu\nu}\, d^4x. \tag{52}$$

The fourth part is the action for the $F_{\mu\nu\lambda}$ and is well known to be[5]

$$S_4 = -k_2 \int d^4x \frac{1}{2}\Sigma_\mu{}^2 = \frac{k_2}{12}\int d^4x\, F^2{}_{\mu\nu\lambda}. \tag{53}$$

The constants k_1 and k_2 are necessary for dimensional reasons $[k_1,k_2 \sim (\text{mass})^{-2}]$.

The fifth part is a mass term for the $A_{\mu\nu}$ field

$$S_5 = -\frac{1}{4}\int [A_{\mu\nu} + (\partial_\mu \Gamma_\nu - \partial_\nu \Gamma_\mu)]^2\, d^4x. \tag{54}$$

The action has the general form

$$S = \int L_1(A) d^4x\, d\theta + \int L_2(A) d^4x,$$

and variation gives the equations of motion

$$\int \frac{\partial L_1}{\partial A}\, d\theta + \frac{\partial L_2}{\partial A} - \partial_\mu \int \frac{\partial L_1}{\partial(\partial_\mu A)}\, d\theta - \partial_\mu \frac{\partial L_2}{\partial(\partial_\mu A)} = 0. \tag{55}$$

Variation with respect to $A_{\mu\nu}$ gives the equation

$$\begin{aligned}\partial_\rho F_{\rho\mu\nu} &= \partial_\rho{}^2 A_{\mu\nu} + \partial_\rho \partial_\mu A_{\nu\rho} + \partial_\rho \partial_\nu A_{\rho\mu} \\ &= -\epsilon_{\mu\nu\kappa\lambda}(\partial_k \Sigma_\lambda - \partial_\lambda \Sigma_k) \\ &= (1/k_2)[-J_{\mu\nu} + A_{\mu\nu} + (\partial_\mu \Gamma_\nu - \partial_\nu \Gamma_\mu)],\end{aligned} \tag{56}$$

with

$$J_{\mu\nu} = k_1 A_{\mu\nu} \int d\theta\, |\phi(x,\theta)|^2 \tag{57}$$

in our particular loop gauge.

Variation with respect to Γ_μ gives the equation

$$\partial_\mu [A_{\mu\nu} + (\partial_\mu \Gamma_\nu - \partial_\nu \Gamma_\mu)] = J'_\nu$$

$$J'_\nu = i\int d\theta\, [\phi(D_\mu \phi)^* - \phi^*(D_\mu \phi)], \tag{58}$$

$$D_\mu = \partial_\mu - iA_\mu \hat{q} - i\Gamma_\mu.$$

The J'_ν corresponds to the U(1)′ group and it would have been zero if our Lagrangian was invariant under loop gauge transformations. However, the terms S_1 and S_5 break loop gauge invariance and the J'_ν is not zero.

We differentiate Eq. (56) and we get

$$\partial_\mu J_{\mu\nu} = \partial_\mu [A_{\mu\nu} + (\partial_\nu \Gamma_\nu - \partial_\nu \Gamma_\mu)]. \tag{59}$$

From Eqs. (58) and (59) we see that we need to prove

$$\partial_\mu J_{\mu\nu} = J'_\nu. \tag{60}$$

We have already explained the physical meaning of this equation. The loop current is not conserved and the strings are open with magnetic monopoles at the end points. The J'_ν is the magnetic current.

In order to prove this equation we consider a loop gauge transformation from our particular loop gauge into another one infinitesimally near and we calculate δL,

$$\phi \to \exp\left[i\int_{(c)}^{x} \delta\Lambda_\mu\, \delta x_\mu\right]\phi,$$

$$\partial_\mu \phi \to \exp\left[i\int_{(c)}^{x} \delta\Lambda_\mu\, \delta x_\mu\right](\partial_\mu + i\,\delta\Lambda_\mu)\phi,$$

$$[\partial_\mu,\partial_\nu]\phi \to \exp\left[i\int_{(c)}^{x} \delta\Lambda_\mu\, \delta x_\mu\right]\{[\partial_\mu,\partial_\nu] + i(\partial_\mu\, \delta\Lambda_\nu - \partial_\nu\, \delta\Lambda_\mu)\}\phi,$$

and

$$\begin{aligned}\delta L &= J'_\mu\, \delta\Lambda_\mu + J_{\mu\nu}(\partial_\nu\, \delta\Lambda_\mu - \partial_\mu\, \delta\Lambda_\nu) \\ &= (J'_\nu - \partial_\mu J_{\mu\nu})\delta\Lambda_\nu - \partial_\nu(J_{\mu\nu}\, \delta\Lambda_\mu).\end{aligned}$$

We get Eq. (60) and $\delta L = -\partial_\nu(J_{\mu\nu}\, \delta\Lambda_\mu)$. Of course, in our model $\delta L = -\partial_\nu(J_{\mu\nu}\, \delta\Lambda_\mu) \neq 0$.

We multiply Eq. (56) by $\epsilon_{\mu\nu\rho\sigma}$ and differentiate to get [use also Eqs. (43) and (44)]

$$(\partial^2 + 1/k_2)\Sigma_\nu = \partial^*_\mu J_{\mu\nu}. \tag{61}$$

These equations are known[5] and the only modification in our model is that the J'_μ and $J_{\mu\nu}$ are given by Eqs. (57) and (58) and not by Eqs. (3) and (4).

Variation with respect to A_μ gives

$$\partial_\mu F_{\mu\nu} = J_\nu, \tag{62}$$

with

$$J_\nu = i\int d\theta\, [(\hat{q}\phi)(D_\mu \phi)^* - (\hat{q}\phi)^*(D_\mu \phi)],$$

$$D_\mu = \partial_\mu - iA_\mu \hat{q} - i\Gamma_\mu, \tag{63}$$

$$\partial_\nu J_\nu = 0. \tag{64}$$

The J_μ is the electric current which can also be considered as the (μ,θ) component of the energy-momentum tensor in the five-dimensional (x^μ,θ) space.

We should point out that a real field $\phi(x,\theta)$ is sufficient to describe the electric charges

$$\phi(x,\theta) = \sum_N \phi_N(x)\exp(iN\theta) + \phi^*_N(x)\exp(-iN\theta).$$

The $\phi_1(x), \phi_1{}^*(x)$ describe the $\pm 1e$, $\phi_2(x), \phi_2{}^*(x)$, $\pm 2e$, etc. The complex field $\phi(x,\theta)$ has double degrees of freedom and describes both electric and magnetic charges.

Variation with respect to ϕ, ϕ^* in the five-dimensional (x^μ,θ) space and in the particular loop gauge $\tau = 0$ gives

$$D_\mu{}^2\phi + A^2{}_{\mu\nu}\,\phi = 0, \tag{65}$$

$$D_\mu{}^2\phi^* + A^2{}_{\mu\nu}\,\phi^* = 0. \tag{66}$$

The $A^2{}_{\mu\nu}\,\phi$ describes the effect of the "Bohm–Aharanov medium" on the field ϕ (in this particular loop gauge).

Following Ref. 1 we can include in the Lagrangian a mass term

$$S_6 = \int |\partial_\theta \phi|^2 \, d^4x \, d\theta$$

$$= \sum_N N^2 \int (|\phi_N|^2 + |\phi_{-N}|^2) d^4x. \tag{67}$$

This term gives higher masses [$\sim O(N^2)$] to higher charges and therefore offers an explanation for their nonobservability in the experiments. This term alters Eqs. (65) and (66) to

$$\begin{aligned} &D_\mu{}^2\phi + A^2{}_{\mu\nu}\,\phi + \partial_\theta{}^2\phi = 0, \\ &D_\mu{}^2\phi^* + A^2{}_{\mu\nu}\,\phi^* + \partial_\theta{}^2\phi^* = 0. \end{aligned} \tag{68}$$

V. CONCLUSIONS

Electric current is associated with a U(1) symmetry and magnetic current is also associated with a U(1)′ symmetry. It is a nontrivial problem, to combine the two U(1) groups in a theory with electric and magnetic charges. In standard electrodynamics without magnetic charges, parallel transport along a closed curve C changes the U(1) phase by $e\int_c A_\mu\,\delta x_\mu = e\int_E f_{\mu\nu}\,\delta\sigma_{\mu\nu}$; this is magnetic flux through the loop C produced by electric charges and described by the potential A_μ. Magnetic strings and magnetic monopoles are in some sense "extra objects" which have to be introduced in a way consistent with the above picture. The Dirac–Wu–Yang approach exploits the fact that $\mathrm{U}(1) = R\,|Z$; it introduces tubes of magnetic flux $2\pi N/e$, which change the phase by $2\pi N$ and leave unchanged the quantity $\exp\{ie\int A_\mu\,\delta x_\mu + i2\pi N\}$.

In this paper we treat the $e^{i\theta} \in \mathrm{U}(1)$ phase quantum mechanically. Parallel transport along a curve C transforms the coset $I_x = \{e^{i\gamma}|\theta\rangle\,|e^{i\gamma} \in \mathrm{U}(1)'\}$ into the coset $I_y = \{e^{i\gamma}|\theta + \int_x^y A_\mu \delta x_\mu\rangle\,|e^{i\gamma} \in \mathrm{U}(1)'\}$. We then ask the question, how should we combine the U(1)′ with the Weyl group in order to interpret the U(1)′ as a group of magnetic charges? We explore various ways as extensions of the Weyl group by U(1)′ and we use them to study the most general mapping between the elements of I_x and I_y. We find that the $g^{-1}\,\partial_\mu\, g$ is an extension-dependent quantity and that a change of the extension leads to a loop gauge transformation and consequently to the potentials and currents that describe magnetic strings and charges (two-form potential, two-form current and three-form gauge field). So the answer to the above question is that we should require covariance under a change of the extension.

The currents $J_{\mu\nu}$ and J'_ν are not the semiclassical currents of Eqs. (3) and (4) but are given in (57) and (58) in terms of the wave function $\phi(x,\theta)$. An important point, which we have not discussed, is the $e \to 0$ semiclassical (for the θ dimension) limit. The operators $\hat{q}$ and $\hat{\theta}$ become c numbers and the U(1)′, which played an important role in our arguments, shrinks into a point. In this limit the currents (57) and (58) should reduce to (3) and (4).

Finally we should mention the work on quantum mechanics in nontrivial topology,[17] which has been inspired by the Bohm–Aharanov experiment and which is a beautiful prototype for the ideas involved in the topological objects. In the Bohm–Aharanov experiment we have a solenoid (singularity) that is a macroscopic classical object (which follows only one world surface in space-time) and that creates a multiply connected space. We should generalize these ideas and study the "Bohm–Aharanov medium" where the tube of magnetic flux follows all the surfaces in space-time. In this paper we have presented a geometrical model for the Bohm–Aharanov medium.

ACKNOWLEDGMENTS

I am grateful to Dr. J. S. Dowker for many helpful discussions and a critical reading of the manuscript. I am also grateful to Dr. R. Bryant for mathematical advice on group extension.

[1] E. Cremmer and J. Scherk, Nucl. Phys. B **103**, 399 (1976); B **118**, 61 (1977); A. Salam and J. Strathdee, Ann. Phys. (NY) **141**, 316 (1982); E. Cremmer, in *Supergravity 81*, edited S. Ferrara and J. G. Taylor (Cambridge U. P., Cambridge, 1982).

[2] H. Weyl, *Theory of Groups and Quantum Mechanics* (Dover, New York, 1950); J. Schwinger, *Quantum Kinematics and Dynamics* (Benjamin, New York, 1970); A. M. Perelomov, Sov. Phys. Usp. **20**, 703 (1977); J. R. Klauder, in *Path Integrals*, edited by G. J. Papadopoulos and J. T. Devreese (Plenum, New York, 1978).

[3] S. MacLane, *Homology* (Springer, Berlin, 1963); A. A. Kirillov, *Elements of the Theory of Representation* (Springer, Berlin, 1976); M. Hall, *Theory of Groups* (MacMillan, London, 1959).

[4] E. Wigner, Ann. Math. **40**, 149 (1939); V. Bargmann, Ann. Math. **59**, 1 (1954); G. W. Mackey, *Induced Representations of Groups and Quantum Mechanics* (Benjamin, New York, 1968); A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).

[5] M. Kalb and P. Ramond, Phys. Rev. D **9**, 2273 (1974); E. Cremmer and J. Scherk, Nucl. Phys. B **72**, 117 (1974); Y. Nambu, Phys. Rev. D **10**, 4262 (1974); Phys. Rep. C **23**, 250 (1976); in *Quark Confinement and Field Theory*, edited by D. R. Stump and D. H. Weingarten (Wiley, New York, 1977); F. Lund and T. Regge, Phys. Rev. D **14**, 1524 (1976); A. Aurilia and F. Legovini, Phys. Lett. B **67**, 299 (1977); A. Aurilia and D. Christodoulou, *ibid.* B **71**, 90 (1977); A. Aurilia, D. Christodoulou, and F. Legovini, *ibid.* B **73**, 429 (1978); A. Aurilia and D. Christodoulou, *ibid.* B **78**, 589 (1978).

[6] J. Scherk and J. H. Schwarz, Nucl. Phys. B **81**, 118 (1974).

[7] P. A. M. Dirac, Proc. R. Soc. London Ser. A **133**, 60 (1931); Phys. Rev. **74**, 817 (1948); T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975); P. Goddard and D. Olive, Rep. Prog. Phys. **41**, 1357 (1978); J. Schwinger, Phys. Rev. **144**, 1087 (1966); **173**, 1536 (1968); D **12**, 3105 (1975); in *Particles, Sources and Fields* (Addison–Wesley, Reading, MA, 1970), Vol. 1; D. Zwazinger, Phys. Rev. D **3**, 880 (1971); **176**, 1489 (1968).

[8] A. M. Polyakov, Phys. Lett. B **103**, 207 (1981).

[9] M. Kato and K. Ogawa, Nucl. Phys. B **212**, 443 (1983); W. Siegel, Phys. Lett. B **151**, 391, 396 (1985); D. J. Gross, J. A. Harvey, E. Martinec, and R. Rohm, Nucl. Phys. B **256**, 253 (1985); T. Banks and M. E. Peskin, Nucl. Phys. B **264**, 513 (1986); A. Neveu, H. Nicolai and P. West, Nucl. Phys. B **264**, 573 (1986); M. Kaku, CCNY-HEP-11 preprint; K. Bardacki, UCB-PTH-85 33 preprint.

[10] E. Witten, Nucl. Phys. B **268**, 253 (1986).

[11] G. t'Hooft, Commun. Math. Phys. **81**, 267 (1981); Acta Phys. Austriaca Suppl. XXII, 53 (1980); in *Proceedings 1980 Scottish University Summer School*, edited by K. Bowler and D. Sutherland (SUSSP, Edinburgh, 1981).

[12] J. L. Gervais and A. Neveu, Phys. Lett. B **80**, 255 (1979); Nucl. Phys. B **153**, 445 (1979); Y. Nambu, Phys. Lett. B **80**, 372 (1979); E. Corrigan and B. Hasslacher, Phys. Lett. B **81**, 781 (1979).

[13] R. Jackiw, Phys. Rev. Lett. **54**, 159 (1985).

[14] S. Mandelstam, Ann. Phys. (NY) **19**, 1 (1962); Y. Makeenko and A. A. Migdal, Sov. J. Nucl. Phys. **32**, 431 (1980); **33**, 882 (1981).

[15] P. G. Bergmann, *Introduction to the Theory of Relativity* (Prentice–Hall, Englewood Cliffs, NJ, 1960).

[16] H. B. Nielsen and P. Olesen, Nucl. Phys. B **61**, 45 (1973).

[17] L. S. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981), (§ 23); J. S. Dowker, J. Phys. A **5**, 936 (1972); K. D. Rothe and J. A. Swieca, Nucl. Phys. B **138**, 26 (1978); B **149**, 237 (1979).

Fractal and nonfractal behavior in Levy flights

Zheming Cheng and Robert Savit
Physics Department, The University of Michigan, Ann Arbor, Michigan 48109

(Received 22 January 1986; accepted for publication 22 October 1986)

The d-dimentional space-continuous time-discrete Markovian random walk with a distribution of step lengths, which behaves like $x^{-(\alpha+d)}$ with $\alpha>0$ for large x, is studied. By studying the density–density correlation function of these walks, it is determined under what conditions the walks are fractal and when they are nonfractal. An ensemble average of walks is considered and the lower entropy dimension D of the set of stopovers of the walks in this ensemble is calculated, and $D=\min\{2,\alpha,d\}$ is found. It is also found that the fractal nature of the walks is related to a finite value of the mean first passage time. The crossover of the correlation function from the fractal to nonfractal regimes is studied in detail. Finally, it is conjectured that these results for the lower entropy dimension apply to a wide class of symmetric Markov processes.

I. INTRODUCTION

The morphology of random fractals has recently become of considerable interest. One of the primary motivations for this interest has been the central role that these morphologies appear to play in a variety of kinetic growth processes. Among major questions to be understood in these processes are the questions of what conditions are necessary and sufficient for fractal growth to occur, and how the crossover to nonfractal growth regimes takes place. Unfortunately, even relatively simple, moderately realistic growth models are sufficiently complicated to render analytic progress toward understanding these questions difficult. Under these circumstances, it is therefore useful to study a much simpler process which exhibits both fractal and nonfractal growth and in which one can make analytic progress both in characterizing the nature of the fractal object generated in the fractal regime, and in studying the crossover between the fractal and nonfractal regions. To this end, we will study the process of Levy flights, which, in a certain sense, exhibit crossover from fractal to nonfractal growth as the step-length exponent of the walk is varied. Although the Hausdorff dimension of the stopovers of a Levy flight is always zero, the lower entropy dimension[1] (LED) for the process is nontrivial and corresponds to our intuitive motion of a "mass dimension." This dimension, defined for an ensemble average of walks (see below) will be used to distinguish between fractal and nonfractal regimes of the walk. Aside from their utility as analog growth processes, Levy flights are also of interest in their own right. Some work on the subject has been done by Mandelbrot,[2] and on the related subject of Weierstrassian random walks by Hughes, Montroll, and Shlesinger and Montroll and Shlesinger.[3] Furthermore, after the work reported in the present paper was completed, we became aware of the work of Hioe[4] in which a number of our results are obtained in the context of a lattice version of Levy flights.

The structure of the rest of this paper is as follows: First, we shall introduce some preliminary notions including a definition of the LED. Then we shall relate this dimension to the density–density corelation function, after which we shall calculate the asymptotic behavior of the density–density correlation function for the processes of interest. We shall end up with an expression for the LED of the stopovers of the Levy flight defined over a certain ensemble, as well as obtaining a relationship between the fractal nature of the Levy flight and the mean first passage time. We will also be able to study in detail the crossover between the fractal and nonfractal regions of the walk as we vary the step-length exponent. We will conclude with several comments and speculations.

The process we will study, a discrete-time continuous-space Levy flight, is a Markovian random walk process controlled by the probability function $P(n+1,\mathbf{x}|n,\mathbf{y})d\mathbf{x}\,d\mathbf{y}$ which is the conditional probability for the walker to be in the region $\mathbf{x}+d\mathbf{x}$ at time step $n+1$, if he was in the region $\mathbf{y}+d\mathbf{y}$ at time n. Here $\mathbf{x}$ and $\mathbf{y}$ are points in a continuous d-dimensional space, $d\mathbf{x}\equiv d^d\mathbf{x}$, $d\mathbf{y}\equiv d^d\mathbf{y}$, and n is an integer. We restrict ourselves to $P(n+1,\mathbf{x}|n,\mathbf{y})=f(\mathbf{x}-\mathbf{y})$, and we will be particularly concerned with cases in which $f(\mathbf{x}-\mathbf{y})\sim|\mathbf{x}-\mathbf{y}|^{-(\alpha+d)}$ for large $|\mathbf{x}-\mathbf{y}|$. The Levy flight is thus a random walk with a variable step length whose size distribution is determined by $f(\mathbf{x}-\mathbf{y})$. To interpret the Levy flight as a "growth process," we imagine placing a particle at the end point of every step. Among the quantities we will discuss is the lower entropy dimension (LED), D, of the collection of these end points or stopovers defined by averaging over a suitable ensemble of walks. This D is a measure of how $N(L)$, the average number of particles contained in a nonempty region of linear dimension L, scales with L: i.e., $N(L)\sim L^{D(L)}$ and is thus consistent, for this process, with our intuitive notion of a mass dimension. If $D(L)$ is independent of L over some range then the system has a well-defined LED over that range.

Before proceeding with the calculation properly, it is useful to carefully define the quantities in which we shall be interested and to clearly state how averages are to be understood. Consider then the Levy flight defined by

$$P_n(\mathbf{x})=\int d\mathbf{y}\, f(\mathbf{x}-\mathbf{y})P_{n-1}(\mathbf{y}), \tag{1}$$

where $P_n(\mathbf{x})$ is the probability density for the nth step to land on point x. We start our process at time $n=0$ at point $\mathbf{x}=0$, so that in terms of the conditional probability defined above,

 0022-2488/87/030592-06$02.50

$$P_n(\mathbf{x}) \equiv P(n,\mathbf{x}|0,0). \tag{2}$$

Now, suppose we have generated a single sample of a Levy flight with a total of m steps. Let $\rho_m(\mathbf{x})d\mathbf{x}$ be the number of stopovers contained in the region $d\mathbf{x}$ about the point $\mathbf{x}$. The density–density correlation function is then

$$C'_m(\mathbf{r};\mathbf{x}) = \rho_m(\mathbf{x}+\mathbf{r})\rho_m(\mathbf{x}). \tag{3}$$

This quantity can be integrated over $\mathbf{r}$ to obtain

$$N'(L;\mathbf{x}) = \int_0^L d^d\mathbf{r}\, C'_m(\mathbf{r};\mathbf{x}), \tag{4}$$

which is the number of points contained in the region of linear dimension L weighted by $\rho_m(\mathbf{x})$, the number of particles at $\mathbf{x}$. Finally, we may average this quantity over a number of such m-step Levy flights and over all starting points $\mathbf{x}$ to obtain

$$\begin{aligned} N(L) &\equiv \langle N'(L;\mathbf{x})\rangle \\ &= \left\langle \int_0^L d^d\mathbf{r}\, C'_m(\mathbf{r};\mathbf{x}) \right\rangle \\ &= \left\langle \int_0^L d^d\mathbf{r}\, \rho_m(\mathbf{x}+\mathbf{r})\rho_m(\mathbf{x}) \right\rangle \\ &= \int_0^L d^d\mathbf{r} \langle \rho_m(\mathbf{x}+\mathbf{r})\rho_m(\mathbf{x})\rangle \\ &= \int_0^L d^d\mathbf{r} \langle C'_m(\mathbf{r};\mathbf{x})\rangle = \int_0^L d^d\mathbf{r}\, C_m(\mathbf{r}), \end{aligned} \tag{5}$$

where $\langle\ \rangle$ means averaging over the ensemble of samples. An explicit procedure for performing this average will be explained below. As we shall see, as a result of our averaging procedure, $N(L)$ and $C_m(\mathbf{r})$ will be independent of x. In any case, the x dependence for large m would be trivial since the process is translationally invariant. Therefore, $N(L)$, the average number of particles contained in a region of linear dimension L, having a behavior like $N(L) \sim L^D$ is equivalent to $C_m(\mathbf{r})$, the average density–density correlation function behaving like $C_m(\mathbf{r}) \sim r^{D-d}$.

II. THE AVERAGE DENSITY–DENSITY CORRELATION FUNCTION

We now want to calculate the average density–density correlation function for the processes in which we are interested. The result of this calcualtion will be an expression for the LED of the Levy flight averaged over a suitable ensemble. We will also be able to relate the fractal nature of the Levy flight to its mean first passage time, and we will be able to study in some detail the crossover from a fractal to nonfractal structure for the walk as we vary the step-length exponent. Unless explicitly stated otherwise in the sequel, when we refer to properties of the Levy flight, it should be understood that these statements refer to quantities averaged over the ensemble of sample flights, the construction of which we now explain.

To do this, we begin by defining a modified correlation function,

$$C_m(\mathbf{r}|j,\mathbf{x}) = \langle \rho_m(\mathbf{x}+\mathbf{r})\rho_m(\mathbf{x})\rangle_{(j,\mathbf{x})},$$

where $\langle\ \rangle_{(j,\mathbf{x})}$ means averaging over those systems in the ensemble in which the jth particle (i.e., the jth vertex of the given path) is between $\mathbf{x}$ and $\mathbf{x}+d\mathbf{x}$. Then

$$C_m(\mathbf{r}|j,\mathbf{x}) = \sum_{l=1}^{m}{}' P(l,\mathbf{r}+\mathbf{x}|j,\mathbf{x}), \tag{6}$$

where the prime on the sum means $l \neq j$. This is just the average particle density at the point $\mathbf{r}+\mathbf{x}$ if the jth particle is at the point $\mathbf{x}$. Averaging over $\mathbf{x}$, we have the correlation function averaged over an ensemble of samples in which the position of the jth particle is taken as one end point of the correlation function: i.e.,

$$C_m(\mathbf{r}|j) = \int d^d\mathbf{x}\, P_j(\mathbf{x}) C_m(\mathbf{r}|j,\mathbf{x}). \tag{7}$$

Using Eq. (1) it is clear that $P(l,\mathbf{x}|m,\mathbf{y}) = P_{l-m}(\mathbf{x}-\mathbf{y})$ for $l \geqslant m$, so that

$$C_m(\mathbf{r}|j) = \sum_{l=1}^{j-1} P_l(\mathbf{r}) + \sum_{l=1}^{m-j} P_l(\mathbf{r}). \tag{8}$$

Finally, if we randomly choose one particle in the object as the origin for calculating the correlation function, it is equally likely to be any of the particles, so that

$$C_m(\mathbf{r}) = \frac{1}{m}\sum_{j=0}^{m} C_m(\mathbf{r}|j) = 2\sum_{l=1}^{m}\left(1 - \frac{l-\frac{1}{2}}{m}\right) P_l(\mathbf{r}). \tag{9}$$

We now want to take $m \to \infty$ in this expression. First we show that $C_m(\mathbf{r})$ and $\Sigma_{l=1}^m P_l(\mathbf{r})$ diverge and converge together as $m \to \infty$. To see this, note that if $C_m(\mathbf{r})$ diverges as $m \to \infty$, then $\Sigma_{l=1}^m P_l(\mathbf{r})$ also diverges since, recalling that $P_l(\mathbf{r}) \geqslant 0$, it follows from Eq. (9) that $\Sigma_{l=1}^m P_l(\mathbf{r}) \geqslant \frac{1}{2} C_m(\mathbf{r})$. Furthermore, we can prove that if $\Sigma_{l=1}^m P_l(\mathbf{r})$ diverges as $m \to \infty$, then so does $C_m(\mathbf{r})$ as follows: If $\Sigma_{l=1}^m P_l(\mathbf{r}) \to \infty$, as $m \to \infty$, then for a given $\mathbf{r}$ there exists, for any L, an M such that $\Sigma_{l=1}^m P_l(\mathbf{r}) \geqslant L$. This means that for $m > 2M$,

$$\begin{aligned} C_m(\mathbf{r}) &> 2\sum_{l=1}^{M}\left(1 - \frac{l}{m}\right) P_l(\mathbf{r}) \\ &> 2\sum_{l=1}^{M}\left(1 - \frac{M}{m}\right) P_l(\mathbf{r}) > 2\sum_{l=1}^{M}\left(1 - \frac{M}{2M}\right) P_l(\mathbf{r}) \\ &> 2\frac{1}{2}L = L. \end{aligned}$$

Therefore, for large enough m, $C_m(\mathbf{r})$ is larger than any preassigned number L, and so diverges as $m \to \infty$.

Finally we note that if $C_m(\mathbf{r})$ converges we have

$$C(\mathbf{r}) = \lim_{m=0} C_m(\mathbf{r}) = 2\sum_{l=1}^{\infty} P_l(\mathbf{r}). \tag{10}$$

The right-hand side of Eq. (10) is twice the mean first passage time for this random walk.

Now we use a Fourier transform to rewrite Eq. (10) as

$$C(\mathbf{r}) = 2\frac{1}{(2\pi)^{d/2}}\int d\mathbf{k} \frac{\tilde{f}(\mathbf{k})}{1-\tilde{f}(\mathbf{k})} e^{-i\mathbf{k}\cdot\mathbf{r}}, \tag{11}$$

where

$$\tilde{f}(\mathbf{k}) = \frac{1}{(2\pi)^{d/2}}\int d\mathbf{r}\, f(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}}$$

is the d-dimensional Fourier transform of $f(\mathbf{r})$. We have used $\tilde{P}_l(\mathbf{k}) = \tilde{f}^l(\mathbf{k})$. If we consider only those processes which are independent of the angular variables, Eq. (11) is reduced to a form of Hankel transform,

$$C(r) = r^{-(d-1)/2}\int_0^\infty dk \frac{\tilde{f}(k)}{1-\tilde{f}(k)} \times k^{(d-1)/2}(kr)^{1/2}J_{(d-2)/2}(kr), \qquad (11')$$

where $r = |\mathbf{r}|$, $k = |\mathbf{k}|$.

Let us now compute $C(r)$ and the LED for Levy flights. We consider walks for which the kernel in Eq. (1) has the form

$$f(r) \sim r^{-d}\sum_{i=0}^{n} b_i r^{-\alpha_i}, \quad \alpha_n > \alpha_{n-1} > \cdots > \alpha_0 \equiv \alpha > 0, \quad b_0 \neq 0, \qquad (12)$$

for large r and some integer $n > 0$ ($\alpha = \infty$ is included as a special case).

It is easy to show that (see Appendix A)

$$\tilde{f}(k) = 1 - \beta k^A + o(k^A) \quad \text{as } k \to 0, \qquad (13)$$

where $A = \min\{2,\alpha\}$. Notice that $\tilde{f}(0) = 1$, otherwise the $P_n(\mathbf{x})$ cannot be interpreted as probabilities.

Using (13) in (11) it is not difficult to determine the necessary and sufficient conditions for the convergence of $C(r)$. We find that $C(r)$ converges (a) for $d \geqslant 3$ and any $\alpha > 0$, (b) for $d = 2$ and $\alpha < 2$, and (c) for $d = 1$ and $\alpha < 1$. Using (13) in (11), we see that for these values of d and α, $C(r) \sim r^{-(d-A)}$ as $r \to \infty$, and since $C(r) \sim r^{D-d}$, $D = A$ for these values of d and α. By Eq. (10), the mean first passage time is also finite for these vlaues of d and α.

For values of d and α for which $C(r)$ is divergent, we need to study $C_m(r)$ in the $m \to \infty$ limit a little more carefully. This is done in some detail in Appendix B. Here we report the results of this calculation. We find that for (d,α) such that $C(r)$ diverges, $\lim_{m\to\infty} C_m(0) \to \infty$, but $\lim_{m\to\infty}[C_m(0) - C_m(r)]$ is a finite function of r. Therefore, it is also possible to extract for this case a value of the LED by rescaling the correlation function by its value at the origin. Defining $C_m(r) = C_m(r)/C_m(0)$, we find $\lim_{m\to\infty} C_m(r) = 1$, and so the LED in this case is $D = d$. This is the case in which the LED of the trail of points left by a typical sample of the Levy flight passages has the naive dimension of space, and is, by Eq. (10), also the case in which the mean first passage time diverges. The value of the LED for all of these cases, for both divergent and convergent values of $C(r)$ can be summarized by the formula $D = \min\{2,\alpha,d\}$. Notice that we can mimic those cases in which $f(r)$ falls faster than a power as $r \to \infty$ by setting $\alpha = \infty$. We then find the usual Gaussian result for short range random walks, namely $D = 2$ for $d \geqslant 2$, and $D = 1$ for $d = 1$.

III. THE CROSSOVER REGIME BETWEEN FRACTAL AND NONFRACTAL

The structure of a typical sample of the Levy flight process, as we can infer from the results of an ensemble average, are markedly different in the fractal and nonfractal regimes. Since, to our knowledge, this is one of the only analytically tractable systems to exhibit this crossover, it is of considerable value to explicitly display the behavior of the correlation function in the crossover regime. This is done in Appendix C. Here we wish to point out some features of this crossover and comment on the qualitative differences in the behavior of a typical Levy flight in the fractal and nonfractal regimes. First, we want to make it clear that there are really three qualitatively different types of behavior possible for the Levy flight: (i) For $D < d \leqslant 2$ and for $D < 2$ and $d \geqslant 3$ the Levy flight is fractal-like and self-similar and the mean first passage time is finite. (ii) For $D = d \leqslant 2$ the Levy flight is nonfractal and space filling and the mean first passage time is infinite. (iii) For $D = 2$ and $d \geqslant 3$ the Levy flight is not space filling, but neither is it fractal. (This case also corresponds to the usual short-range finite step length random walk above two dimensions.) Because the walk is not space filling the mean first passage time is finite in this case, also.

The dynamics for case (i) differs markedly from the dynamics for cases (ii) and (iii). In cases (ii) and (iii) in which the step length distribution, $f(r)$, falls relatively rapidly, there will be no very large jumps and the stopovers will tend to congregate near the origin of the walk with the distribution of steps forming a Gaussian-like distribution which grows smoothly in width (and for $d \leqslant 2$, in height) at time goes by. For $d = 1$ and 2 the phase space is restricted enough so that these dynamics will cause $C_m(0)$ to diverge as $m \to \infty$ causing the mean first passage time to be infinite. For $d \geqslant 3$ there are enough random walk paths to prevent $C_m(0)$ from diverging as $m \to \infty$, and so the mean first passage time is finite. If, on the other hand, $f(r)$ does not fall rapidly enough, as is the situation in case (i), the dynamics is very different. In this case very large jumps will be possible, and the whole space will be sampled, although not densely. Indeed, in computer simulations of fractal Levy flights it is observed that the fractal structure is generated by the walker spending some time in a given region of space, then taking a single very large step to a far distant region, spending some time there, and repeating the process in a scale invariant way. This dynamics differs markedly from the smoothly spreading Gaussian distribution of cases (ii) and (iii). In terms of the density–density correlation function, we show in Appendix B that for $d = 1,2$, if we set $\alpha = d - \epsilon$, then for small positive ϵ, $C(r) \sim (1/\epsilon)r^{-\epsilon}$. Thus $C(r) \to \infty$ as $\epsilon \to 0^+$ and $[C(r) - C(0)] \sim \ln r$ for large r and $\epsilon = 0$, a behavior reminiscent of simple crossover effects in critical phenomena. This paradigm is worth keeping in mind as one studies more realistic and complex growth processes with fractal–nonfractal crossover.

IV. SUMMARY

In this paper we have analyzed the structure of Levy flights in the continuum. Using the lower entropy dimension as a criterion, we have found that the set of stopover points can exhibit both fractal and nonfractal behavior depending on the value of d, the number of dimensions in which the walk is embedded, and α, the power with which the jump distribution falls off asymptotically. We were also to exhibit in detail the behavior of an ensemble average Levy flights at the fractal–nonfractal crossover point. We showed furthermore that if the mean first passage time diverges, the LED is equal to d, and the typical Levy flight (understood as a representative of our ensemble) is not fractal-like. If the mean

first passage time is finite, then the typical Levy flight will not be space filling and will generally be fractal unless $d \geqslant 3$ and $\alpha \geqslant 2$, in which case the dimension of the walk will be $D = 2$, just as for the ordinary random walk with fixed, finite step length.

We have analyzed the Levy flight for the specific step size distribution of Eq. (10). However, a careful examination of the derivation of our results clearly suggests an interesting generalization. We believe that the expression for the lower entropy dimension of the stopovers of this random walk, $D = \min\{2,\alpha,d\}$, will be correct for any symmetric distribution $f(r)$ where α is defined by

$$\alpha = \sup\left\{\alpha' \left| \int |\mathbf{x}|^{\alpha} f(|\mathbf{x}|) d^d\mathbf{x} < \infty \right.\right\}.$$

The random Levy flight we have studied has a very rich structure, but, using the techniques of this paper, is amenable to considerable analysis. Such models should prove to be simple but useful archetypes in the study of fractal kinetic growth processes.

ACKNOWLEDGMENTS

We are grateful to Z. Schuss for helpful discussions and comments and to B. Mandelbrot for a stimulating correspondence.

This work was supported by the Department of Energy under Grant No. DE-FG02-85ER45189. One of us (R. S.) also gratefully acknowledges the partial support of an Alfred P. Sloan Foundation Research Fellowship during the early stages of this work.

APPENDIX A: LEADING BEHAVIOR OF $\tilde{f}(k)$ FOR SMALL k

In this Appendix we show that for

$$f(r) \sim r^{-d} \sum_{i=0}^{n} b_n r^{-\alpha_i}; \quad \alpha_n > \alpha_{n-1} > \cdots > \alpha_0 \equiv \alpha > 0,$$

$b_0 \neq 0$ for large r, we have

$$\tilde{f}(k) = 1 - \beta k^A + o(k^A), \quad \text{as } k \to 0,$$

where $A = \min\{2,\alpha\}$.

Sketch of proof: Without loss of generality, let us consider the case $f(r) = r^{-(d+\alpha)} \sum_{i=0}^{\infty} c_i r^{-i}$ for $\alpha > 0$, r large. By using the integral representation of $J_\nu(x)$ for $d \geqslant 2$ we have

$$\tilde{f}(k) = c \int_0^\infty dr\, f(r) r^{d-1} \int_0^1 dy (1-y^2)^{(d-3)/2} \cos kyr, \tag{A1}$$

where c is a normalization constant. Equation (A1) can be rewritten as

$$\tilde{f}(k) = 1 - c \int_0^1 dy (1-y^2)^{(d-3)/2} \times \left\{ \int_0^\infty dr\, f(r) r^{d-1} (1 - \cos kyr) \right\}. \tag{A2}$$

Let us first concentrate on the integral,

$$\int_0^\infty dr\, f(r) r^{d-1} (1 - \cos kyr)$$

in (A2). We divide the integral into two parts by some large number R above which the expansion of $f(r)$ around $r = \infty$ is valid, then expand the integrands properly, we have

$$\int_0^\infty dr\, f(r) r^{d-1} (1 - \cos kyr) = \sum_{i=1}^{\infty} (ky)^{2i} \left[\int_0^R dr \frac{(-1)^i f(r)}{(2i)!} r^{2i+d-1} \right] + \sum_{i=0}^{\infty} c_i \int_R^\infty dr\, r^{-(1+\alpha+i)} (1 - \cos rky).$$

Define

$$e_i(R) = \int_0^R dr \frac{(-1)^i f(r)}{(2i)!} r^{2i+d-1}, \tag{A3}$$

then it is easy to see $e_i(R)$'s are finite for any R for $\infty > R > 0$. Next we divide the integral in the second summation into two parts by $(1/ky)(> R)$, then expand $(1 - \cos rky)$ in the first part and rescale the integral variable in the second part, and then we have

$$\begin{aligned}
&\int_R^\infty dr\, r^{-(1+\alpha+i)} (1 - \cos rky) \\
&= -\int_{1/ky}^R dr\, r^{-(1+\alpha+i)} (1 - \cos rky) + \int_{1/ky}^\infty dr\, r^{-(1+\alpha+i)} (1 - \cos rky) \\
&= -\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{(2j)!} (ky)^{2j} \int_{1/ky}^R dr\, r^{-(1+\alpha+i)+2j} + (ky)^{-(\alpha+i)} \int_1^\infty dr\, r^{-(1+\alpha+i)} (1 - \cos r) \\
&= -\sum_{j=1}^{\infty} (ky)^{2j} \frac{(-1)^{j+1}}{(2j)!} \left[\frac{R^{2j-(1+\alpha+i)+1}}{2j-(1+\alpha+i)+1} - \frac{(ky)^{-2j+(1+\alpha+i)-1}}{2j-(1+\alpha+i)+1} \right] \\
&\quad + (ky)^{-(1+\alpha+i)+1} \int_1^\infty dr\, r^{-(1+\alpha+i)} (1 - \cos r) \\
&= -\sum_{j=1}^{\infty} (ky)^{2j} g_j(R,i) + (ky)^{-(\alpha+i)} h_i,
\end{aligned} \tag{A4}$$

where

$$g_j(R,i)=\frac{(-1)^{j+1}}{(2j)!}\frac{R^{2j-(1+\alpha+i)+1}}{2j-(1+\alpha+i)+1},$$

and

$$h_i=\sum_{j=1}^{\infty}\frac{(-1)^{j+1}}{(2j)!}\frac{1}{2j-(1+\alpha+i)+1}+\int_1^{\infty}dr\,r^{-(1+\alpha+i)}(1-\cos r).$$

There could be a $\ln(ky)$ term for $\alpha=$ integer in the above procedure, but it will not be the leading term, so it will not affect our derivation.

Now we go back to (A3), and we found

$$\int_0^{\infty}dr f(r)r^{d-1}(1-\cos rky)=\sum_{i=1}^{\infty}(ky)^{2j}\left[e_i(R)-\sum_{j=0}^{\infty}g_i(R,j)c_j\right]+\sum_{i=0}^{\infty}c_ih_i(ky)^{-(\alpha+i)}. \tag{A5}$$

Then we see

$$\tilde f(k)=1-c\sum_{i=1}^{\infty}k^{2i}\left\{\left[e_i(R)-\sum_{j=0}^{\infty}g_i(R,j)c_j\right]\times\int_0^1(1-y^2)^{(d-3)/2}y^{2i}\,dy\right\}-c\sum_{i=0}^{\infty}k^{-(\alpha+i)}\left[c_ih_i\int_0^1(1-y^2)^{(d-3)/2}y^{2i}\,dy\right]. \tag{A6}$$

Since $c_0\neq 0$

$$h_0=\sum_{j=1}^{\infty}\frac{(-1)^{j+1}}{(2j)!}\frac{1}{2j-\alpha}\neq 0.$$

The leading term in the second summation is in order of $k^{-\alpha}$. The leading term in the first summation is k^{2i} for some integer $i>0$. If $\alpha>2$, from the probability theory we know that the second monent exists, therefore, $\tilde f(k)=1-\beta k^2+o(k^2)$. From all the above procedures, we have shown for $d\geqslant 2$,

$$\tilde f(k)=1-\beta k^A+o(k^A)\ \text{with}\ A=\min\{2,\alpha\}\ \text{and}\ \beta\neq 0.$$

The proof for $d=1$ is very similar (and also simpler).

APPENDIX B: FINITENESS OF $C(0)-C(r)$

In this Appendix we show that if $C(r)=\lim_{m\to\infty}C_m(r)$ diverges, then $\lim_{m\to\infty}[C_m(0)-C_m(r)]$ is a finite function of r so that $\lim_{m\to\infty}C_m(r)=1$, where $C_m(r)=C_m(r)/C_m(0)$.

The Fourier transform of $C_m(r)$ may be written

$$\tilde C_m(k)=2\sum_{l=1}^{m}\left(1-\frac{l}{m}\right)[\tilde f(k)]^l=2\left[\tilde f\frac{H_m}{H_1}-\frac{\tilde f}{m}\frac{H_m}{H_1^2}+\frac{\tilde f^{m+1}}{H_1}\right], \tag{B1}$$

where $H_m(k)\equiv 1-[f(k)]^m$.

We need to examine the cases $d=1$ and $d=2$ separately.

(a) For $d=1$,

$$C_m(0)-C_m(r)\sim\int_0^{\infty}dk\,[1-\cos(kr)]\times\left\{\tilde f\frac{H_m}{H_1}-\frac{\tilde f}{m}\frac{H_m}{H_1^2}+\frac{\tilde f^{m+1}}{H_1}\right\}. \tag{B2}$$

Recalling that $\tilde f(k)<1$ for $k>0$ and $\tilde f(0)=1$, it is clear that for $m\to\infty$ only the first term in the curly brackets survives, so

$$C(0)-C(r)\sim\int_0^{\infty}dk\,(1-\cos(kr))\frac{\tilde f(k)}{H_1}. \tag{B3}$$

For $k\to 0$, the right-hand side of (B3) behaves like $\int_0(K^2/K^4)dk$ and so is convergent. For $k\to\infty$ the right-hand side of (B3) is also convergent, having the behavior $\int^{\infty}\tilde f(k)(1-\cos kr)dk$. Therefore $C(0)-C(r)$ is a finite function of r.

(b) For $D=2$; after integrating over the angular degrees of freedom,

$$C(0)-C(r)\sim\int_0^{\infty}k\,dk\frac{\tilde f(k)}{H_1}\times\left[\int_0^1(1-y^2)^{-1/2}(1-\cos(kry))dy\right]. \tag{B4}$$

For $k\to 0$ the right-hand side of (B4) has the behavior $\int_0(K^3/K^4)dk$, which is convergent, and for $k\to\infty$, the right-hand side of (B4) behaves like

$$\int^{\infty}k\,dk\tilde f(k)\left[\int_0^1(1-y)^{-1/2}(1-\cos(kry))dy\right],$$

which is also convergent. Therefore $C(0)-C(r)$ is a finite function of r in this case also.

APPENDIX C: LEADING BEHAVIOR OF $C(r)$ IN CROSSOVER REGIME

In this Appendix we study the crossover between the fractal and nonfractal regimes by examining the leading behavior of $C(r)$ for large r and values of α close to the critical crossover value.

(i) $d=1$. Here the critical value of α is $\alpha=1$. Let $\alpha=1-\epsilon$.

(a) $\epsilon\leqslant 0$. In this case we know from the results of Appendix A that $\lim_{m\to\infty}C_m(r)=\infty$ and $C(r)/C(0)=1$, which we interpret as implying nonfractal behavior with the LED $D=d=1$.

(b) $\epsilon>0$.

$$C(r)=\lim_{m\to\infty}C_m(r)=\int_0^{\infty}\frac{\tilde f(k)}{1-\tilde f(k)}\cos(kr)dk=r^{-\epsilon}\int_0^{\infty}\frac{(\tau/r)^{1-\epsilon}}{1-\tilde f(\tau/r)}\tilde f(\tau/r)\frac{\cos\tau}{\tau^{1-\epsilon}}d\tau. \tag{C1}$$

We now want to show that the leading behavior of the integrals as $r\to\infty$ is a constant proportional to $1/\epsilon$. To do this we note that

$$\frac{(\tau/r)^{1-\epsilon}}{1-\tilde{f}(\tau/r)}\tilde{f}\left(\frac{\tau}{r}\right)\equiv I\left(\frac{\tau}{r}\right) \tag{C2}$$

is bounded and that

$$\lim_{\tau\to\infty}\frac{\cos\tau}{\tau^{1-\epsilon}}=0. \tag{C3}$$

From this we can show that

$$\lim_{r\to\infty}\int_0^\infty I\left(\frac{\tau}{r}\right)\frac{\cos\tau}{\tau^{1-\epsilon}}d\tau=\int_0^\infty\left[\lim_{r\to\infty}I\left(\frac{\tau}{r}\right)\right]\frac{\cos\tau}{\tau^{1-\epsilon}}d\tau, \tag{C4}$$

and, since $I(0)$ is a finite constant, the integral in (C4) has the behavior

$$\int_0^\infty\left[\lim_{r\to\infty}I\left(\frac{\tau}{r}\right)\right]\frac{\cos\tau}{\tau^{1-\epsilon}}d\tau\sim\int_0^\infty\frac{\cos\tau}{\tau^{1-\epsilon}}d\tau\sim\frac{1}{\epsilon}.$$

The leading behavior of $C(r)$ for r large and $\epsilon>0$ is thus

$$C(r)\sim(1/\epsilon)r^{-\epsilon}.$$

Note that as $\epsilon\to0$ for large r,

$$C(r)\sim(1/\epsilon)[1-\epsilon\ln r]=(1/\epsilon)-\ln r.$$

Here we see explicitly that as $\epsilon\to0^+$, $C(r)$ consists of a divergent piece plus a finite function of r, which at the crossover point is proportional to $\ln r$.

(ii) $d=2$. The derivation of the behavior of $C(r)$ in this case is quite similar to the one-dimensional case. Defining $\alpha=2-\epsilon$, we have, as before, nonfractal behavior with the LED $D=d=2$ for $\epsilon\leqslant0$. For $\epsilon>0$ we can write

$$C(r)=\int_0^\infty k\,dk\frac{\tilde{f}(k)}{1-\tilde{f}(k)}\int_0^1(1-y^2)^{1/2}\cos(rky)dy$$

$$=r^{-\epsilon}\int_0^\infty d\tau\frac{(\tau/r)^{2-\epsilon}}{1-\tilde{f}(\tau/r)}\tilde{f}\left(\frac{\tau}{r}\right)\frac{J_0(\tau)}{\tau^{1-\epsilon}}. \tag{C5}$$

As before

$$\frac{(\tau/r)}{1-\tilde{f}(\tau/r)}\tilde{f}\left(\frac{\tau}{r}\right)$$

is bounded, and

$$\lim_{\tau\to\infty}\frac{J_0(\tau)}{\tau^{1-\epsilon}}=0,$$

so that

$$\lim_{r\to\infty}\int_0^\infty d\tau\frac{(\tau/r)^{2-\epsilon}}{1-\tilde{f}(\tau/r)}\tilde{f}\left(\frac{\tau}{r}\right)\frac{J_0(\tau)}{\tau^{1-\epsilon}}$$

$$=\int_0^\infty d\tau\left[\lim_{r\to\infty}\frac{(\tau/r)^{2-\epsilon}}{1-\tilde{f}(\tau/r)}\tilde{f}\left(\frac{\tau}{r}\right)\right]\frac{J_0(\tau)}{\tau^{1-\epsilon}}\sim\frac{1}{\epsilon}.$$

Therefore for small positive ϵ, the leading behavior of $C(r)$ for large r is

$$C(r)\sim(1/\epsilon)r^{-\epsilon}.$$

[1] B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, San Francisco, 1983), p. 359, and references therein.

[2] B. Mandelbrot, see Ref. 1, p. 288ff.

[3] B. D. Huges, E. W. Montroll, and M. F. Shlesinger, J. Stat. Phys. **28**, 111 (1982); E. W. Montroll and M. F. Shlesinger, in *Nonequilibrium Phenomena II: From Stochastics to Hydrodynamics*, edited by J. L. Lebowitz and E. W. Montroll (North-Holland, Amsterdam, 1984), p. 1.

[4] F. T. Hioe, in *Random Walks and Their Application in the Physical and Biological Sciences*, AIP Conf. Proc. No. 109, edited by M. F. Shlesinger and B. J. West (American Institute of Physics, New York, 1984), p. 85.

The hydrogen atom in phase space

L. Chetouani and T. F. Hammann
Faculté des Sciences et Techniques, Université de Haute Alsace, 4 rue des Frères Lumière, 68093 Mulhouse Cédex, France

(Received 22 May 1986; accepted for publication 12 November 1986)

The Hamiltonian of the three-dimensional hydrogen atom is reduced, in parabolic coordinates, to the Hamiltonians of two bidimensional harmonic oscillators, by doing several space-time transformations,separating the movement along the three parabolic directions (ξ,η,ϕ), and introducing two auxiliary angular variables ψ and ψ', $0\leqslant\psi$, $\psi'\leqslant 2\pi$. The Green's function is developed into partial Green's functions, and expressed in terms of two Green's functions that describe the movements along both the ξ and η axes. Introducing auxiliary Hamiltonians allows one to calculate the Green's function in the configurational space, via the phase-space evolution function of the two-dimensional harmonic oscillator. The auxiliary variables ψ and ψ' are eliminated by projection. The thus-obtained Green's function, save for a multiplicating factor, coincides with that calculated following the path-integral formalism.

I. WEYL FORMALISM

The Weyl correspondence (1927), as denoted by $A\leftrightarrow a(p,q)$, relates any operator

$$A=\int \alpha(u,v)e^{-(i/\hbar)(Qu+Pv)}\,du\,dv$$

of a Hilbert space to a phase-space function[1]

$$a(p,q)=\int \alpha(u,v)e^{-(i/\hbar)(qu+pv)}\,du\,dv,$$

the Heisenberg uncertainty principle, which is mathematically due to the noncommutativity of the observables, in the Hilbert space, being expressed in the Weyl formalism, by the Wigner distribution function, or quasiprobability, not everywhere positive, associated with the states of a physical system.[2] Two interesting formulations of Weyl's ideas have been proposed, by Kastler[3] and by Bayen *et al.*[4] In its most developed formulations, this theory is an alternative to Schrödinger wave mechanics, Heisenberg matrix mechanics, or Feynman functional mechanics.

Gracia-Bondia[5] has recently calculated the hydrogen atom spectrum and the Green's function in the configurational space, by making use of the well-known Kustaanheimo–Stiefel transformation, which has already been employed in celestial mechanics. His phase-space Green's function is that of the unconstrained four-dimensional harmonic oscillator.

Our aim is to calculate the hydrogen-atom spectrum and its Green's function, in parabolic coordinates.[6] These physical coordinates are useful especially when the system has a prevailing direction, for instance, the electric field direction in the Stark effect.[7]

By performing several space-time transformations, separating the parabolic variables in the Hamiltonian, and introducing the auxiliary variables ψ and ψ', it is possible to determine the H-atom spectrum from that of the two-dimensional harmonic oscillator (Sec. II). The generalized Green's function is obtained as the Fourier transform of the evolution function (Sec. III):

$$g(E)=-\frac{i}{\hbar}\int_0^\infty dt\, e^{iEt/\hbar}\mathscr{E}\left(-\frac{i}{\hbar}H_w t\right), \tag{1}$$

with $\mathscr{E}(-(i/\hbar)H_w t)\leftrightarrow U=\exp\{-(i/\hbar)tH\}$. $H_w\leftrightarrow H$. Here

$$\mathscr{E}\left(-\frac{i}{\hbar}tH_w\right)=1-\frac{i}{\hbar}H_w t+\frac{1}{2}\left(-\frac{it}{\hbar}\right)^2 H_w * H_w+\cdots$$

is a solution of the following equation:

$$i\hbar\frac{\partial\mathscr{E}}{\partial t}=H_w*\mathscr{E}=\mathscr{E}*H_w.$$

By taking into account the association between the product of the two operators A and B,

$$A\leftrightarrow a(\mathbf{p},\mathbf{q}),\quad B\leftrightarrow b(\mathbf{p},\mathbf{q}),$$

and a phase-space function defined by

$$A\cdot B\leftrightarrow a(\mathbf{p},\mathbf{q})*b(\mathbf{p},\mathbf{q})$$

$$=a(\mathbf{p},\mathbf{q})\exp\left\{\frac{\hbar}{2i}\left[\frac{\partial}{\partial\mathbf{p}}\frac{\partial}{\partial\mathbf{q}}-\frac{\partial}{\partial\mathbf{q}}\frac{\partial}{\partial\mathbf{p}}\right]\right\}b(\mathbf{p},\mathbf{q})$$

$$=\sum_{\substack{r,s,l,m\\k,n}}\left(\frac{i\hbar}{2}\right)^{r+s+l+m+k+n}\frac{(-)^{s+m+n}}{r!s!l!m!k!n!}\frac{\partial^{r+s+l+m+k+n}a}{\partial^r x\,\partial^s p_x\,\partial^l y\,\partial^m p_y\,\partial^k z\,\partial^n p_z}\frac{\partial^{r+s+l+m+k+n}b}{\partial^r p_x\,\partial^s x\,\partial^l p_y\,\partial^m y\,\partial^k p_z\,\partial^n z}, \tag{2a}$$

 0022-2488/87/030598-07$02.50

in a six-dimensional space, we express the Green's function in terms of Green's functions that describe the movement along the three axes.

The evolution function is related to the propagator K by the following equation:

$$K(\mathbf{q}_f,\mathbf{q}_i;t)=\frac{1}{(2\pi\hbar)^n}\int d^n p\,\exp\left\{\frac{i\mathbf{p}}{\hbar}(\mathbf{q}_f-\mathbf{q}_i)\right\}\times\mathscr{E}\left(-\frac{i}{\hbar}H_w t\right)\Big|_{\mathbf{q}=(\mathbf{q}_f+\mathbf{q}_i)/2}, \tag{2b}$$

where $2n$ is the phase-space dimension.

II. SPECTRUM OF THE THREE-DIMENSIONAL H ATOM

In Cartesian coordinates, the classical H-atom Hamiltonian is written as follows:

$$H_{\rm cl}=(1/2M)\{p_x^2+p_y^2+p_z^2\}-\alpha/r,\quad r^2=x^2+y^2+z^2, \tag{3}$$

M being the mass of the electron, and $\mathbf{p}=(p_x,p_y,p_z)$ and $\mathbf{q}=(x,y,z)$ satisfying the Hamilton equations

$$\frac{d\mathbf{p}}{dt}=-\frac{\partial H_{\rm cl}}{\partial\mathbf{q}}\quad\text{and}\quad\frac{d\mathbf{q}}{dt}=\frac{\partial H_{\rm cl}}{\partial\mathbf{p}}. \tag{4}$$

It is obviously impossible to calculate the H-atom spectrum from Eq. (3), because of the occurrence of the Coulombian term, which is indefinitely derivable and gives thus an infinite-order differential equation or an insolvable integrodifferential equation for the evolution function.

It is possible to find a coordinate change that transforms the Hamiltonian (3) into that of an harmonic oscillator, the evolution function of which can be calculated and is known.[4] We employ the space-time transformations defined in Refs. 6 and 7.

In parabolic coordinates,

$$x=(\xi\eta)^{1/2}\cos\phi,$$
$$y=(\xi\eta)^{1/2}\sin\phi,\quad 0\leqslant\xi,\eta<\infty,$$
$$z=\tfrac{1}{2}(\xi-\eta),\quad 0\leqslant\phi\leqslant 2\pi,$$

the Hamiltonian [Eq. (3)] is written as follows:

$$H_{\rm cl}=\frac{2}{M(\xi+\eta)}\{p_\xi^2\xi+p_\eta^2\eta\}+\frac{p_\phi^2}{2M\xi\eta}-\frac{2\alpha}{\xi+\eta}. \tag{5}$$

A. First time transformation (q,p,t)→(q,p,s)

By means of the "time" transformation

$$\frac{dt}{ds}=\frac{r(s)}{2}=\frac{\xi(s)+\eta(s)}{4}, \tag{6}$$

the Hamilton equations, which describe the evolution of $\mathbf{p}=(p_\xi,p_\eta,p_\phi)$ and $\mathbf{q}=(\xi,\eta,\phi)$ in the coordinate system $(\mathbf{q},\mathbf{p},t)$, become

$$\frac{d\mathbf{p}}{dt}=\frac{d\mathbf{p}}{ds}\cdot\frac{ds}{dt}=-\frac{\partial H_{\rm cl}}{\partial\mathbf{q}}\ \text{or}\ \frac{d\mathbf{p}}{ds}=-\frac{r}{2}\frac{\partial H_{\rm cl}}{\partial\mathbf{q}},$$

$$\frac{d\mathbf{q}}{dt}=\frac{d\mathbf{q}}{ds}\cdot\frac{ds}{dt}=\frac{\partial H_{\rm cl}}{\partial\mathbf{p}}\ \text{or}\ \frac{d\mathbf{q}}{ds}=\frac{r}{2}\frac{\partial H_{\rm cl}}{\partial\mathbf{p}},$$

in the coordinate system $(\mathbf{q},\mathbf{p},s)$.

The transformation (6) thus appears not entirely canonical. However, if

$$E=\frac{1}{2}M\dot{\mathbf{r}}^2(t)-\frac{\alpha}{r}=\frac{1}{2}M\dot{\mathbf{r}}^2(s)\frac{4}{r^2(s)}-\frac{\alpha}{r}$$

is the energy of the system and $H_{\rm cl}\simeq E$, there can be defined a pseudo-Hamiltonian $\mathscr{H}=(r/2)(H_{\rm cl}-E)$ which is nil, in the weak sense of Dirac, and which keeps the Hamilton equations unchanged:

$$\frac{d\mathbf{p}}{ds}=-\frac{r}{2}\frac{\partial H_{\rm cl}}{\partial\mathbf{q}}=-\frac{\partial}{\partial\mathbf{q}}\left(\frac{r}{2}(H_{\rm cl}-E)\right)=-\frac{\partial}{\partial\mathbf{q}}\mathscr{H},$$

$$\frac{d\mathbf{q}}{ds}=\frac{r}{2}\frac{\partial H_{\rm cl}}{\partial\mathbf{p}}=\frac{\partial}{\partial\mathbf{p}}\left(\frac{r}{2}(H_{\rm cl}-E)\right)=\frac{\partial}{\partial\mathbf{p}}\mathscr{H}.$$

Thus this pseudo-Hamiltonian $\mathscr{H}$, determining the movement in the coordinate system $(\mathbf{q},\mathbf{p},\mathbf{s})$, is

$$\mathscr{H}=\frac{r}{2}(H_{\rm cl}-E)=\frac{p_\xi^2\xi}{2M}+\frac{p_\eta^2\eta}{2M}+\frac{p_\phi^2}{8M}\left(\frac{1}{\xi}+\frac{1}{\eta}\right)-\frac{\alpha}{2}-\frac{E}{4}(\xi+\eta)\simeq 0. \tag{7}$$

The canonical variables $\mathbf{p}$ and $\mathbf{q}$, occurring in (7), satisfy the Hamilton equations. The "time" transformation eliminates the denominator term $(\xi+\eta)$ of (5), in introducing a new classical pseudopotential $-(E/4)(\xi+\eta)$, and in making $\alpha/2$ act as a pseudoenergy.

B. Separation of the movements

The angular variable ϕ is cyclic, thus $dp_\phi/ds=0$, and $p_\phi=p_{0\phi}$ is a constant of movement.

The equation $p_\phi=l_z=xp_y-yp_x$, allows the constant $p_{0\phi}$ to be calculated in the quantum case[4]:

$$\text{Spectrum}\{p_\phi\}=\text{Spectrum}\{l_z\}=m\hbar,$$

where $m=0,\pm 1,\pm 2,\pm 3,\ldots$.

Then the pseudo-Hamiltonian is the sum of two pseudo-Hamiltonians $\mathscr{H}_\xi$ and $\mathscr{H}_\eta$:

$$\mathscr{H}=\mathscr{H}_\xi+\mathscr{H}_\eta-\alpha/2\simeq 0, \tag{8}$$

with

$$\mathscr{H}_x=\frac{p_x^2 x}{2M}+\frac{\hbar^2m^2}{8Mx}-\frac{E}{4}x,\quad x=(\xi,\eta). \tag{9}$$

These equations separate the movements along the positive directions ξ and η.

The quantum eigenvalue $\hbar m$ can replace the generalized momentum p_ϕ; we let $\hbar\to 0$ and $m\to\infty$ as $\hbar m=$ const.

The pseudo-Hamiltonian (9) describes the movement of a particle having a variable mass M/x, in two pseudopotentials: the Coulombian potential $\hbar^2m^2/8Mx$ ($\hbar\to 0$, with $\hbar m=$ const), and the potential $-Ex/4$ of an electric field with the intensity $E/4$. It will be shown that every pseudo-Hamiltonian $\mathscr{H}_x$ is a constant of the movement. The evolution of $\mathscr{H}_x$ is indeed described by $(x=\xi,\eta)$,

$$\frac{d\mathscr{H}_x}{ds}=\{\mathscr{H}_x,\mathscr{H}\}_{\xi,\eta,p_\xi,p_\eta}=0,$$

the Poisson brackets being understood in the weak sense of

Dirac. Then, $\mathscr{H}_\xi \approx \beta_1$, $\mathscr{H}_\eta \approx \beta_2$, and, taking (8) into account,

$$\beta_1 + \beta_2 = \alpha/2. \tag{10}$$

Because the pseudo-Hamiltonians $\mathscr{H}_\xi$ and $\mathscr{H}_\eta$ contain some Coulombian terms, both constants β_1 and β_2 cannot be calculated now. It is clear that the pseudo-Hamiltonians $\mathscr{H}_\xi$ and $\mathscr{H}_\eta$ govern the movements along the axes ξ and η, respectively:

$$\frac{dx}{ds} = \frac{\partial \mathscr{H}}{\partial p_x} = \frac{\partial(\mathscr{H}_\xi + \mathscr{H}_\eta - \alpha/2)}{\partial p_x} = \frac{\partial \mathscr{H}_x}{\partial p_x},$$

$$\frac{\partial p_x}{ds} = -\frac{\partial \mathscr{H}}{\partial x} = -\frac{\partial(\mathscr{H}_\xi + \mathscr{H}_\eta - \alpha/2)}{\partial x} = -\frac{\partial \mathscr{H}_x}{\partial x},$$

$$x = (\xi,\eta) \text{ and } p_x = (p_\xi, p_\eta).$$

C. Second time transformation (q,p,s)→(q,p,θ)

The time transformation

$$\frac{ds}{d\theta} = \frac{1}{x(\theta)}, \tag{11}$$

i.e.,

$$\frac{ds}{d\theta_1} = \frac{1}{\xi(\theta_1)} \quad \text{and} \quad \frac{ds}{d\theta_2} = \frac{1}{\eta(\theta_2)},$$

aims at making constant the variable mass M/x in the kinetic terms of Eq. (9). This time transformation amounts to introducing, according to the procedure described in Sec. II A, two new pseudo-Hamiltonians, $\mathscr{H}'_1(\xi,p_\xi,\theta_1)$ and $\mathscr{H}'_2(\eta,p_\eta,\theta_2)$, nil in the weak sense of Dirac:

$$\mathscr{H}'_1 = (1/\xi)(\mathscr{H}_\xi - \beta_1) = \mathscr{H}_1 - E/4 \approx 0, \tag{12a}$$

$$\mathscr{H}'_2 = (1/\eta)(\mathscr{H}_\eta - \beta_2) = \mathscr{H}_2 - E/4 \approx 0, \tag{12b}$$

where

$$\mathscr{H}_i = \frac{p_x^2}{2M} + \frac{\hbar^2 m^2}{8Mx^2} - \frac{\beta_i}{x} \approx \frac{E}{4},$$

$$\hbar \to 0, \quad m \to \infty, \quad \hbar m = \text{const}, \tag{13}$$

$$x = (\xi,\eta), \quad p_x = (p_\xi, p_\eta), \quad i = (1,2).$$

The $\mathscr{H}_i$ being the new Hamiltonians governing the movements, the Hamilton equations

$$\frac{d\xi}{d\theta} = \frac{1}{\xi}\frac{\partial \mathscr{H}_\xi}{\partial p_\xi} = \frac{\partial}{\partial p_\xi}\left(\frac{1}{\xi}(\mathscr{H}_\xi - \beta_1)\right)$$

$$= \frac{\partial}{\partial p_\xi}\left(\mathscr{H}_1 - \frac{E}{4}\right) = \frac{\partial \mathscr{H}_1}{\partial p_\xi},$$

$$\frac{\partial p_\xi}{\partial \theta} = -\frac{1}{\xi}\frac{\partial \mathscr{H}_\xi}{\partial \xi} = -\frac{\partial}{\partial \xi}\left(\frac{1}{\xi}(\mathscr{H}_\xi - \beta_1)\right)$$

$$= -\frac{\partial}{\partial \xi}\left(\mathscr{H}_1 - \frac{E}{4}\right) = -\frac{\partial \mathscr{H}_1}{\partial \xi}$$

govern the movement along the positive axis ξ, in the coordinate system (ξ,p_ξ,θ_1). Similar equations are obtained for the movement along the axis η, in the coordinate system (η,p_η,θ_2), by changing $(\xi,p_\xi,\theta_1,\mathscr{H}_\xi,\mathscr{H}_1,\beta_1)$ into $(\eta,p_\eta,\theta_2,\mathscr{H}_\eta,\mathscr{H}_2,\beta_2)$.

Thus the transformation (11) makes the masses constant in the kinetic terms of (13), but it results in the occurrence of two new Coulombian potentials β_i/x, $i = (1,2)$, $x = (\xi,\eta)$, in Eq. (13). The potential $\hbar^2 m^2/8Mx$, which is Coulombian in (9), becomes centrifugal in (13). The occurrence of Coulombian terms in (13) makes it impossible to use this equation in order to determine the constants β_i $(i = 1,2)$.

D. Third transformation

The transformation $(\xi,p_\xi) \to (u,p_u)$, defined by

$$\xi = u^2 \quad \text{and} \quad p_\xi = p_u/2u, \tag{14}$$

changes the Hamiltonian $\mathscr{H}_1$ (13) into

$$\tilde{\mathscr{H}}_1 = \frac{p_u^2}{8Mu^2} + \frac{\hbar^2 m^2}{8Mu^4} - \frac{\beta_1}{u^2} \approx \frac{E}{4}$$

$$\left(\lim_{\substack{\hbar \to 0 \\ m \to \infty}} \hbar m = \text{const}\right). \tag{15}$$

The movement equations of Eq. (15) are, in the coordinate system (u,p_u,θ_1),

$$\frac{du}{d\theta_1} = \frac{\partial \tilde{\mathscr{H}}_1}{\partial p_u} \quad \text{and} \quad \frac{dp_u}{d\theta_1} = -\frac{\partial \tilde{\mathscr{H}}_1}{\partial u}.$$

In Eq. (15), the kinetic term has still a variable mass $4Mu^2$. It can be made constant following the usual procedure, by introducing a new time transformation $(\theta_1 \to \tau)$, which is defined by

$$\frac{d\theta_1}{d\tau} = 4u^2(\tau). \tag{16}$$

It provides a nil pseudo-Hamiltonian

$$K'_1 = 4_u^2(\tilde{\mathscr{H}}_1 - E/4) = K_1 - 4\beta_1 \approx 0;$$

or the pseudo-Hamiltonian

$$K_1 = p_u^2/2M + \hbar^2 m^2/2Mu^2 - Eu^2 \approx 4\beta_1, \tag{17}$$

(with the usual limit rules for $\hbar^2 m^2$), which governs the movement along the positive axis u in the system (u,p_u,τ):

$$\frac{du}{d\tau} = \frac{\partial K_1}{\partial p_u} \quad \text{and} \quad \frac{dp_u}{d\tau} = -\frac{\partial K_1}{\partial u}.$$

By doing the same space-time transformation in the case of the movement along the axis η,

$$\eta = v^2, \quad p_\eta = \frac{p_v}{2v}, \quad \frac{d\theta_2}{d\tau'} = 4v^2(\tau'),$$

the Hamiltonian $\tilde{\mathscr{H}}_2$ and the pseudo-Hamiltonian K_2 are obtained:

$$\tilde{\mathscr{H}}_2 = \frac{p_v^2}{8Mv^2} + \frac{\hbar^2 m^2}{8Mv^4} - \frac{\beta_1}{v^2} \approx \frac{E}{4}, \tag{18}$$

$$K_2 = \frac{p_v^2}{2M} + \frac{\hbar^2 m^2}{2Mv^2} - Ev^2 \approx 4\beta_2, \tag{19}$$

$\hbar^2 m^2$ being the constraint to the usual limit rules.

The pseudo-Hamiltonians K_1 and K_2 describe the two-dimensional isotropic harmonic oscillator $(E < 0)$, by setting

$$p_\phi = p_\psi = \hbar m \quad \text{in Eq. (17)},$$

$$p_\phi = p_{\psi'} = \hbar m \quad \text{in Eq. (19)},$$

$$\text{with } \lim_{\substack{\hbar\to 0\\ m\to\infty}} \hbar m = \text{const.}$$

Then

$$H_{\text{osc}} = \frac{p_u^2}{2M} + \frac{p_\psi^2}{2Mu^2} - Eu^2$$

$$= \frac{p_X^2 + p_Y^2}{2M} - E(X^2 + Y^2) \approx 4\beta_1, \tag{20a}$$

with

$$u^2 = X^2 + Y^2, \quad X = u\cos\psi, \quad Y = u\sin\psi,$$

and

$$H'_{\text{osc}} = \frac{p_v^2}{2M} + \frac{p_{\psi'}^2}{2Mv^2} - Ev^2$$

$$= \frac{p_{X'}^2 + p_{Y'}^2}{2M} - E(X'^2 + Y'^2) \approx 4\beta_2, \tag{20b}$$

with

$$v^2 = X'^2 + Y'^2, \quad X' = v\cos\psi', \quad Y' = v\sin\psi'.$$

These equations allow the constants β_1 and β_2 to be calculated[4] in the quantum case:

$$4\beta_1 = \hbar[-2E/M]^{1/2}[N_1 + \tfrac{1}{2} + N_2 + \tfrac{1}{2}], \quad (N_1,N_2) = 0,1,2,3,\ldots,\infty, \tag{21a}$$

with the condition

$$\text{Spectrum}\{p_\phi\} = \hbar m = \hbar(N_1 - N_2);$$

and

$$4\beta_2 = \hbar[-2E/M]^{1/2}(N_3 + \tfrac{1}{2} + N_4 + \tfrac{1}{2}), \quad (N_3,N_4) = 0,1,2,\ldots,\infty, \tag{21b}$$

with the condition

$$\text{Spectrum}\{p_\psi\} = \hbar m = \hbar(N_3 - N_4).$$

Accounting for Eq. (10), this gives us

$$E_n = -M\alpha^2/2\hbar^2 n^2,$$

$$n = n_1 + n_2 + |m| + 1 = 1,2,3,\ldots,\infty,$$

$$(n_1,n_2) = 0,1,2,3,\ldots,\infty, \quad m = 0, \pm 1, \pm 2,\ldots, \tag{22}$$

which is the well-known H-atom spectrum.

III. GREEN'S FUNCTION OF THE HYDROGEN ATOM

The Hamiltonian operator corresponding to (5) is written

$$\hat{H} = -\frac{\hbar^2}{2M}\frac{1}{\sqrt{g}}\partial_\alpha\sqrt{g}g^{\alpha\beta}\partial_\beta - \frac{\alpha}{r},$$

and as functions of the moments $\hat{p}_\alpha = (\hbar/i)g^{-1/4}\partial_\alpha g^{+1/4}$, conjugate to the $\hat{q}_\alpha$,

$$\hat{H} = (1/2M)g^{-1/4}\hat{p}_\alpha\sqrt{g}g^{\alpha\beta}\hat{p}_\beta g^{-1/4} - \alpha/r,$$

whereby the quadratic differential space element is

$$ds^2 = \frac{\xi+\eta}{4}\left[\frac{d\xi^2}{\xi} + \frac{d\eta^2}{\eta}\right] + \eta\xi\, d\phi^2.$$

As a result of the time transformation (6), a new, nil pseudo-Hamiltonian, is obtained:

$$\hat{\mathscr{H}} = (1/2M)\hat{p}_\alpha\sqrt{g}g^{\alpha\beta}\hat{p}_\beta + \sqrt{g}(-\alpha/r - E) = 0.$$

By redefining a Hilbert space by the scalar product without measure

$$\langle\psi_1|\psi_2\rangle = \int \psi_1^*\psi_2\, d\xi\, d\eta\, d\phi,$$

which amounts to changing $\hat{p}_\alpha$ into $-i\hbar\partial_\alpha$ in the aforesaid Hamiltonian, the Weyl transform can be obtained:

$$\hat{\mathscr{H}} \leftrightarrow \mathscr{H}_w = \frac{1}{2M}p_\alpha * \sqrt{g}g^{\alpha\beta} * p_\beta + \sqrt{g}\left(-\frac{\alpha}{r} - E\right)$$

$$= \frac{1}{2M}\left\{p_\xi * \xi * p_\xi + p_\eta * \eta * p_\eta + p_\phi * \frac{\xi+\eta}{4\xi\eta} * p_\phi\right\} - \frac{\alpha}{2} - \frac{E}{4}(\xi+\eta)$$

$$= \frac{1}{2M}\{p_\xi^2\xi + p_\eta^2\eta\} + \frac{p_\phi^2}{8M}\left(\frac{1}{\xi} + \frac{1}{\eta}\right) - \frac{\alpha}{2} - \frac{E}{4}(\xi+\eta) = 0. \tag{23}$$

The Weyl-transformed $\mathscr{H}_w$ thus coincides with the classical expression $\mathscr{H}$ [Eq. (7)]. The generalized Green's function is obtained from Eq. (1), in the space-time system $(\mathbf{p},\mathbf{q},s)$, by setting

$$G(E) = \frac{i}{\hbar}\int_0^\infty dt\, e^{+iEt/\hbar}\mathscr{C}\left(-\frac{i}{\hbar}H_w t\right)$$

$$= -\frac{i}{\hbar}\int_0^\infty ds\, \mathscr{C}\left(-\frac{i}{\hbar}\mathscr{H}\cdot s\right). \tag{24}$$

A. Decomposition into partial generalized Green's functions and separation of the movements

The variable ϕ being cyclic, its conjugated momentum p_ϕ is constant:

$$i\hbar\frac{dp_\phi}{ds} = -[\mathscr{H} * p_\phi - p_\phi * \mathscr{H}] = 0.$$

As a result, p_ϕ can only have discrete values.

By using Eqs. (1) and (2b) in order to separate the angular movement from other movements by means of integration over p_ϕ, the expression (24) becomes

$$G(\phi_f,\phi_i;E) = \frac{-i}{\hbar}\int_{-\infty}^{+\infty}\frac{dp_\phi}{2\pi\hbar}e^{(i/\hbar)p_\phi(\phi_f-\phi_i)} \times \int_0^\infty ds\, \mathscr{C}\left(-\frac{i}{\hbar}\mathscr{H}\cdot s\right)\Big|_{\phi=(\phi_f+\phi_i)/2}$$

$$= \frac{-i}{2\pi\hbar}\sum_{m=-\infty}^{+\infty} e^{im(\phi_f-\phi_i)} \times \int_0^\infty ds\, \mathscr{C}\left(-\frac{i}{\hbar}\mathscr{H}^m\cdot s\right), \tag{25}$$

where $\mathscr{H}^m$ is expression (8) without the conditions $\hbar\to 0$ and $m\to\infty$. The purely quantum potential $(\hbar^2m^2/8M) \times (1/\xi + 1/\eta)$ is thus introduced, and the independence of $\mathscr{C}$ from ϕ is taken into account.

As $\mathscr{H}_\xi * \mathscr{H}_\eta = \mathscr{H}_\eta * \mathscr{H}_\xi = \mathscr{H}_\xi\mathscr{H}_\eta$, with $\hbar\neq 0$ and m an integer, then

$$\mathscr{E}\left(\frac{-i}{\hbar}\mathscr{H}^m\cdot s\right)=e^{ias/2\hbar}\mathscr{E}\left(-\frac{i}{\hbar}\mathscr{H}_\xi\cdot s\right)\mathscr{E}\left(-\frac{i}{\hbar}\mathscr{H}_\eta\cdot s\right).$$

Each evolution function $\mathscr{E}(-i/\hbar\mathscr{H}_x\cdot s)$ is also the Fourier transform of a pseudo-Green's function:

$$\mathscr{E}\left(-\frac{i}{\hbar}\mathscr{H}_x s\right)=\frac{i}{2\pi}\int_{-\infty}^{+\infty}d\beta_i\, e^{-i\beta_i s/\hbar}g_E^m(\beta_i).$$

This allows (25) to be written as follows:

$$G(\phi_f,\phi_i;E)=\frac{i}{(2\pi)^3\hbar}\sum_{m=-\infty}^{+\infty}e^{im(\phi_f-\phi_i)}\int_0^\infty ds\, e^{ias/2\hbar}\times\left\{\left[\int_{-\infty}^{+\infty}d\beta_1\, e^{-(i/\hbar)\beta_1 s}g_E^m(\beta_1)\right]\times\left[\int_{-\infty}^{+\infty}d\beta_2\, e^{-(i/\hbar)\beta_2 s}g_E^m(\beta_2)\right]\right\}.$$

The integration with respect to s leads one to write

$$G(\phi_f,\phi_i;E)=\frac{-1}{(2\pi)^3}\sum_{m=-\infty}^{+\infty}e^{im(\phi_f-\phi_i)}\times\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}d\beta_1\,d\beta_2\,\frac{g_E^m(\beta_1)g_E^m(\beta_2)}{\alpha/2-\beta_1-\beta_2+i0},$$

and, taking into account the analyticity of the g_E^m's,

$$G(\phi_f,\phi_i;E)=\frac{1}{(2\pi)^2}\sum_{m=-\infty}^{+\infty}e^{im(\phi_f-\phi_i)}\times\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}d\beta_1\,d\beta_2\,\delta\left(\frac{\alpha}{2}-\beta_1-\beta_2\right)\times g_E^m(\beta_1)g_E^m(\beta_2).\tag{26}$$

We thus succeeded in expressing the function $G(\phi_f,\phi_i;E)$ as a function of two monodimensional pseudo-Green's functions g_E^m, depending on the variables ξ and η, respectively.

B. Green's function of the hydrogen atom

In the space-time systems $(\xi,p_\xi;\theta_1)$ and $(\eta,p_\eta;\theta_2)$ resulting from the transformation (11), an evolution function $\mathscr{E}$ can be associated with each $g_E^m(\beta_i)$:

$$g_E^m(\beta_i)=\frac{-i}{\hbar}\int_0^\infty d\theta_i\,\mathscr{E}(-i\mathscr{H}'_{iw}\theta_i),\quad i=(1,2),\tag{27}$$

where $\mathscr{H}'_{iw}$ is the Weyl transform of the operator $\hat{\mathscr{H}}'_i$,

$$\hat{\mathscr{H}}'_i\leftrightarrow\mathscr{H}'_{iw}=\frac{1}{\sqrt{x}}*\frac{p_x^2x}{2M}*\frac{1}{\sqrt{x}}+\frac{\hbar^2m^2}{8Mx^2}-\frac{\beta_i}{x}-\frac{E}{4}=\frac{p_x^2}{2M}+\frac{\hbar^2}{8Mx^2}(m^2-1)-\frac{\beta_i}{x}-\frac{E}{4}=\mathscr{H}_{iw}-E/4=0,\quad x=(\xi,\eta),\quad p_x=(p_\xi,p_\eta).$$

Furthermore

$$\mathscr{H}_{iw}=\frac{p_x^2}{2M}+\frac{\hbar^2(m^2-1)}{8Mx^2}-\frac{\beta_i}{x},\quad i=(1,2),$$

differs from the classical expression (13), only by the pure quantum correction $-\hbar^2/8Mx^2$, which is independent from the potential and only depends upon the transformation (11); but $\mathscr{H}_{iw}$ coincides with (13) at the limit $\hbar\to0$, $m\to\infty$, while $\hbar m=$ const.

Let

$$g_E^m(\beta_i)=-\frac{i}{\hbar}\int_0^\infty d\theta_i\, e^{iE\theta_i/4\hbar}\mathscr{E}\left(-\frac{i}{\hbar}\mathscr{H}_{iw}\theta_i\right)=-\frac{i}{\hbar}\int_0^\infty d\theta_i\, e^{iE\theta_i/4\hbar}\mathscr{E}\left(-\frac{i}{\hbar}\tilde{\mathscr{H}}_{iw}\theta_i\right),\tag{28}$$

where $\tilde{\mathscr{H}}_{iw}$ is the Weyl transform of the operator $\tilde{\hat{\mathscr{H}}}_i$ which is obtained by the punctual transformations (14), $(\xi,p_\xi;\theta_1)\to(u,p_u;\theta_1)$, and $(\eta,p_\eta;\theta_2)\to(v,p_v;\theta_2)$.

For example, for $\tilde{\mathscr{H}}_{1w}$ we get

$$\tilde{\hat{\mathscr{H}}}_1\leftrightarrow\tilde{\mathscr{H}}_{1w}=\frac{1}{8M}\left(\frac{p_u}{u}\right)*\left(\frac{p_u}{u}\right)+\frac{\hbar^2(m^2-1)}{8Mu^4}-\frac{\beta_1}{u^2}=\frac{p_u^2}{8Mu^2}+\frac{\hbar^2}{8Mu^4}\left(m^2-\frac{3}{4}\right)-\frac{\beta_1}{u^2}=\frac{E}{4},$$

which is different from the classical expression (15) only by the quantum correction $-3\hbar^2/32Mu^4$, but coincides with (15) at the limit $\hbar\to0$, $m\to\infty$, $\hbar m=$ const.

The time transformation $(u,p_u;\theta_1)\to(u,p_u,\tau)$ has no influence upon (28),

$$g_E^m(\beta_1)=-\frac{i}{\hbar}\int_0^\infty d\tau\, e^{i(4/\hbar)\beta_1\tau}\mathscr{E}\left(-\frac{i}{\hbar}K_{1w}\tau\right),$$

where

$$K_{1w}=\frac{1}{2M}u*\frac{p_u^2}{u^2}*u+\frac{\hbar^2}{2Mu^2}\left(m^2-\frac{3}{4}\right)-Eu^2=\frac{p_u^2}{2M}+\frac{\hbar^2}{2Mu^2}\left(m^2-\frac{1}{4}\right)-Eu^2=4\beta_1,$$

and is again different from (15) by a quantum correction $-\hbar^2/8Mu^2$, but coincides with (15) when $\hbar\to0$, $m\to\infty$, $\hbar m=$ const.

Finally, Eq. (26) becomes

$$G(\phi_f,\phi_i;E)=\frac{-i}{(2\pi)^2\hbar^2}\sum_{m=-\infty}^{+\infty}e^{im(\phi_f-\phi_i)}\times\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}d\beta_1\,d\beta_2\left\{\delta\left(\frac{\alpha}{2}-\beta_1-\beta_2\right)\times\int_0^\infty d\tau\, e^{(4i/\hbar)\beta_1\tau}\mathscr{E}\left(-\frac{i}{\hbar}K_{1w}\tau\right)\times\int_0^\infty d\tau'\, e^{(4i/\hbar)\beta_2\tau'}\mathscr{E}\left(-\frac{i}{\hbar}K_{2w}\tau'\right)\right\}.\tag{29}$$

By integrating with respect to β_1, β_2, and τ',

$$G(\phi_f,\phi_i;E)=\frac{-i}{8\pi\hbar}\sum_{m=-\infty}^{+\infty}e^{im(\phi_f-\phi_i)}\int_0^\infty d\tau\, e^{2i\alpha\tau/\hbar}\times\mathscr{E}\left(-\frac{i}{\hbar}K_{1w}\tau\right)\mathscr{E}\left(-\frac{i}{\hbar}K_{2w}\tau\right).\tag{30}$$

Now passing into the configuration space by integrating with respect to p_u and p_v [u and v vary from $-\infty$ to $+\infty$, whereby the movement is limited by an infinite barrier when $(u,v)<0$],

$$G(u_f,v_f,\phi_f,u_i,v_i,\phi_i;E) = -\frac{i}{8\pi\hbar^2}\sum_{m=-\infty}^{+\infty} e^{im(\phi_f-\phi_i)}\int_0^\infty d\tau\, e^{2i\alpha\tau/\hbar}\left\{\frac{1}{2\pi\hbar}\int_{-\infty}^{+\infty} dp_u\, e^{(i/\hbar)p_u(u_f-u_i)}\right.$$
$$\left.\times\mathscr{E}\left(-\frac{i}{\hbar}K_{1w}\tau\right)\Big|_{u=(u_f+u_i)/2}\frac{1}{2\pi\hbar}\int_{-\infty}^{+\infty} dp_v\, e^{(ip_v/\hbar)(v_f-v_i)}\mathscr{E}\left(-\frac{i}{\hbar}K_{2w}\tau\right)\Big|_{v=(v_f+v_i)/2}\right\}. \quad (31)$$

Following (2b)

$$\frac{1}{2\pi\hbar}\int_{-\infty}^{+\infty} dp_u\, e^{(ip_u/\hbar)(u_f-u_i)}\mathscr{E}\left(-\frac{i}{\hbar}K_{1w}\tau\right)\Big|_{u=(u_f+u_i)/2} = P(u_f,u_i;\tau),$$

where $P(u_f,u_i;\tau)$ is the propagator, which is written in the Feynman path integral formalism[8] as follows:

$$P(u_f,u_i;\tau) = \int \mathscr{D}u(\sigma)\mathscr{D}p_u(\sigma)\exp\left(\frac{i}{\hbar}\int_0^\tau d\sigma(p_u\dot{u}-K_{1w})\right)$$
$$= \int \mathscr{D}u(\sigma)\mathscr{D}p_u(\sigma)\exp\left[\frac{i}{\hbar}\int_0^\tau d\sigma\left(p_u\dot{u}-\frac{p_u^2}{2M}-\frac{\hbar^2}{2Mu^2}\left(m^2-\frac{1}{4}\right)+Eu^2\right)\right].$$

Introducing an auxiliary Hamiltonian[9] $(p_\psi^2-\hbar^2/4)/2Mu^2$, where ψ is an auxiliary angular variable $(0\leqslant\psi\leqslant 2\pi)$, allows the projection method[10] to be used:

$$P(u_f,u_i;\tau) = (u_fu_i)^{1/2}\int_0^{2\pi} d\psi_f\, e^{im\psi_f}\bar{P}(u_f,\psi_f,u_i,0;\tau),$$

where

$$\bar{P}(u_f,\psi_f,u_i,0;\tau) = \frac{1}{(u_fu_i)^{1/2}}\int \mathscr{D}(u,\psi)\mathscr{D}(p_u,p_\psi)\exp\left(\frac{i}{\hbar}\int_0^\tau d\sigma(p_u\dot{u}+p_\psi\dot{\psi}-H_{\rm osc})\right)$$
$$= \int \mathscr{D}X\mathscr{D}Y\mathscr{D}p_X\mathscr{D}p_Y\exp\left(\frac{i}{\hbar}\int_0^\tau d\sigma(p_X\dot{X}+p_Y\dot{Y}-H_{\rm osc})\right),$$

$H_{\rm osc}$ being the classical "Hamiltonian" [Eq. (20a)]. Naturally $\bar{P}$ can be related to the evolution function of the bidimensional harmonic oscillator

$$\bar{P}(u_f,\phi_f,u_i,0;\tau) = \frac{1}{(2\pi\hbar)^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} dp_X\, dp_Y\, e^{(i/\hbar)[p_X(X_f-X_i)+p_Y(Y_f-Y_i)]}\mathscr{E}\left(-\frac{i}{\hbar}H_{\rm osc}\tau\right)\Big|_{X=(X_f+X_i)/2,\ Y=(Y_f+Y_i)/2} \quad (32)$$

with[4]

$$\mathscr{E}\left(-\frac{i}{\hbar}H_{\rm osc}\tau\right) = \frac{1}{\cos^2(\omega\tau/2)}\exp\left\{-\frac{2i}{\hbar\omega}H_{\rm osc}\tan\frac{\omega\tau}{2}\right\},$$

and $E = -\frac{1}{2}M\omega^2\leqslant 0$.

Integrating with respect to p_X and p_Y, developing $\bar{P}$ into partial propagators, and integrating with respect to ψ_f lead us to write

$$P(u_f;u_i;\tau) = \frac{-i}{\hbar}M\omega\frac{(u_fu_i)^{1/2}}{\sin(\omega\tau)}$$
$$\times\exp\left\{\frac{i}{2\hbar}M\omega(u_f^2+u_i^2)\cot(\omega\tau)\right\}$$
$$\times I_m\left(\frac{iM\omega u_fu_i}{\hbar\sin(\omega\tau)}\right),$$

where I_m is the modified Bessel function.[11]

By reintroducing the initial variables $\xi = u^2$, $\eta = v^2$, the H-atom Green's function finally becomes

$$G(\mathbf{r}_f,\mathbf{r}_i;E) = \frac{iM^2\omega^2}{8\pi\hbar^3}(\xi_f\xi_i\eta_f\eta_i)^{1/4}\sum_{m=-\infty}^{+\infty} e^{im(\phi_f-\phi_i)}$$
$$\times\int_0^\infty \frac{d\tau\, e^{2i\alpha\tau/\hbar}}{\sin^2(\omega\tau)}I_m\left(\frac{-iM\omega(\xi_f\xi_i)^{1/2}}{\hbar\sin(\omega\tau)}\right)$$
$$\times I_m\left(\frac{-iM\omega(\eta_f\eta_i)^{1/2}}{\hbar\sin(\omega\tau)}\right)$$
$$\times\exp\left\{\frac{iM\omega}{2\hbar}(\xi_f+\xi_i+\eta_f+\eta_i)\cot(\omega\tau)\right\}.$$

It coincides, save for the factor $4(\xi_f\xi_i\eta_f\eta_i)^{1/4}$, with the result obtained by means of the functional formalism.[6] This factor is due to the time transformations $\theta_1\to\tau$, $\theta_2\to\tau'$ [Eq. (16)], and was ignored in Eq. (29).

The identity[12]

$$\sum_{m=-\infty}^{+\infty} e^{im\phi}I_m(z)I_m(z') = I_0[(z^2+z'^2-2zz'\cos\phi)^{1/2}]$$

allows the Green's function [Eq. (32)] to be written in a compact form,[5,7] after changing $\omega\to i\omega$.

IV. CONCLUSION

We obtained the H-atom spectrum by classical transformations of the Hamiltonian, which lead to the classical Hamiltonians of two bidimensional harmonic oscillators, with well known spectra and evolution functions, in the phase space [the Kepler problem with SU(2) ⊗ SU(2) symmetry].

It seems difficult to obtain the Green's function in the phase space only as a function of physical coordinates, like the parabolic coordinates, if one avoids the Kustaanheimo–Stiefel transformation.

Thus, before doing the Weyl transformation, we have written in the Hilbert space the Hamiltonian operators obtained from different time–space transformations of the Schrödinger equation.

The Weyl transformations of these Hamiltonians enable us to obtain the evolution functions.

The Green's function was obtained by a Fourier expansion [Eq. (26)]:

$$G(\phi_f,\phi_i;E) = \frac{i}{(2\pi)^2}\sum_m e^{im(\phi_f-\phi_i)} \times\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} d\beta_1\, d\beta_2\, \delta\left(\frac{\alpha}{2}-\beta_1-\beta_2\right) \times g_E^m(\beta_1)g_E^m(\beta_2),$$

where the Dirac distribution takes Eq. (10) into account, and $g_E^m(\beta_1)$ and $g_E^m(\beta_2)$ are the pseudo-Green's functions corresponding to the projected harmonic oscillator Hamiltonians [(20a) and (20b)]. The intimate relation[8] between the Weyl formalism and the Feynman path integral formalism, particularly with respect to the Coulombian problem, should be noted.[5,9] It would be interesting to get the spectrum of the H atom in the case of spherical coordinates.

[1]S. De Groot, *La Transformation de Weyl et la fonction de Wigner* (Les Presses de l'Université de Montréal, Montreal, 1974); M. S. Marinov, Phys. Rep. **60**, 1 (1980).
[2]J. E. Moyal, Proc. Cambridge Philos. Soc. **45**, 99 (1949).
[3]D. Kastler, Commun. Math. Phys. **1**, 14 (1965); G. Loupias and M. Solé, *ibid.* **2**, 31 (1966).
[4]F. Bayen, M. Flato. C. Frondstal, A. Lichnerowicz, and D. Sternheimer, Ann. Phys. (NY) **111**, 61, 111 (1978).
[5]J. M. Gracia-Bondia, Phys. Rev. A **30**, 691 (1984).
[6]L. Chétouani and T. F. Hammann, Nuovo Cimento B, to be published.
[7]L. Chétouani and T. F. Hammann, Phys. Rev. A **34**, 4737 (1986).
[8]M. M. Mizrahi, J. Math. Phys. **16**, 2201 (1975); P. Sharan, Phys. Rev. D **20**, 414 (1979).
[9]I. H. Duru, and H. Kleinert, Phys. Lett. B **84**, 185 (1979); Fortschr. Phys. **30**, 401 (1982); I. H. Duru, Phys. Lett. A **112**, 421 (1985).
[10]J. M. Gipson, Phys. Rev. Lett. **48**, 1511 (1982).
[11]E. Whittaker and G. Watson, *A Course of Modern Analysis* (Cambridge U.P., Cambridge, 1952), p. 372.
[12]H. Bateman, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. 2, p. 101, Eq. (31).

The generalized Morse oscillator in the SO(4,2) dynamical group scheme

A. O. Barut
Department of Physics, University of Colorado, Boulder, Colorado 80309

A. Inomata
Department of Physics, State University of New York at Albany, Albany, New York 12222

R. Wilson
Department of Mathematics, University of Texas, San Antonio, Texas 78285

(Received 3 April 1986; accepted for publication 10 September 1986)

A family of the Morse oscillators with certain quantized coupling constants are described as composite objects in the framework of the SO(4,2) dynamical group scheme. Although a single Morse oscillator can be solved by the subgroup SO(2,1) of SO(4,2), this SO(2,1) is not the spectrum generating group, the set of all energy levels is given by the representation of another particular one-parameter subgroup of SO(4,2), which is the dynamical group of a single Morse oscillator. The continuous spectra of this oscillator and other variations of the Morse potential are also discussed by making an analytic continuation from the Morse potential well to the Morse barrier.

I. INTRODUCTION

Recently, some interest has been revived in the Morse potential problem from both the physical application and the calculational technique points of view.[1-6] In particular, the SO(2,1) algebraic approach has been successfully applied to the study of the Morse oscillator model for molecular and nuclear anharmonic vibrations.[1-3] Although the algebraic method has proved powerful for many quantum problems,[7-9] there is no general procedure for constructing a necessary algebra. It is certainly desirable to have a scheme within which a class of problems, if not all, can be treated in a unified manner. A possible candidate for such a general scheme is the one based on the SO(4,2) dynamical group.[7,10] In this paper, we reexamine the Morse problem in one dimension in the context of the SO(4,2) scheme.

The Hamiltonian of the Morse system is given by

$$H = (1/2m)p^2 + Ae^{-2ax} - Be^{-ax}, \tag{1.1}$$

where a is a positive constant. For the standard Morse oscillator,[11] the constants A and B are positive and related by $2A = B$. In our discussion, we include the nonpositive real values of A and B for generality and for other possible applications. In the coordinate representation, the Schrödinger equation, $H|\Psi\rangle = E|\Psi\rangle$, is written as

$$\left[\frac{d^2}{d\xi^2} - \frac{2mA}{a^2}e^{-2\xi} + \frac{2mB}{a^2}e^{-\xi} + \frac{2mE}{a^2}\right]\Psi(\xi) = 0, \tag{1.2}$$

where we have set $\hbar = 1$ and $\xi = ax$. Usually, this equation is solved for the energy eigenvalues and the corresponding wave functions.

However, in this paper, we treat the Morse oscillator as a composite system obeying a wave equation based on the dynamical group SO(4,2).[7] A large number of relativistic and nonrelativistic quantum systems belong to this general framework.[8] In Sec. II, we discuss the SO(4,2) scheme for the Morse system, and show that the basic equation of the SO(4,2) scheme can be reduced in a particular representation to the Schrödinger equation for a family of Morse systems (1.2). In Sec. III, we obtain algebraically the discrete and continuous energy spectrum of the Morse oscillator ($A > 0$). We also find the energy eigenfunctions from the basic equation in special representations. Sections IV and V cover the Morse barriers with $A < 0$ and $A = 0$, solutions for which are related to those of the oscillator case ($A > 0$) by tilt transformations. Some basic and necessary information of the dynamical group SO(4,2) and its most degenerate unitary irreducible representation are put together in the Appendix. Throughout the paper, we denote the special orthogonal group in N dimensions by SO(N) and its associated algebra by so(N).

II. SO(4,2) SCHEME FOR THE MORSE SYSTEM

The Morse oscillator is a composite system that, we assume, obeys the wave equation[7]

$$W|\Psi\rangle = 0, \tag{2.1}$$

with a relativistically covariant wave operator

$$W = \alpha_1 P^\mu \Gamma_\mu + \alpha_2 P^2 + \alpha_3 P^2 S + \beta S + \gamma, \tag{2.2}$$

constructed on the carrier space of SO(4,2) $\otimes$ $T(3,1)$. In (2.2), Γ_μ ($\mu = 0, 1, 2, 3$) and S are 5 of 16 operators in the algebra of SO(4,2), P_μ are the generators of the space-time translation $T(3,1)$, and $P^2 = P_\mu P^\mu = M^2$ is the invariant mass squared. We characterize the system by selecting the parameters of (2.2) to be

$$\alpha_1 = [(a^2/2m) + A]/M, \quad \alpha_2 = \alpha_3 = 0, \qquad \beta = (a^2/2m) - A, \quad \gamma = -B. \tag{2.3}$$

In the rest frame where $P_\mu = (M, 0, 0, 0)$, the wave equation (2.1) takes the form

$$[(a^2/2m + A)\Gamma_0 + (a^2/2m - A)S - B]|\Psi\rangle = 0. \tag{2.4}$$

The physical state in a moving frame can be obtained by boosting the rest frame solution. However, confining our interest in a nonrelativistic Morse system, we consider the choice (2.3) as a characterization of the system in the nonre-

 0022-2488/87/030605-07$02.50

lativistic limit. For a relativistic Morse oscillator, we would have to make a different selection of parameters. In (2.3) the parameter α_1 depends on the invariant mass of the whole Morse system. A similar situation occurs in other composite systems.[8]

To find the rest frame solution of (2.4) we perform the tilt transformation,

$$|\Phi\rangle = e^{-i\theta T}|\Psi\rangle. \tag{2.5}$$

The tilt angle θ may be fixed appropriately to specify the group state $|\Phi\rangle$. The tilt operator T, and other two operators, Γ_0 and S, form the so(2,1) subalgebra of so(4,2),

$$[\Gamma_0,S] = iT, \quad [T,\Gamma_0] = iS, \quad [S,T] = -i\Gamma_0. \tag{2.6}$$

The group states $|\Phi\rangle$ are chosen to be the basis states of the most degenerate unitary irreducible representation of SO(4,2). With respect to the subgroup SO(2,1), they are the eigenstates of the Casimir operator $Q^2 = \Gamma_0^2 - S^2 - T^2$,

$$Q^2|\Phi\rangle = \varphi(\varphi+1)|\Phi\rangle, \tag{2.7}$$

and are simultaneously eigenstates of a linear combination of Γ_0 and S, which is to be specified by a particular choice of the tilted angle θ,

$$e^{-i\theta T}We^{i\theta T}|\Phi\rangle = 0, \tag{2.8}$$

or

$$[((a^2/2m)e^{\theta} + Ae^{-\theta})\Gamma_0 + ((a^2/2m)e^{\theta} - Ae^{-\theta})S - B\,]|\Phi\rangle = 0. \tag{2.9}$$

In the ξ-representation given by (A17) and (A18) in the Appendix, we can express (2.9) as

$$\left[\frac{d^2}{d\xi^2} - \frac{2mA}{a^2}k^2e^{-2\theta}e^{-2\xi} + \frac{2mB}{a^2}ke^{-\theta}e^{-\xi} - \left(\varphi+\frac{1}{2}\right)^2\right]\Phi(\xi) = 0. \tag{2.10}$$

This equation coincides with the Schrödinger equation (1.2) for the Morse problem provided that the following identifications are made:

$$k = e^{\theta}, \tag{2.11}$$

and

$$E = -\left(\frac{a^2}{2m}\right)\left(\varphi+\frac{1}{2}\right)^2 = -\left(\frac{a^2}{2m}\right)\left[\varphi(\varphi+1)+\frac{1}{4}\right]. \tag{2.12}$$

Thus we see that finding a solution of (2.9) under conditions (2.11) and (2.12) amounts to solving the Schrödinger equation of (1.2).

In previous applications of the SO(4,2) scheme, the generators of the so(2,1) subalgebra are related to the energy operator, and a fixed value of the parameter μ, i.e., a fixed single representation of so(2,1) contains all the energy levels. In contrast, for the Morse system, we reverse the above procedure. First, we fix the eigenvalue of an appropriate SO(2,1) operator through the wave equation (2.9) and let the energy dependent parameter μ, i.e., the Casimir operator of SO(2,1), vary. This procedure allows us to obtain a complete description of the Morse system in one dimension including bound and scattering states. Here,

$$\mu^2 = -\varphi(\varphi+1) = (2mE/a^2) + \tfrac{1}{4}. \tag{2.13}$$

The eigenvalue of Γ_0 is now fixed by geometry of the Morse potential, whereas the energy E is related to the eigenvalues of the Casimir operator of SO(2,1). Hence we see that for each physical state we use a different representation of SO(2,1). In other words, we consider a family of representations T_φ of SO(2,1) and take one state from each corresponding to a fixed value of Γ_0. Therefore, SO(2,1) is not the spectral generating group of the system. The energy levels of a given Morse oscillator form a representation of a subgroup of SO(4,2) commuting with Γ_0, which is in general SO(4). Since the energy spectrum is nondegenerate for the one-dimensional case, we have only a one-parameter subgroup of SO(4), i.e., the one generated by L_{34}, which lowers and raises φ by one unit for a fixed value of L_{12}. This accounts for the finite number of bound states. Furthermore, we see that, for the Morse oscillator in three dimensions, the full SO(4) representation will be the underlying Hilbert space of states.

III. THE MORSE OSCILLATOR ($A>0$, $B>0$)

To discuss the Morse oscillator ($A>0$, $B>0$) in the SO(4,2) scheme, we start with the basic wave equation (2.9) tilted in the rest frame, and condition (2.13). Choosing the tilt angle θ in (2.9) to be

$$\theta = \tfrac{1}{2}\ln(2mA/a^2), \tag{3.1}$$

we reduce (2.9) into the form

$$[(2a^2A/m)^{1/2}\Gamma_0 - B]|\Phi\rangle = 0. \tag{3.2}$$

Of course, (3.2) is free of representation, so that it can be handled in any representation. However, it is important to notice that solutions in different representations are not all physically equivalent.

The compact operator Γ_0 in (3.2) can be diagonalized as[7,9]

$$\Gamma_0|\Phi_n\rangle = n|\Phi_n\rangle, \tag{3.3}$$

where the discrete eigenvalue n is either bounded below

$$D^+(\varphi): n = -\varphi + s \quad (\varphi < -\tfrac{1}{2}), \tag{3.4}$$

or bounded above

$$D^-(\varphi): n = \varphi - s \quad (\varphi < -\tfrac{1}{2}), \tag{3.5}$$

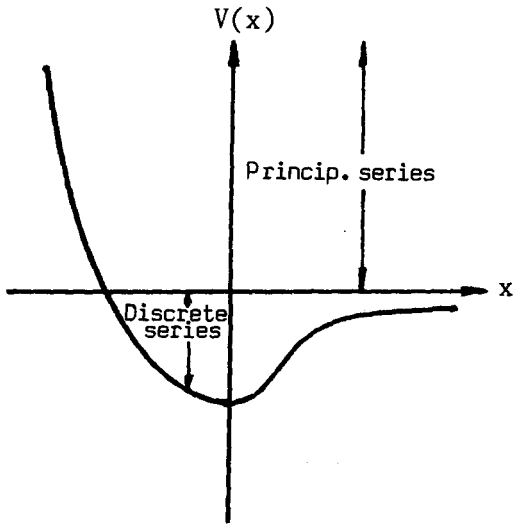
with $s = 0, 1, 2,...$, as shown in (A24) and (A25). On the basis, however, the dynamical equation (3.2) fixes the eigenvalue of Γ_0 to be

$$n = B(m/2a^2A)^{1/2}. \tag{3.6}$$

This is a point fixed in the homogeneous space $SO(2,1)/\Gamma_0$. We see from (3.6) that n is fixed. Then for a fixed energy E, φ is fixed. Hence by (3.4) s is fixed. Thus, as we stated above, we have one physical state in each representation of SO(2,1).

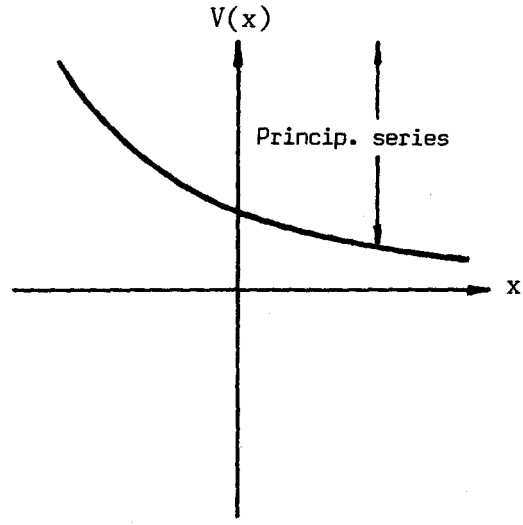
If $B>0$, i.e., if $n>0$, then the $D^+(\varphi)$ representation (3.4) is appropriate, but the $D^-(\varphi)$ representation (3.5) is not. Using (3.4) and (3.6) in (2.12), we obtain the discrete energy spectrum for the bound states in the Morse potential (see Fig. 1),

$$E_s = -\frac{a^2}{2m}\left[B\left(\frac{m}{2a^2A}\right)^{1/2} - \left(s+\frac{1}{2}\right)\right]^2, \tag{3.7}$$



A > 0, B > 0

FIG. 1. The Morse oscillator with $A>0$ and $B>0$: The principal continuous series is for scattering states and the discrete series for bound states.



A ≧ 0, B < 0

FIG. 2. The Morse barrier with $A\geqslant 0$ and $B<0$: Only the principal series is involved for the positive energy continuum.

where

$$s=0,\ 1,\ 2,\ldots <B(m/2a^2A)^{1/2}-\tfrac{1}{2}.$$

In particular, setting $B=2A$ in (3.7) yields the standard result,

$$E_s=-\frac{a^2}{2m}\left[\left(\frac{2mA}{a^2}\right)^{1/2}-\left(s+\frac{1}{2}\right)\right]^2, \tag{3.8}$$

with

$$s=0,\ 1,\ 2,\ldots <(2mA/a^2)^{1/2}-\tfrac{1}{2}.$$

The scattering states of the Morse oscillator ($A>0$, $B>0$) correspond to the continuous values of φ in the principal series of representation (A27),

$$D_p(\varphi):\varphi=-\tfrac{1}{2}+i\sigma \quad (\sigma \text{ real}), \tag{3.9}$$

thus belonging to the continuous energy spectrum,

$$E=a^2\sigma^2/2m. \tag{3.10}$$

For this oscillator, there is no continuum corresponding to the value of φ in the supplementary series (A28). Again for each energy we have a different representation of SO(2,1).

For $B<0$, i.e., for $n<0$, the $D^+(\varphi)$ representation (3.4) is not applicable, but the $D^-(\varphi)$ representation (3.5) would appear consistent. The $D^-(\varphi)$ series, however, leads us to the discrete energy spectrum

$$E'_s=-(a^2/2m)[-B(m/2a^2A)^{1/2}-(s+\tfrac{1}{2})]^2,$$

$$s=0,\ 1,\ 2,\ldots <-B(m/2a^2A)^{1/2}.$$

This is physically undesirable since the Morse system with $B<0$ does not have a potential well in which the system is bound (see Fig. 2). In fact, there are no finite eigenfunctions belonging to the discrete spectrum E'_s. All allowed states of this Morse system ($A>0$, $B<0$) correspond, like the scattering states of the Morse oscillator ($A>0$, $B>0$), to the values of φ in the principal series (3.9), carrying the same continuous energy spectrum as (3.10). The limiting case where $B=0$ can also be included in the continuous spectrum belonging to the principal series (3.9).

To find the energy eigenfunctions, we employ Γ_0 in the R-representation (A14) to reduce (3.3) to the Whittaker equation[12]

$$\left[\frac{d^2}{dR^2}-\frac{(\varphi+\frac{1}{2})^2-\frac{1}{4}}{R^2}+\frac{n}{R}-\frac{1}{4}\right]\Phi(R)=0, \tag{3.11}$$

whose solutions are given in terms of the Whittaker functions or the confluent hypergeometric functions.[13] For the bound states with $\varphi=s-n$ ($B>0$), we choose a solution regular at $R=0$. Transforming the R variable back into the x variable, we obtain the wave function

$$\Phi_s(x)=M_{n,n-s-1/2}(2ke^{-ax}) \tag{3.12}$$

or

$$\Phi_s(x)=(2k)^{n-s}e^{-(n-s)ax}\times\exp(-ke^{-ax})F(-s,2n-ns;2ke^{-ax}), \tag{3.13}$$

where $n=B(m/2a^2A)^{1/2}$ as fixed by (3.5) and

$$k=(2mA/a^2)^{1/2}, \tag{3.14}$$

as chosen by (3.1) via (2.11). For the continuous energy states with $\varphi=-\frac{1}{2}+i\sigma$ (σ real), the finite solution in the valid range of x is

$$\Phi_E(x)=W_{n,i\sigma}(2ke^{-ax}), \tag{3.15}$$

where $\sigma=(2mE/a^2)^{1/2}$ and n are given by (3.6) with $B\lesseqgtr 0$. For $B<0$, if one insists on using the $D^-(\varphi)$ discrete series, one can get from (3.12), by setting $\varphi=n+s$,

$$\Phi_{E'}(R)=M_{n,\pm(n+s+1/2)}(R).$$

However, these solutions are divergent at either $R=0$ or $R=\infty$, which we consider unphysical.

Alternatively, using the ρ-representation (A20), we can write (3.2) as a differential equation of the Infeld–Hull type,[9]

$$\left[\frac{d^2}{d\rho^2}+\frac{c_1}{\rho^2}+c_2\rho^2+c_3\right]\Phi(\rho)=0, \tag{3.16}$$

with

$$c_1 = -4(\varphi+\tfrac{1}{2})^2 + \tfrac{1}{4} = (8mE/a^2) + \tfrac{1}{4}, \tag{3.17}$$

$$c_2 = -4k^2 = -8mA/a^2, \tag{3.18}$$

$$c_3 = 8k(mB^2/2a^2A)^{1/2} = 8mB/a^2. \tag{3.19}$$

The discrete energy spectrum (3.6) can immediately be obtained from the formula of Wybourne,[9]

$$4s + 2 + (1-4c_1)^{1/2} = c_3(-c_2)^{-1/2}. \tag{3.20}$$

For the continuous spectrum, we have to go back to (2.12) with (3.9). The corresponding energy eigenfunctions (3.12) and (3.15) follow via the solutions of (3.16)

$$\Phi(x) = \rho^{1/2}\Phi(\rho), \tag{3.21}$$

with $\rho = \exp(-\frac{1}{2}ax)$.

In comparison with the usual radial wave equation, we notice that in the R-variable equation (3.11), the energy appears as an "angular momentum" and n as a "coupling constant." The coupling constant, being quantized, represents a family of Morse oscillators. We may say that SO(2,1) is the dynamical group of states of a family of different Morse oscillators belonging to the same energy E, whereas the dynamical group of a given single oscillator is the group SO(3). Thus, even for a one-dimensional system, the larger group SO(4,2) seems to be necessary, which contains both the above SO(3) and SO(2,1) groups.

At this point, it may be relevant to remark that the Morse oscillator can also be treated exactly by path integration.[4] This is due to the fact that the Morse oscillator is reducible to the Infeld–Hull form, which has been known to be path integrable.[14]

IV. THE MORSE BARRIER ($A<0$)

The Morse system (2.9) with $A>0$ and $B>0$ has bound states and scattering states. However, if $A>0$ and $B<0$ or if $A<0$, there are no bound states. In the case where $A<0$ (see Figs. 3 and 4), the choice of the tilt angle (3.1) is irrelevant. Thus, instead, we choose

$$\theta = \tfrac{1}{2}\ln(-2mA/a^2), \tag{4.1}$$

reducing (2.9) into

$$[(-2a^2A/m)^{1/2}S - B]|\Phi\rangle = 0. \tag{4.2}$$

Since S is a noncompact operator, its diagonalization yields a continuous eigenvalue[7]

$$S|\Phi_\nu\rangle = \nu|\Phi_\nu\rangle, \tag{4.3}$$

where $-\infty < \nu < \infty$. On the $|\Phi_\nu\rangle$ basis, (4.2) fixes the value of ν to be

$$\nu = B(-m/2a^2A)^{1/2}, \tag{4.4}$$

which is a point on the real projective line $SO(2,1)/S$. No real discrete spectrum of φ corresponds to this case. The positive and negative energy continua belong, respectively, to the principal series of representation (A27),

$$\varphi = -\tfrac{1}{2} + i\sigma \quad (\sigma \text{ real}), \tag{4.5}$$

and the supplementary series (A28),

$$\varphi = \text{real}. \tag{4.6}$$

However, a complex discrete spectrum may also be obtained by considering the $D^+(\varphi)$ discrete series of representation with $n \to iv$ in (3.4). The complex energy thus obtained is

$$E_s = (a^2/2m)[B(-m/2a^2A)^{1/2} + i(s+\tfrac{1}{2})]^2, \quad s = 0, 1, \ldots, \tag{4.7}$$

and this corresponds to the discrete tunneling of the positive energy continuum into the potential hill. Besides multiple reflections inside the hill, the trapped waves with energy given by $\mathrm{Im}\, E_s$ would escape outside of the hill.

In the R-representation (A15), the tilted equation (4.2) can be expressed as

$$\left[\frac{d^2}{dR^2} - \frac{\varphi(\varphi+1)}{R^2} + \frac{\nu}{R} + \frac{1}{4}\right]\Phi(R) = 0, \tag{4.8}$$

with $\nu = B(-m/2a^2A)^{1/2}$. If we set $R' = -iR$, then (4.8)

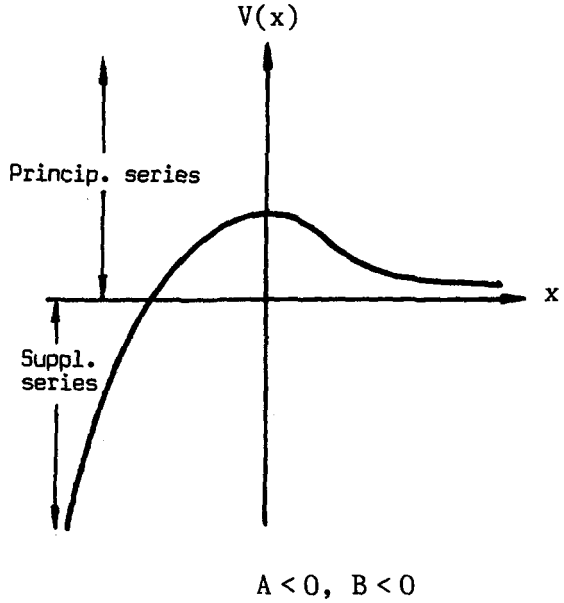


FIG. 3. The Morse barrier with $A<0$ and $B<0$: The principal series is for the positive energy continuum and the supplementary series for the negative continuum.

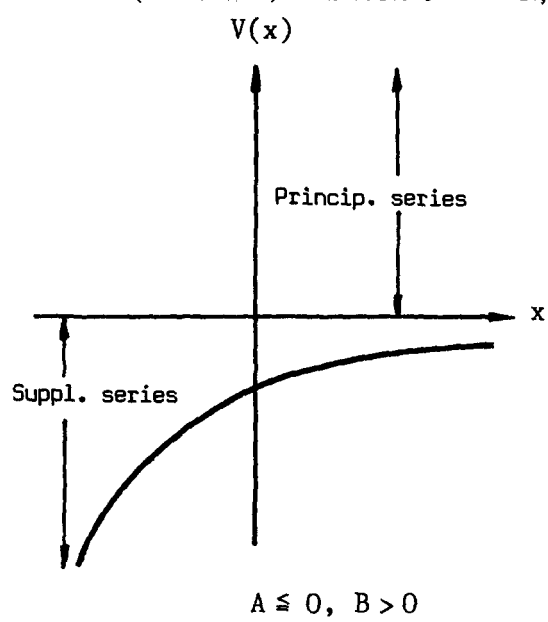


FIG. 4. The Morse barrier with $A<0$ and $B>0$: The principal series is for the positive continuum and the supplementary series for the negative continuum.

becomes

$$\left[\frac{d^2}{dR'^2}-\frac{(\varphi+\frac{1}{2})^2-\frac{1}{4}}{R'^2}+\frac{i\nu}{R'}-\frac{1}{4}\right]\Phi(R')=0, \tag{4.9}$$

which is identical in form with the Whittaker equation (3.11). Since we have defined $R=2ke^{-ax}$ with $k=(2mA/a^2)^{1/2}$ in (3.11), we have to recognize that $R'=-iR=(-8mA/a^2)^{1/2}e^{-ax}$ is now a real variable ranging from zero ($x\to\infty$) to infinity ($x\to-\infty$). For the positive energy continuum, regardless of whether $B\gtrless 0$, the solution is of the form

$$\Phi(x)=CM_{i\nu,i\sigma}(2k'e^{-ax})+C'M_{i\nu,-i\sigma}(2k'e^{-ax}), \tag{4.10}$$

and for the negative energy continuum

$$\Phi(x)=CM_{i\nu,\sigma}(2k'e^{-ax})+C'M_{i\nu,-\sigma}(2k'e^{-ax}), \tag{4.11}$$

where

$$\nu=B\left(\frac{m}{2a^2|A|}\right)^{1/2},$$

$$\sigma=\left(\frac{2m|E|}{a^2}\right)^{1/2},\quad k'=\left(\frac{2m|A|}{a^2}\right)^{1/2}.$$

The constants C and C' have to be appropriately chosen to meet the boundary conditions.

V. THE MORSE BARRIER ($A=0$)

In the case where $A=0$ (see Figs. 2 and 4), the basic equation (2.4) or the tilted equation (2.9) with $\theta=0$ takes the form

$$[(a^2/2m)(\Gamma_0+S)-B]|\Phi\rangle=0. \tag{5.1}$$

Although θ may be kept arbitrary, we can set $\theta=0$ without loss of generality. The continuous eigenvalue λ of the noncompact operator Γ_0+S is fixed by (5.1) to be

$$\lambda=2mB/a^2. \tag{5.2}$$

Again, as in the case of $A<0$, if $B>0$, the principal and supplementary series of representation correspond to the positive and negative energy continua, respectively. For $B<0$, there are no negative energy states.

In the ρ-representation, (5.1) can be written as

$$\left[\frac{d^2}{d\rho^2}-\frac{4(2\varphi+1)^2-1}{4\rho^2}+\frac{8mB}{a^2}\right]\Phi(\rho)=0, \tag{5.3}$$

by setting $k=1$ in (A20) and (A21). A solution of (5.3) is given in terms of cylindrical functions,

$$\Phi(\rho)=\rho^{1/2}Z_{2\varphi+1}(2k''\rho), \tag{5.4}$$

where $k''=(2mB/a^2)^{1/2}$. Thus, we obtain the positive energy solution

$$\Phi(x)=e^{-(1/4)ax}[CJ_{2i\sigma}(2k''e^{-(1/2)ax})+C'N_{2i\sigma}(2k''e^{-(1/2)ax})], \tag{5.5}$$

and the negative energy solution

$$\Phi(x)=e^{-(1/4)ax}[CJ_{2\sigma}(2k''e^{-(1/2)ax})+C'N_{2\sigma}(2k''e^{-1/2ax})], \tag{5.6}$$

where $\sigma=(2m|E|/a^2)^{1/2}$. Clearly, the constant $k''=(2mB/a^2)^{1/2}$ is real if $B>0$ and imaginary if $B<0$. For $B<0$, the negative energy solution (5.6) must vanish in order to remain finite for the entire range of x. Thus, there is no negative energy continuum.

VI. FINAL REMARK

We see that for different ranges of the parameters A, B, and energy E one has to use different representations of SO(2,1) and to make different SO(2,1) generators diagonal (see Figs. 1–4). It is known that also in the H atom we use two replicas of the SO(4,2) representations, one for the discrete spectrum and one for the continuous spectrum in which different generators are diagonalized. For the one-dimensional Morse oscillator the full extent of the SO(4,2) does not come into play because we transformed the system (2.1) into the rest frame. It will come in when the three-dimensional and moving system is considered.

APPENDIX: REALIZATIONS OF THE GENERATORS OF SO(2,1)⊂SO(4,2)

Here we briefly describe some properties of the most degenerate representation of SO(4,2) and provide some physical realizations of the SO(2,1) generators relevant to the discussions in the text.[7,12,13]

The 15 operators L_{AB} of the dynamical group SO(4,2) form the Lie algebra so(4,2)

$$[L_{AB},L_{CD}]=i(g_{AD}L_{BC}-g_{AC}L_{BD}+g_{BC}L_{AD}-g_{BD}L_{AC}), \tag{A1}$$

where $A,B=1,2,\ldots,6$, $g_{11}=g_{22}=g_{33}=g_{44}=-g_{55}=-g_{66}=-1$ and $g_{AB}=0$ for $A\neq B$. The algebra so(4,2) contains the angular momentum vector $\mathbf{L}(L_{23},L_{31},L_{12})$, the Lenz–Runge vector $\mathbf{A}(L_{14},L_{24},L_{34})$, the Lorentz boost vector $\mathbf{M}(L_{15},L_{25},L_{35})$, the current vector $\mathbf{\Gamma}(L_{16},L_{26},L_{36})$, and the remaining three operators $\Gamma_0=L_{56}$, $S=L_{46}$, and $T=L_{45}$, which form the subalgebra so(2,1). Here, $\Gamma_\mu(\mathbf{\Gamma},\Gamma_0)$ is a Lorentz four-vector and T is a Lorentz scalar.

In general, the basis of a unitary irreducible representation is labeled by the eigenvalues of the nine invariant operators. However, we confine ourselves to the most degenerate unitary irreducible representation whose basis is characterized by the eigenvalues of only four operators. To realize the generators of SO(4,2), six real variables are needed in general, whereas only four are sufficient in the degenerate representation. If polar coordinate variables (r,θ,ϕ,ψ) are used, the radial variable r and the other angular variables can be separated in realizing the operators. Let us set $\langle r,\theta,\phi,\psi|\Phi\rangle=\Phi(r)D_m{}^l{}_{l_0}(\theta,\phi,\psi)$ with

$$L^2D_m{}^l{}_{l_0}(\theta,\phi,\psi)=l(l+1)D_m{}^l{}_{l_0}(\theta,\phi,\psi), \tag{A2}$$

$$L_{12}D_m{}^l{}_{l_0}(\theta,\phi,\psi)=mD_m{}^l{}_{l_0}(\theta,\phi,\psi), \tag{A3}$$

$$L_0D_m{}^l{}_{l_0}(\theta,\phi,\psi)=l_0D_m{}^l{}_{l_0}(\theta,\phi,\psi), \tag{A4}$$

where

$$[L_0,L_{AB}]=0;\quad l_0=0,1,2,\ldots,\text{ or }\tfrac{1}{2},\tfrac{3}{2},\tfrac{5}{2},\ldots;$$

$$l=|l_0|,|l_0|+1,|l_0|+2,\ldots,N;$$

and

$$m = -l, -l+1,...,l-1,l.$$

The Morse system in one dimension may be characterized by $l=0$ and hence described by the representation with $l_0=0$. If the compact operator Γ_0 is diagonalized with the eigenvalue n, then the upper limit N of l is $n-1$. On the other hand, if a noncompact operator of so(2,1) is diagonalized with a continuous eigenvalue, then N is infinity. Now we are able to represent the generators of SO(2,1) in terms of the radial variable r alone and treat them as operators acting on $\Phi(r)$. For example,

$$\Gamma_0\Phi(r) = \frac{1}{2}r\left[-\frac{d^2}{dr^2} - \frac{2}{r}\frac{d}{dr} + \frac{l(l+1)}{r^2} + 1\right]\Phi(r), \tag{A5}$$

$$S\Phi(r) = \frac{1}{2}r\left[-\frac{d^2}{dr^2} - \frac{2}{r}\frac{d}{dr} + \frac{l(l+1)}{r^2} - 1\right]\Phi(r), \tag{A6}$$

$$T\Phi(r) = -i\left[r\frac{d}{dr} + 1\right]\Phi(r). \tag{A7}$$

It is easy to show from (A5)–(A7) that

$$Q^2\Phi(r) = l(l+1)\Phi(r), \tag{A8}$$

where $Q^2 = \Gamma_0^2 - S^2 - T^2$ is the Casimir operator of SO(2,1). Clearly the eigenvalue $\varphi(\varphi+1)$ of Q^2 coincides with that of L^2,

$$\varphi(\varphi+1) = l(l+1). \tag{A9}$$

Therefore we can write (A5), (A6), and (A7) as

$$\Gamma_0^{(r)}\Phi(r) = \frac{1}{2}r\left[-\frac{d^2}{dr^2} - \frac{2}{r}\frac{d}{dr} + \frac{\varphi(\varphi+1)}{r^2} + 1\right]\Phi(r), \tag{A10}$$

$$S^{(r)}\Phi(r) = \frac{1}{2}r\left[-\frac{d^2}{dr^2} - \frac{2}{r}\frac{d}{dr} + \frac{\varphi(\varphi+1)}{r^2} - 1\right]\Phi(r), \tag{A11}$$

$$T^{(r)}\Phi(r) = -i\left[r\frac{d}{dr} + 1\right]\Phi(r), \tag{A12}$$

satisfying

$$Q^2\Phi(r) = \varphi(\varphi+1)\Phi(r). \tag{A13}$$

The realization of the generators is by no means unique. We can change variables and operands successively to obtain various representations. To this end, in each step, we construct the generators $L^{(y)} \in \mathrm{so}(2,1)$ in such a way that $L^{(x)}\Phi(x) = f(y)L^{(y)}\Phi(y)$ when $\Phi(x) = f^{(y)}\Phi(y)$ under the mapping $x \to y$. First, let $\Phi(r) = R^{-1}\Phi(R)$ with $R = 2r$ $(0 < R < \infty)$. Then, we have in the R-representation,

$$\Gamma_0^{(R)}\Phi(R) = \left[-R\frac{d^2}{dR^2} + \frac{\varphi(\varphi+1)}{R} + \frac{1}{4}R\right]\Phi(R), \tag{A14}$$

$$S^{(R)}\Phi(R) = \left[-R\frac{d^2}{dR^2} + \frac{\varphi(\varphi+1)}{R} - \frac{1}{4}R\right]\Phi(R), \tag{A15}$$

$$T^{(R)}\Phi(R) = -iR\frac{d}{dR}\Phi(R). \tag{A16}$$

Next, let $\Phi(R) = R^{1/2}\Phi(\xi)$, with $R = 2ke^{-\xi}$ $(-\infty < \xi < \infty)$. Then, in the ξ-representation,

$$\Gamma_0^{(\xi)}\Phi(\xi) = \frac{1}{2k}e^{\xi}\left[-\frac{d^2}{d\xi^2} + \varphi(\varphi+1) + \frac{1}{4} + k^2e^{-2\xi}\right]\Phi(\xi), \tag{A17}$$

$$S^{(\xi)}\Phi(\xi) = \frac{1}{2k}e^{\xi}\left[-\frac{d^2}{d\xi^2} + \varphi(\varphi+1) + \frac{1}{4} - k^2e^{-2\xi}\right]\Phi(\xi), \tag{A18}$$

$$T^{(\xi)}\Phi(\xi) = i\left[\frac{d}{d\xi} - \frac{1}{2}\right]\Phi(\xi). \tag{A19}$$

Furthermore, letting $\Phi(\xi) = \rho^{1/2}\Phi(\rho)$ with $\rho = e^{-1/2\xi}$ $(0 < \rho < \infty)$, we have the ρ-representation

$$\Gamma_0^{(\rho)}\Phi(\rho) = \frac{1}{8k}\left[-\frac{d^2}{d\rho^2} + \frac{(4\varphi+1)(4\varphi+3)}{4\rho^2} + 4k^2\rho^2\right]\Phi(\rho), \tag{A20}$$

$$S^{(\rho)}\Phi(\rho) = \frac{1}{8k}\left[-\frac{d^2}{d\rho^2} + \frac{(4\varphi+1)(4\varphi+3)}{4\rho^2} - 4k^2\rho^2\right]\Phi(\rho), \tag{A21}$$

$$T^{(\rho)}\Phi(\rho) = -\frac{1}{4}i\left[2\rho\frac{d}{d\rho} + 1\right]\Phi(\rho). \tag{A22}$$

Suppose the SO(4,2) symmetry is broken so that the eigenvalue of Q^2 is shifted by

$$\varphi(\varphi+1) = l(l+1) - \mu^2. \tag{A23}$$

Certainly, Γ_0, S, and T, realized with this shifted eigenvalue (A23), do not satisfy the so(4,2) algebra (A1). Nevertheless, they still form an so(2,1) algebra. Therefore the realizations of the SO(2,1) generators with φ, (A10)–(A22) are useful in dealing with the case of symmetry breaking. The parameter μ^2 in (A23) indicates the degree of the symmetry breaking of SO(4,2) to SO(3) ⊗ SO(2,1).

There are four types of unitary irreducible representations of so(2,1): (i) positive discrete series $D^+(\varphi)$,

$$n = -\varphi, -\varphi+1, -\varphi+2,..., \quad \varphi \text{ real and negative;} \tag{A24}$$

(ii) negative discrete series $D^-(\varphi)$,

$$n = \varphi, \varphi-1, \varphi-2,..., \quad \varphi \text{ real and negative;} \tag{A25}$$

(iii) supplementary continuous series $D_s(\varphi)$,

$$-1 + |E_0| \leqslant \varphi \leqslant -|E_0|, \quad |E_0| < \tfrac{1}{2}, \quad n = E_0, E_0 \pm 1, E_0 \pm 2,...; \tag{A26}$$

and (iv) principal continuous series $D_p(\varphi)$,

$$\varphi = -\tfrac{1}{2} + i\sigma, \quad \sigma \text{ real.} \tag{A27}$$

In the above representations, n is the eigenvalue of Γ_0. In the broken symmetry case (A23), the supplementary series need to be modified as

$$(\varphi + \tfrac{1}{2})^2 \leqslant (\tfrac{1}{2} - |E_0|^2)^2 - \mu^2. \tag{A28}$$

[1]R. D. Levine and C. E. Wulfman, Chem. Phys. Lett. **60**, 372 (1979).
[2]M. Berrondo and A. Palma, J. Phys. A **13**, 773 (1980), and references therein.
[3]O. S. van Roosmalen, thesis, Rijksuniversiteit Gronigen, 1982.
[4]P. Y. Cai, A. Inomata, and R. Wilson, Phys. Lett. A **96**, 117 (1983).
[5]Y. Alhassid, F. Gürsey, and F. Iachello, Ann. Phys. (NY) **148**, 346 (1983).
[6]C. C. Sun and Z. H. Zeng, J. Math. Phys. **24**, 1482 (1983).
[7]A. O. Barut, *Dynamical Groups and Generalized Symmetries in Quantum Theory* (Univ. Canterbury P., Christchurch, 1972).
[8]C. Fronsdal, Phys. Rev. **156**, 1653, 1665 (1967).
[9]B. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974).
[10]A. O. Barut, H. Beker, and D. Dibekci, J. Bogazici Univ. (Turkey) **8–9**, 11 (1981); A. O. Barut and H. Beker, Phys. Rev. Lett. **50**, 1560 (1983).
[11]P. M. Morse, Phys. Rev. **34**, 57 (1929); D. ter Haar, *ibid.* **70**, 222 (1946).
[12]E. T. Whittaker and G. N. Watson, *Modern Analysis* (Cambridge, U.P., London, 1965), 4th ed.
[13]H. Bucholz, *The Confluent Hypergeometric Function* (Springer, New York, 1969).
[14]D. Peak and A. Inomata, J. Math. Phys. **10**, 1423 (1969); W. Langguth and A. Inomata, J. Math. Phys. **20**, 499 (1979).
[15]A. O. Barut and G. L. Bornzin, J. Math. Phys. **12**, 841 (1971).
[16]A. O. Barut, C. K. E. Schneider, and R. Wilson, J. Math. Phys. **20**, 2244 (1979).

On a class of 6j coefficients with one multiplicity index for groups SP(2N), SO(2N), and SO(2N+1)

Marcin Cerkaski
Department of Theoretical Physics, Institute of Nuclear Physics, ul. Radzikowskiego 152, 31-342 Kraków, Poland and Joint Institute for Nuclear Research, Dubna, USSR

(Received 17 June 1986; accepted for publication 1 October 1986)

An important class of 6j symbols for the groups SP(2N), SO(2N), and SO(2N + 1) with one nontrivial multiplicity index is investigated. An appropriate choice of a basis in the multiplicity space is made and the so-called canonical form for 6j symbols is obtained. Their expressions depending on the roots of an N th-order equation and explicit expressions for some simple class of representations are obtained.

I. INTRODUCTION

The problem of finding matrix elements of generators for simple classical algebras B_{N+1}, C_{N+1}, and D_{N+1} with reduction on subalgebras B_N, C_N, and D_N is important for applications of mathematical methods. The branching rules for the above reduction are known.[1,2] The appropriate sets of the so-called missing label operators are found.[3,4] Applying the generalized Wigner–Eckart theorem to commutation relations between the generators, one obtains equations for reduced matrix elements. In this approach the 6j symbols for chosen subgroups occurring in the above system should be calculated. An important subclass of these 6j symbols is found in the present work. In Sec. II the general remarks on our class of 6j symbols are made, and the symmetry properties of 3j symbols entering into 6j are discussed. In Sec. III we show that 6j symbols may be found from the unitarity properties if some reasonable basis in the multiplicity space is chosen. In Sec. IV their general expression depending on the roots of N th-order equations is obtained. In Sec. V singular cases connected with reduction of the dimension of the multiplicity space are discussed. Also, we have found explicit expressions for a simple class of representations (see Sec. VI).

The same class of 6j symbols for the SU(N) group may be considered, and we hope that our approach may be extended also to this group.

II. GENERAL REMARKS ON THE INVESTIGATED 6j SYMBOLS

The class of 6j symbols to be treated in this paper shortly denoted by $\phi(\Omega)_{\alpha\alpha}$ is of the form

$$\phi(\Omega)_{\alpha\alpha} = \phi(\Omega,\Omega'_\alpha)_a = \begin{Bmatrix} 1^* & 1 & \Lambda \\ \Omega & \Omega & \Omega'_\alpha \end{Bmatrix}_{..A}, \tag{2.1}$$

where the symbols Ω's ($\Omega = (\Omega_1,\Omega_2,...,\Omega_N)$) label a unitary irreducible representation (UIR) belonging to groups SP(2N), SO(2N), or SO(2N + 1) and Ω^* is the complex conjugate representation. The representation $(\Omega_1,\Omega_2,...,\Omega_p,0,0,...,0)$ is written shortly $(\Omega_1,\Omega_2,...,\Omega_p)$ and we sometimes omit the parentheses in our notation if the meaning of the label is obvious, as in example (2.1). The Λ representation is one of (0), (11), (2), and the A is the multiplicity index, so, in the Kronecker product of representations $\Omega\times\Lambda$, the Ω representation may be found more than once. Other multiplicity indices do not occur in our symbols, so we use dots in their place. Here, we use the same definition and the notation of 6j symbols as in Ref. 5. In expression (2.1) index in the left-hand side is fixed from the Ω'_α representation in the following way:

$$\Omega'_\alpha = (\Omega_1,\Omega_2,...,\Omega_{|\alpha|-1},\Omega_{|\alpha|}+\epsilon_\alpha,\Omega_{|\alpha|+1},...,\Omega_N), \tag{2.2a}$$

where $\epsilon_\alpha = 1(-1)$ if the $\alpha>0$ ($\alpha<0$). For the case SO(2N) one more Ω'_0 representation should be added,

$$\Omega'_0 = (\Omega_1,\Omega_2,...,\Omega_N). \tag{2.2b}$$

Hence the range α is $\pm 1, \pm 2,..., \pm N, (0)$, respectively, for cases SP(2N), SO(2N), (SO(2N + 1)) and the reduction of α's must be done if some Ω'_α is not an allowed label for the studied group [see (2.5a)–(2.5e) and text below]. Let us introduce the notation

$$\Lambda_+ = (1,1), \quad \Lambda_- = (2,0), \quad \Lambda_0 = (0), \tag{2.3a}$$

for groups SO(2N) and SO(2N + 1) or

$$\Lambda_+ = (2,0), \quad \Lambda_- = (1,1), \quad \Lambda_0 = (0), \tag{2.3b}$$

for the SP(2N) group [here Λ_+ is an adjoint representation for all cases]. The case of the SO(4) group is rather peculiar. In this case the three-dimensional representations (1,1) and (1, − 1) are used instead of the $\Lambda_+ = (1,1) + (1,-1)$ and some modifications of expressions to be obtained here should be done.

The dimensions of the multiplicity space referred to triads $\{\Omega,\Omega^*,\Lambda_\pm\}$ are denoted by $d_\pm$. All symbols of our class may be written if one uses index a instead of $\Lambda_\pm A_\pm, \Lambda_0$, where

$$a = A_+, \quad \text{for } \Lambda = \Lambda_+, \tag{2.4a}$$

$$a = -1, \quad \text{for } \Lambda = \Lambda_0, \tag{2.4b}$$

$$a = -A_- - 1, \quad \text{for } \Lambda = \Lambda_-, \tag{2.4c}$$

for cases SP(2N) and SO(2N). For the SO(2N + 1) group (2.4a) and (2.4b) coincide, but if $\Omega_N \neq \frac{1}{2}$ (see Sec. V) then (2.4c) should be changed: $a = -2, -3,..., -d_-, 0$.

In what follows we will understand $\phi(\Omega)_{\alpha\alpha}$ as a square matrix, the dimension of which depends on Ω. For that reason we introduce δ_α symbols

$$\delta_\alpha = \begin{cases} 1, & \text{if } \Omega_\alpha > \Omega_{\alpha+1}, \\ 0, & \text{if } \Omega_\alpha = \Omega_{\alpha+1}, \end{cases} \tag{2.5a}$$

 0022-2488/87/030612-06$02.50

where $\alpha = 1,2,...,N-1$ [or $\alpha = 1,2,...,N-2$ for the SO(2N) group]

$$\delta_N = \begin{cases} 1, & \text{if } \Omega_N > 0, \\ 0, & \text{if } \Omega_N = 0, \end{cases} \tag{2.5b}$$

for the SP(2N) and SO(2N+1) groups, and for the SO(2N) group we have

$$\delta_{N-1} = \begin{cases} 1, & \text{if } \Omega_{N-1} > |\Omega_N|, \\ 0, & \text{if } \Omega_{N-1} = \pm\Omega_N, \end{cases} \tag{2.5c}$$

$$\delta_N = \begin{cases} 1, & \text{if } \Omega_{N-1} > 0, \\ 0, & \text{if } \Omega_{N-1} = 0. \end{cases} \tag{2.5d}$$

For the SO(2N+1) group we use

$$\delta_0 = \begin{cases} 1, & \text{if } \Omega_N \neq 0, \tfrac{1}{2}, \\ 0, & \text{otherwise}. \end{cases} \tag{2.5e}$$

Now we may introduce Δ_α symbols indicating rows of $\phi(\Omega)_{\alpha a}$ that are equal to zero and should be removed,

$$\Delta_1 = 1, \quad \Delta_{-1} = \delta_1, \quad \Delta_0 = \delta_N - \delta_0,$$
$$\Delta_\alpha = \delta_{\alpha-1}, \quad \Delta_{-\alpha} = \delta_\alpha, \quad \text{for } \alpha = 2,3,...,N.$$

The dimensions d_+, d_- introduced above are

$$d_+ = \sum_{\alpha=1}^{N} \delta_\alpha, \tag{2.6a}$$

$$d_- = \sum_{\alpha=1}^{N-1} \delta_\alpha + \delta_0. \tag{2.6b}$$

These formulas will be proved in Sec. V.

The unitarity conditon for our 6j symbols may be written

$$\sum_a D_\alpha D_a \phi(\Omega)^*_{\alpha a} \phi(\Omega)_{\beta a} = \delta_{\alpha\beta}\Delta_\alpha, \tag{2.7a}$$

$$\sum_\alpha D_\alpha D_a \phi(\Omega)^*_{\alpha a} \phi(\Omega)_{\alpha b} = \delta_{ab}, \tag{2.7b}$$

$$D_a = D_{\Lambda_a} = \begin{cases} \tfrac{1}{2}(2f+1-\eta)(2f+1), & \text{if } a \geqslant 1, \quad (2.8a) \\ 1, & \text{if } a = -1, \quad (2.8b) \\ \tfrac{1}{2}(2f+1-3\eta)(2f+1), & \text{if } a = 0 \text{ or } a \leqslant -2, \quad (2.8c) \end{cases}$$

where D_M is the dimension of UIR M and $D_{\Omega'_\alpha}$ is denoted by D_α. We have

$$f = \begin{cases} N, & \text{for SP}(2N), \\ N-1, & \text{for SO}(2N), \\ N-\tfrac{1}{2}, & \text{for SO}(2N+1), \end{cases}$$

$$\eta = \begin{cases} 1, & \text{for SP}(2N), \\ -1, & \text{for SO}(2N), \text{SO}(2N+1). \end{cases}$$

The following three classes of 3j symbols occur in 6j:

(a) $(\Omega,0,\Omega^*)_{m0m'}$,

(b) $(\Omega,\Lambda_\pm,\Omega^*)_{A_\pm m_1 m_2 m_3}$,

(c) $(\Omega,1,\Omega'^*_\alpha)_{m_1 m_2 m_3}$.

The symbols of class (a) are proportional to the elements of the metrics tensor

$$(\Omega)_{mm'} = ((\Omega)^{mm'})^* = (D_\Omega)^{1/2}(\Omega,0,\Omega^*)_{m0m'},$$

that is used to raise and lower m index in the usual way,

$$F(\Omega)^m = (\Omega)^{mm'}F(\Omega^*)_{m'},$$
$$F(\Omega)_m = F(\Omega^*)^{m'}(\Omega^*)_{m'm}.$$

All representations for groups SP(2N), SO(2N+1), and SO(4N) are self-complex-adjoint, but for the group SO(4N+2) we have

$$(\Omega_1,\Omega_2,...,\Omega_{N-1},\Omega_N)^* = (\Omega_1,\Omega_2,...,\Omega_{N-1},-\Omega_N).$$

It is possible for all classes to use real 3j coefficients that have very simple symmetry properties:

$$(\Omega^{P(1)},\Omega^{P(2)},\Omega^{P(3)})_{Am_{P(1)}m_{P(2)}m_{P(3)}} = (-1)^{S_P([\Omega^1]+[\Omega^2]+[\Omega^3])}(\Omega^1,\Omega^2,\Omega^3)_{Am_1m_2m_3}. \tag{2.9}$$

Here $S_P = 1$ for odd and $S_P = 0$ for even permutations P

$$[\Omega] = \frac{1}{2}\sum (-1)^{i+1}\Omega_i, \tag{2.10a}$$

for the SP(2N) group and

$$[\Omega] = |\Omega_2|, \tag{2.10b}$$

for the SO(2N) and SO(2N+1) groups ($N \geqslant 2$). For coefficients of class (c), relations (2.9) may be obtained if appropriate phase conventions are chosen. The same is true for classes (a) and (b) if Ω is not equivalent to Ω^* [the SO(4N+2) case if $\Omega_N \neq 0$], but if $\Omega \simeq \Omega^*$ the requirements (2.9) are not trivial[6] and should be proved. We do not touch upon this question in this paper [here the plethysm relations[7] $(\Omega)\otimes\{\rho\}$, for $\{\rho\} = \{2\},\{11\}$ must be investigated]. We obtain the following symmetry relations for 6j symbols if all entering 3j symbols fulfill conditions (2.9):

$$\begin{Bmatrix} \Omega_1 & \Omega_2 & \Omega_3 \\ \omega_1 & \omega_2 & \omega_3 \end{Bmatrix}_{A_1A_2A_3A_0} = \begin{Bmatrix} \Omega_2 & \Omega_1 & \Omega_3 \\ \omega_2^* & \omega_1^* & \omega_3^* \end{Bmatrix}_{A_2A_1A_3A_0} = \begin{Bmatrix} \Omega_2 & \Omega_3 & \Omega_1 \\ \omega_2 & \omega_3 & \omega_1 \end{Bmatrix}_{A_2A_3A_1A_0} = \cdots, \tag{2.11a}$$

$$\begin{Bmatrix} \Omega_1 & \Omega_2 & \Omega_3 \\ \omega_1 & \omega_2 & \omega_3 \end{Bmatrix}_{A_1A_2A_3A_0} = \begin{Bmatrix} \Omega_1^* & \omega_2 & \omega_3^* \\ \omega_1^* & \Omega_2 & \Omega_3^* \end{Bmatrix}_{A_0A_3A_2A_1} = \begin{Bmatrix} \omega_1 & \omega_2^* & \Omega_3^* \\ \Omega_1 & \Omega_2^* & \omega_3 \end{Bmatrix}_{A_2A_1A_0A_3} = \cdots, \tag{2.11b}$$

$$\begin{Bmatrix} \Omega_1 & \Omega_2 & \Omega_3 \\ \omega_1 & \omega_2 & \omega_3 \end{Bmatrix}^*_{A_1A_2A_3A_0} = \begin{Bmatrix} \Omega_1^* & \Omega_2^* & \Omega_3^* \\ \omega_1^* & \omega_2^* & \omega_3^* \end{Bmatrix}^{A_1A_2A_3A_0}$$
$$= G(\Omega_1^*\omega_2\omega_3^*)^{A_1A_1'} \times G(\omega_1^*\Omega_2^*\omega_3)^{A_2A_2'} G(\omega_1\omega_2^*\Omega_3^*)^{A_3A_3'} \times G(\Omega_1\Omega_2\Omega_3)^{A_0A_0'} \begin{Bmatrix} \Omega_1^* & \Omega_2^* & \Omega_3^* \\ \omega_1^* & \omega_2^* & \omega_3^* \end{Bmatrix}_{A_1'A_2'A_3'A_0'}. \tag{2.11c}$$

The multiplicity metric tensor $G(\Omega,\Omega^*,\Lambda^*_\pm)$ [see (2.11c) and (2.1)] is unitary,[5] and symmetric for the above case. Here it does not depend on the order $\{\Omega,\Omega^*,\Lambda^*_\pm\}$. The one-dimensional metric tensors corresponding to the triads $\{1^*,1,\Lambda_\pm\}$ and $\{\Omega^*,1,\Omega'_\alpha\}$ are chosen to be equal to 1, hence, we may omit the three dots [see (2.1)] connected with ordinary $3j$ symbols. Taking into account that $\Lambda^*_\pm$ are equivalent to $\Lambda_\pm$ and (2.11a)–(2.11c), we may rewrite the unitarity conditon (2.7b) in the form

$$\sum_\alpha D_\alpha D_a \phi(\Omega)_{a\alpha}\phi(\Omega)_{\alpha b} = G(\Omega,\Omega^*,\Lambda_a)_{ab} . \tag{2.12}$$

Matrix $\phi(\Omega)$ to be obtained in Sec. IV is real, hence comparing (2.7b) and (2.12) we find that the tensor G is a unit tensor

$$G(\Omega,\Omega^*,\Lambda_a)_{ab} = \delta_{ab} . \tag{2.13}$$

In the next section the definition of the tensor operators (2.14) $T(\Gamma)_m$ and the generalized Wigner–Eckart theorem[8] will be used (2.15):

$$[A(\Lambda_+)_{m_1}, T(\Gamma)_{m_2}]$$

$$= \sum_{m'} \langle \Gamma m' | A(\Lambda_+)_{m_1} | \Gamma m_2 \rangle T(\Gamma)_{m'} , \tag{2.14}$$

$$\langle \Omega_1 m_1 | T(\Omega_2)_{m_2} | \Omega_3 m_3 \rangle$$

$$= \sum_{A m'_1} (\Omega_1)^{m_1 m'_1} (\Omega_1^* \Omega_2 \Omega_3)_{A m_1 m_2 m_3}$$

$$\times (\Omega_1 || T(\Omega_2) || \Omega_3)^A , \tag{2.15}$$

where $A(\Lambda_+)_m$ are generators of the investigated group. Expressions (2.15) are manifestly invariant under any unitary transformations acting in the multiplicity space, but for $T(\Omega)_m = A(\Lambda_+)_m$ it is more reasonable to do the choice

$$(\Omega || A(\Lambda_+) || \Omega)^A = \delta^A_1 (\tfrac{1}{2} C_\Omega D_\Omega)^{1/2} . \tag{2.16}$$

This condition is invariant with respect to $U(d_+ - 1)$ transformations which do not touch the index $A = 1$. These transformations will be used in the next section. The function C_Ω is an eigenvalue of the second-order Casimir operator, and we have

$$C_\Omega = 2 \times \sum_{\alpha=1}^{N} (\omega_\alpha^2 - g_\alpha^2) , \tag{2.17}$$

where

$$\omega_\alpha = \Omega_\alpha + g_\alpha , \tag{2.18a}$$

$$g_\alpha = f + 1 - \alpha . \tag{2.18b}$$

III. BASIC ASSUMPTIONS OF THE PRESENTED APPROACH

In this section we find one ($a = 1$) column for the $\phi(\Omega)$ matrix and some simple relations. This permits us to calculate the whole matrix.

If we take definition (2.14) for $\Gamma = (1)$ between states $\langle \Omega^* m |$ and $|\Omega^* m'\rangle$ and we make assumption (2.16), after rather simple calculations we obtain

$$(C_\Omega D_\Omega)^{1/2} \begin{Bmatrix} 1^* & 1 & \Lambda_+ \\ \Omega & \Omega & \Omega'_\alpha \end{Bmatrix}_{\cdot\cdot 1\cdot}$$

$$+ (C_{\Omega'_\alpha} D_{\Omega'_\alpha})^{1/2} \begin{Bmatrix} 1 & 1^* & \Lambda_+ \\ \Omega'_\alpha & \Omega'_\alpha & \Omega \end{Bmatrix}_{\cdot\cdot 1\cdot}$$

$$= P_1 \Delta_\alpha \left(\frac{C_{(1)}}{D_{\Lambda_+}} \right)^{1/2} . \tag{3.1}$$

Here (2.9), (2.15), and a simple tensor sum relation

$$\sum_m F(\Omega)^m G(\Omega)_m$$

$$= (-1)^{[\Omega]} \sum_m F(\Omega^*)_m G(\Omega^*)^m$$

has been applied, too. Taking into consideration that representations (1) and (1*) are equivalent for investigated groups, one finds a very similar expression to the well-known SO(3) formula

$$\phi(\Omega)_{\alpha 1} = P_1 N_1 \psi(\Omega)_{\alpha 1} , \tag{3.2a}$$

$$P_a = (-1)^{[\Omega] + [\Omega'_\alpha] + [\Lambda_a] + [(1)]} , \tag{3.2b}$$

$$N_1 = 2 \times (C_\Omega D_\Omega C_{(1)} D_{(1)})^{-1/2} = (C_\Omega D_\Omega D_{\Lambda_+})^{-1/2} , \tag{3.2c}$$

$$\psi(\Omega)_{\alpha 1} = \Delta_\alpha / 4 (C_\Omega + C_{(1)} - C_{\Omega'_\alpha}) = \Delta_\alpha (f - \omega_\alpha) , \tag{3.2d}$$

where the following generalization of (2.18a) and (2.18b) is done:

$$\omega_\alpha = \epsilon_\alpha \omega_{|\alpha|} , \tag{3.3a}$$

and for the SO($2N + 1$) case the component ω_0 is included

$$\omega_0 = -\tfrac{1}{2} . \tag{3.3b}$$

In the same way we obtain other $6j$ coefficients

$$\begin{Bmatrix} \Lambda^*_\pm & \Lambda_\pm & \Lambda_+ \\ \Omega & \Omega & \Omega \end{Bmatrix}^A_{B1\cdot}$$

$$= (-1)^{2[\Omega] + [\Lambda_+] + [\Lambda_\pm]} \delta^A_B$$

$$\times (C_\Omega D_\Omega C_{\Lambda_\pm} D_{\Lambda_\pm})^{-1/2} \frac{C_{\Lambda_\pm}}{2} . \tag{3.4}$$

Let us introduce new symbols

$$H^a_b = (C_\Omega D_\Omega D_a D_b D_{\Lambda_+})^{1/2} F^a_b , \tag{3.5}$$

$$F^a_b = \sum_\alpha P_a P_b P_1 D_\alpha \phi(\Omega)^*_{a\alpha} \phi(\Omega)_{\alpha\Lambda} \phi(\Omega)_{\alpha b} , \tag{3.6}$$

$$\psi(\Omega)_{a\alpha} = P_a (C_\Omega D_\Omega D_a)^{1/2} \phi(\Omega)_{a\alpha} . \tag{3.7}$$

The symbols F^a_b are a simple extension to a nonsimple reducible group of the so-called second kind $9j$ symbols, and they may be expressed by two $6j$ symbols,

$$F^a_b = (-1)^{2[\Omega] + 2[(1)]} \begin{Bmatrix} \Lambda^*_a & \Lambda_b & \Lambda_+ \\ 1 & 1 & 1 \end{Bmatrix} \begin{Bmatrix} \Lambda^*_a & \Lambda_b & \Lambda_+ \\ \Omega & \Omega & \Omega \end{Bmatrix}^A_{B1} . \tag{3.6'}$$

The $3j$ symbols entering into the first $6j$ symbol are ordinary, hence we omit the multiplicity indices and the indices A (B) in the second one depend on the a (b) [(2.4a)–(2.4c) and text below].

Lemma 1: Eigenvalues of the Hermitian H matrix are the $\psi(\Omega)_{\alpha 1}$ coefficients.

Lemma 2: (a) The following equations for the $\psi(\Omega)_{aa}$ are satisfied:

$$\psi(\Omega)_{aa}U_{aa}=\psi(\Omega)_{\alpha 1}U_{\alpha 1}\,, \tag{3.8}$$

where $U_\alpha=(U_{\alpha 1},U_{\alpha 2},\ldots,U_{\alpha 2f+1-\eta})$ are eigenvectors of the H matrix or else the following is true.

(b) The H matrix has the reducible form and here Eq. (3.8) touches only the index a, α occurring at the same block.

Proposition 1: (a) If relations (2.16) are preserved and all δ_α symbols (2.5a)–(2.5e) are equal to 1, then the matrix H may be chosen in the so-called canonical form

$$H=H^D+G\,, \tag{3.9a}$$

where H^D is the diagonal matrix and G is of the form

$$G=\begin{bmatrix}0 & g\\ g^+ & 0\end{bmatrix}. \tag{3.9b}$$

Here all elements of the 0 matrix are equal to zero and the g is a square $N\times N$-dimensional matrix for the SP($2N$) and SO($2N$) case

$$g=\begin{bmatrix}X_1 & X_2 & X_3 & \cdots & & X_N\\ 0 & Q_1 & 0 & \cdots & & 0\\ \vdots & \vdots & \vdots & \cdots & & \vdots\\ 0 & 0 & 0 & \cdots & 0 & Q_N\end{bmatrix}, \tag{3.9c}$$

and g is an $N\times(N+1)$-dimensional matrix for the SO($2N+1$) case

$$g=\begin{bmatrix}X_1 & X_2 & X_3 & \cdots & & X_N & X_0\\ 0 & Q_2 & 0 & \cdots & & 0 & 0\\ \vdots & \vdots & \vdots & \cdots & & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 0 & Q_N & 0\end{bmatrix}. \tag{3.9d}$$

For the diagonal elements of H we have

$$H^a_a=\begin{cases}\frac{1}{2}(2f+1+\eta)\,, & \text{if } a\geqslant 1\,,\\ 0\,, & \text{if } a=-1\,,\\ \frac{1}{2}(2f+1-\eta)\,, & \text{if } a\leqslant -2 \text{ [or } a=0\\ & \text{for the SO}(2N+1) \text{ case]}\,.\end{cases} \tag{3.10}$$

Here the following order of index a is established: $1,2,\ldots,d_+,-1,-2,\ldots,-d_-,(0)$.

(b) If some δ_α symbols are equal to zero, then the elements of H outside the range of a [see (2.6a) and (2.6b), (2.4a)–(2.4c) and the text below] vanish.

The proof of Lemma 1 is very simple. Here we obtain the relation $\mathrm{Tr}(H^n)=\Sigma_\alpha(\psi_{\alpha 1})^n$ immediately from the unitarity relations (2.7a). Lemma 2 is derived from the relations

$$(\psi(\Omega)_{\alpha 1})^n=\sum_a (H^n)^a_1\psi(\Omega)_{aa}\,,$$

which are also obtained from the unitarity relations (2.7a).

Proof for Proposition 1: (a) The whole space related with index a is decomposed into three subspaces $\mathscr{A}_+,\mathscr{A}_0,\mathscr{A}_-$ connected with three multiplicity spaces appropriate to the triads $\{\Omega,\Omega^*,\Lambda_+\},\{\Omega,\Omega^*,\Lambda_0\},\{\Omega,\Omega^*,\Lambda_-\}$ [see (2.4a)–(2.4c)], and a more general multiplicity transformation is $U^+(d_+)\times U^0(1)\times U^-(d_-)$. In fact, we may use only the $U^+(d_+-1)\times U^0(1)\times U^-(d_-)$ transformation if relations (2.16) are preserved. Next, the diagonal blocks H^+_+, H^0_0, H^-_-, and H^+_0, H^0_+ are obtained by using (3.4)–(3.6b), (3.2a)–(3.2d), (2.17), (2.18a), and (2.18b). The off-diagonal blocks H^-_0, H^0_- are equal to zero from the general properties of $3j$ coefficient. The appropriate choice for transformations $U^+(d_+-1)$ and $U^-(d_-)$ may be done, and the off-diagonal block $H^\pm$ according to (3.9c) or (3.9d) may be obtained. The proof of point b will be put off to Sec. V.

IV. THE OFF-DIAGONAL ELEMENTS OF THE H MATRIX AND THE SOLUTION FOR THE $\phi(\Omega)$ MATRIX

The dispersion equations for the H matrix may be written in the form

$$-\frac{\Pi_{\omega_\alpha\in[\omega]}[2(\omega-\omega_\alpha)]}{\Pi_{q_b\in[q]}[2\omega+1-q_b]}=-1+\sum_{a=1}^{N(0)}\frac{4|X_a|^2}{(2\omega+1-q_a)(2\omega+1-q_{-a})}\,, \tag{4.1}$$

where the $(2f+1-\eta)$-dimensional set

$$[\omega]=[\omega_1,\omega_2,\ldots,\omega_N,\omega_{-1},\omega_{-2},\ldots,\omega_{-N},(\omega_0)]$$

is defined by (2.18a) and (2.18b), (3.3a) and (3.3b) and we have

$$q_1=2f+1\,, \tag{4.2a}$$

$$q_{-1}=-\eta\,, \tag{4.2b}$$

$$q_a=-q_{-a}=2(Q_a^2+\tfrac{1}{4})^{1/2}\cdot \text{ for } a=2,3,\ldots,N, \tag{4.2c}$$

$$q_0=-1\,. \tag{4.2d}$$

Here $q_0(q_{-0}\equiv q_{-1}=1)$ is added only for the SO($2N+1$). Comparing the residues for different poles of the function in the left- and right-hand sides of Eqs. (4.1), we receive two different expressions for all terms $X_a=X(q_a)$, $(X_a\equiv X_{-a}=X(q_{-a}))$;

$$|X_a|^2=-\frac{1}{4}\prod_{\omega_\alpha\in[\omega]}[2\omega_\alpha+1-q_a]\times\left(\prod_{\substack{q_b\in[q]\\ q_b\neq q_a,q_{-a}}}[q_b-q_a]\right)^{-1}, \tag{4.3a}$$

except the case SO($2N+1$) when for $a=-1$ we have

$$|X_{-1}|^2-f|X_0|^2=\frac{1}{8}\frac{\Pi_{\alpha=1}^N(2\omega_\alpha)^2}{\Pi_{b=2}^N(q_b^2-1)}. \tag{4.3b}$$

Similar modifications of (4.3a) should be done for the singular cases, i.e., when q_a (for $a=\pm 2,\pm 3,\ldots,\pm N$) takes values q_1, q_{-1}, or q_0 or when they are equal to each other. A more simple formula for X_1 is derived if (2.17), (2.18a), (3.2a)–(3.2d), and (3.4)–(3.6b) are used:

$$X_1=\left\{\frac{2}{(2f+1-\eta)}\sum_{\alpha=1}^N(\omega_\alpha^2-g_\alpha^2)\right\}^{1/2}. \tag{4.3c}$$

Next, comparing X_a with X_{-a} for $a = 2,3,...,N$ we obtain the following equation for q:

$$1 = \prod_{\alpha=1}^{N} \left[\frac{(2\omega_\alpha + 1 - q)(2\omega_{-\alpha} + 1 - q)}{(2\omega_\alpha + 1 + q)(2\omega_{-\alpha} + 1 + q)}\right] \times \begin{cases} (q_1 + q)(q_{-1} + q)/(q_1 - q)(q_{-1} - q)\,, \\ -(q_1 + q)/(q_1 - q)\,, \end{cases} \tag{4.4}$$

appropriate for the cases SP($2N$), SO($2N$) (up case), or SO($2N+1$) (down case). For the down case we get $2N-2$ roots $q_2 = -q_{-2},...,q_N = -q_{-N}$, likewise as for the up case when one unphysical root $q = 0$ should be removed. If $q_a = -q_{-a}$, $a = 2,3,...,N$, are found, then all elements of the H are known, hence the components of the vectors U_α may be calculated immediately and from (3.2a)–(3.2d), (3.7), (3.8), and (2.10a) and (2.10b) we obtain

$$\phi(\Omega)_{\alpha-1} = P_{-1}(D_\Omega(2f + 1 - \eta))^{-1/2}, \tag{4.5a}$$

$$\phi(\Omega)_{\alpha a} = \eta 4 Q_a \{(2\omega_\alpha + 1)^2 - q_a^2\}^{-1} X_a \phi(\Omega)_{\alpha 1}\,, \tag{4.5b}$$

$$\phi(\Omega)_{\alpha-a} = 2(2\omega_\alpha + 1 + \eta)\{(2\omega_\alpha + 1)^2 - q_a^2\}^{-1} \times (D_{\Lambda_+}/D_{\Lambda_-})^{1/2} X_a \phi(\Omega)_{\alpha 1}\,, \tag{4.5c}$$

$$\phi(\Omega)_{\alpha 0} = \{\omega_\alpha + 1\}^{-1} X_0 (D_{\Lambda_+}/D_{\Lambda_-})^{1/2} \phi(\Omega)_{\alpha 1}\,, \tag{4.5d}$$

where the last component $\phi(\Omega)_{\alpha 0}$ only for the SO($2N+1$) case is included, and $Q_a = \frac{1}{2}(q_a^2 - 1)^{1/2}$ for $a = 2,3,...,N$ [see (4.2c)]. The phase for factors X_a ($a = 2,3,...,N$) may be chosen [see (3.9c),and (3.9d), and proof of Proposition 1] so we may let $X_a = (|X_a|^2)^{1/2}$.

Proposition 2: All roots of Eq. (5.3) are real and if $2\omega_p + 1 \geqslant q_1 \geqslant 2\omega_{p+1} + 1$ then the positive roots values are bounded by inequalities

$$2\omega_{a-1} - 1 \geqslant q_a \geqslant 2\omega_a + 1\,, \quad \text{for } a = 2,3,...,p\,,$$

$$2\omega_b + 1 \geqslant q_b \geqslant 2\omega_b - 1\,, \quad \text{for } b = p+1, p+2,...,N\,,$$

and q_b tends to $2\omega_b - 1$ ($q_{-b} \to 2\omega_{-b} + 1$) if $2\omega_{b-1} - 1$ tends to $2\omega_b + 1$.

From Proposition 2 follows that matrix elements $\phi(\Omega)_{\alpha a}$ are real and that right-hand sides of the relations (4.3a) are positive. The above conclusions also remain true for singular cases.

V. SINGULAR CASE FOR *6j* SYMBOLS

In this section we investigate all cases when H and $\phi(\Omega)$ matrices are not, in fact, $(2f + 1 - \eta)$ dimensional. The full agreement of the results obtained below with rules (2.5a)–(2.5d), (2.6a), and (2.6b) assumed by us in Sec. II is obtained.

Definition: We shall say that we have a singular case for the coefficients *6j* if any one of the following relations holds:

$$2\omega_{-\gamma} + 1 = -(2\omega_{\gamma+1} + 1)\,, \tag{5.1a}$$

$$2\omega_{-N} + 1 = 0\,, \tag{5.1b}$$

$$2\omega_{-N} + 1 = q_{-1}\,, \tag{5.1c}$$

$$2\omega_{-N} + 1 = q_0\,. \tag{5.1d}$$

The following representation leads to the singular cases A, B, C, D:

(A)	$\Omega_i = \Omega_{i+1}$,	or if [for the SO($2N$) group only] $\Omega_{N-1} = -\Omega_N$,
(B)	$(\Omega) = (\Omega_1,\Omega_2,...,\Omega_{N-1},0)$,	for the SO($2N+1$) group,
(C1)	$(\Omega) = (\Omega_1,\Omega_2,...,\Omega_{N-1},0)$,	for the SP($2N$) group,
(C2a)	$(\Omega) = (\Omega_1,\Omega_2,...,\Omega_{N-1},0)$,	for the SO($2N$) group,
(C2b)	$(\Omega) = (\Omega_1,\Omega_2,...,\Omega_{N-2},0,0)$,	for the SO($2N$) group,
(D)	$(\Omega) = (\Omega_1,\Omega_2,...,\Omega_{N-1},\frac{1}{2})$,	for the SO($2N+1$) group.

The separation of set $[\omega]$ into two subsets should be done as follows:

$$[\omega] \to [\omega]_{\text{in}} \oplus [\omega]_{\text{out}}\,,$$

where $[\omega]_{\text{out}} = [\omega_{-\gamma},\omega_{\gamma+1}]$ for case A, $[\omega]_{\text{out}} = [\omega_{-N},\omega_0]$ ($\omega_{-N} = \omega_0 = -\frac{1}{2}$) for case B, $[\omega]_{\text{out}} = [\omega_{-N}]$ for cases C1, C2a, and D, $[\omega]_{\text{out}} = [\omega_{-N+1},\omega_N,\omega_{-N}]$ for case C2b (here we simultaneously have case A with respect to the pair ω_{-N+1},ω_N).

Note 1: If more than one pair of $[\omega_{-\gamma},\omega_{\gamma+1}]$ satisfying (5.1A) is found, or if (5.1b) or (5.1c) or (5.1d) is found simultaneously with (5.1a), then all of the above components should be included in $[\omega]_{\text{out}}$.

Now the following results are immediately found.

(1) The expressions $2\omega + 1$ for all ω belonging to $[\omega]_{\text{out}}$ are the roots of Eq. (4.4) for singular A, B, and all C cases. Hence for all cases the separation of $[q]$ into two parts should be done (5.2a), and the elements $[q]_{\text{out}}$ are obtained from (5.2b), where all $\omega \in [\omega]_{\text{out}}$ are used:

$$[q] \to [q]_{\text{in}} \oplus [q]_{\text{out}}\,, \tag{5.2a}$$

$$q = 2\omega + 1\,. \tag{5.2b}$$

(2a) For cases A and B we have $X_N (\equiv X_{-N}) = 0$ (here the notation $[q]_{\text{out}} = [q_N,q_{-N}]$ is used).

(2b) For case D we obtain $X_0 = 0$ ($[q]_{\text{out}} = [q_0]$).

(2c) For cases C1 and C2a we have $Q_N = 0$ (here the notation $[q]_{\text{out}} = [q_N]$ is used). For case C2b we obtain $Q_{N-1} = 0$, $X_N = 0$ ($[q]_{\text{out}} = [q_{N-1},q_N,q_{-N}]$).

(3) Except for case C2a we have $\phi(\Omega)_{\alpha 1} = 0$ for all α such that $\omega_\alpha \in [\omega]_{\text{out}}$ [see (2.5a)–(2.5e) and the text below].

(4) For all singular cases the sets $[\omega]_{\text{in}}$, $[q]_{\text{in}}$ may be used in Eq. (4.3a) instead of $[\omega]$ and $[q]$ if X_b ($q_b \in [q]_{\text{in}}$) is calculated.

If the results of points (2) and (3) are substituted into

expressions (4.5b)–(4.5d), and also if Note 1 is taken into account, we obtain that the dimension of the square matrix $\phi(\Omega)$ is reduced [except for case (C2a)]. The range of index a may be chosen according to (2.6a) and (2.6b) and (2.4a)–(2.4c) [see also the text below (2.4c)]. The range of the α index is reduced to such α that ω_α belongs to $[\omega]_{\text{in}}$.
to $[\omega]_{\text{in}}$.

For the case (C2a) the range of α remains unchanged for all equations (4.5a)–(4.5c) and the following modification for components $\phi(\Omega)_{\alpha N}$ should be done (here $Q_N = 0$):

$$\phi(\Omega)_{\alpha N} = (\delta_{\alpha N} - \delta_{-\alpha N})(2D_{\Lambda_+} D_\alpha)^{-1/2}. \tag{5.3}$$

Formula (5.3) is obtained from the unitarity requirements (2.7b) rather than from (4.5b). It should be noticed for the above case that two eigenvalues of H, $\psi(\Omega)_{N1}$ and $\psi(\Omega)_{-N1}$, are equal and one of them belongs to the one-dimensional block H^N_N [see Lemma 2(b)].

VI. SOLUTION FOR A SIMPLE CLASS OF REPRESENTATIONS

In this section we find explicit expressions for the cases when $(\Omega) = (\Omega_1^{p_1}\Omega_2^{p_2})$. If the results of the previous section will be applied, we obtain that $[\omega]_{\text{in}}$ contains only four [or five for the SO($2N+1$) case, $\omega_0 = -\frac{1}{2}$] elements noted here:

$$\omega_\alpha = \omega'_\alpha + \tfrac{1}{2}(p_\alpha - 1), \tag{6.1}$$

$$\omega'_{\pm 1} = \pm[\Omega_1 + \tfrac{1}{2}(p_1 + 2p_2 + \delta)], \tag{6.2a}$$

$$\omega'_{\pm 2} = \pm[\Omega_2 + \tfrac{1}{2}(p_2 + \delta)], \tag{6.2b}$$

where $\delta = 1, 0, -1$, respectively, for cases SP($2N$), SO($2N+1$), and SO($2N$). The equation for $q = q_2$ becomes quadratic, and the following formulas are found:

$$(2\omega_\alpha + 1)^2 - q_2^2 = p_\alpha(2\omega'_\alpha + p_\alpha + \delta)(2\omega'_\alpha - p_\alpha - 2p_\beta - \delta) \times\{(\omega'_\alpha + \tfrac{1}{2}N)^2 - \omega'^2_\beta\}/S, \tag{6.3}$$

$$S = \tfrac{1}{2}C_\Omega = p_1\omega'^2_1 + p_2\omega'^2_2 + \tfrac{1}{4}N[(N+\delta)^2 + p_1p_2], \tag{6.4}$$

$$Q_2^A = \frac{1}{4}\left[N\prod_{\alpha=1}^{2}\{(2\omega'_\alpha)^2 - (p_\alpha + \eta)^2\}/S\right]^{1/2}, \tag{6.5}$$

$$X_2^A = [p_1p_2/N \times H^4(\omega'_1,\omega'_2)/S]^{1/2}, \tag{6.6}$$

$$H^4(\omega'_1,\omega'_2) = [(\omega'_1 + \tfrac{1}{2}N)^2 - \omega'^2_2][(\omega'_1 - \tfrac{1}{2}N)^2 - \omega'^2_2], \tag{6.7}$$

$$Q_2^B = \tfrac{1}{4}[\Pi_B/S]^{1/2}, \tag{6.8}$$

$$\Pi_B = \frac{1}{2}\left(N - \frac{1}{2}\right)\prod_{\alpha=1}^{2}\{(2\omega'_\alpha)^2 - (p_\alpha + 1)^2\} + \frac{1}{2}\left(N + \frac{1}{2}\right)\prod_{\alpha=1}^{2}\{(2\omega'_\alpha)^2 - (p_\alpha - 1)^2\}, \tag{6.9}$$

$$X_2^B = X_2^A\left[N\prod_{\alpha=1}^{2}\{(2\omega'_\alpha)^2 - p_\alpha^2\}/\Pi_B\right]^{1/2}, \tag{6.10}$$

$$X_0^B = \frac{1}{4}\left[\frac{1}{2N+1}\prod_{\alpha=1}^{2}\{(2\omega'_\alpha)^2 - (p_\alpha + 1)^2\}\right]^{1/2}(Q_2^B)^{-1}. \tag{6.11}$$

Index A in the above expressions is referred to the SP($2N$) and SO($2N$) cases and index B to the SO($2N+1$) case. In formula (6.3) we let $\omega'^2_\beta = \omega'^2_2$, $p_\beta = p_2$ if $\alpha = \pm 1$, and $\omega'^2_\beta = \omega'^2_1$, $p_\beta = p_1$ if $\alpha = \pm 2$.

The above expressions should be substituted into (4.5b) and (4.5c) or into (4.5b)–(4.5d) for the SO($2N+1$) case. Also if (3.2a)–(3.2d) and (4.5a) will be added, then we obtain all elements of the $\phi(\Omega)$ matrix.

[1] D. P. Zhelobenko, "Klassiceskije grupy. Spektralnyj analiz konecnomiernyh predstavlenij," U.M.N. XVII, No. 1, 27 (1962).
[2] V. Amarin, U. Dozzio, and C. Oleari, "Proof an algorithm of the brunching multiplicity SO($2N$)↦SO($2N$)×U(1)," J. Math. Phys. **25**, 2140 (1984).
[3] A. M. Bincer, "Missing label operators in the reduction Sp($2N$)↘SP($2N-2$)×SP(2)," J. Math. Phys. **21**, 67 (1980).
[4] A. M. Bincer, "Missing label operators in the reduction O(p)↘O($p-2$)×O(2)," J. Math, Phys. **24**, 1695 (1983).
[5] J. R. Derome and W. T. Sharp, "Racach algebra for an arbitrary group," J. Math. Phys. **6**, 1584 (1965).
[6] J. R. Derome, "Symmetry properties of the 3j-symbols for an arbitrary group," J. Math. Phys. **7**, 612 (1966).
[7] B. Wyboune, *Symmetry Principles and Atomic Spectroscopy* (Wiley–Interscience, New York, 1970).
[8] A. O. Barut and R. Raczka, *Theory of Group Representations and Applications* (Polish Scientific, Warsaw, 1977).

SU(2) and SU(1,1) time-ordering theorems and Bloch-type equations

G. Dattoli and A. Torre[a)]
ENEA, Dip. TIB–Divisione Fisica Applicata, C. R. E. Frascati, C. P. 65–00044 Frascati, Rome, Italy

(Received 14 April 1986; accepted for publication 24 September 1986)

Algebraic time-ordering techniques for SU(2) and SU(1,1) coherence preserving Hamiltonians are reviewed. The link with Bloch-type equations is pointed out and the extension of the method to higher groups is briefly discussed.

I. INTRODUCTION

The search for methods that allow the analytical treatment of many problems in quantum optics is under active consideration. In fact, the numerical analysis of the dynamical behavior of a quantum system undergoing a strong and time-dependent interaction may be expensive and it could miss, in some cases, the essential features of the problem. Analytical methods, whenever possible, can therefore offer a more appropriate solution, providing also a deeper understanding of the physics problem under study.

We recall that exact solutions have indicated their powerfulness by elucidating, e.g., the dynamic of three-level atoms[1] or by providing a more clear understanding of the physical features of two-level atoms interacting with symmetric pulses.[2] These are just two examples that have effectively shown how a clever mathematical formulation of a physics problem has been a precious tool to indicate new and previously unsuspected features and to prove the underlying connection to other seemingly unrelated fields.

The usefulness of rigorous algebraic methods applied to the time-ordering problems has been emphasized by the present authors in a number of recently published papers.[3–6] In particular, it has been shown that the analytical expression of the evolution operator can be obtained for Hamiltonians written as time-dependent linear combination of the generators of the SU(2) (Ref. 3) and SU(1,1) (Ref. 4) groups.

Furthermore, the method has been applied to more complicated time-dependent Hamiltonians involving the SU(1,1) and Weyl–Heisenberg groups[5] and later SU(3) (Ref. 6).

The keynote of the above papers has been the rediscovery, and the suitable rehandling, of the Wei–Norman algebraic method.[7]

It is, however, well known that the dynamic of a quantum system ruled by a SU(2) or SU(1,1) Hamiltonian can be treated using, in the Heisenberg picture, a set of equations known as the torque Bloch equations.[8,9] A natural question can be therefore the following: "Is it possible to recover a Bloch-type dynamic from the characteristic equations of the SU(2) and SU(1,1) time-ordering procedure?"

In this paper we show that this is possible and we dwell on this aspect of the problem because it may allow a generalization of the ordering theorems, e.g., the SU(n) case, in a rather straightforward way. We will finally add some comments on the possibility of getting a "global" exact solution for the problem of SU(2) and SU(1,1) dynamics. By global exact solution we mean the analytical expression of the evolution operator and a closed form of the wave functions describing the time behavior of SU(2) and SU(1,1) states.

The paper consists of three sections. In Sec. II we show how Bloch-type equations can be derived from the characteristic equations of the time-ordering procedure. In Sec. III we indicate under what conditions a full analytical expression of the evolution operator can be obtained and we also calculate the most general form of the wave functions for both SU(2) and SU(1,1) states. Finally in the Appendix we show how the characteristic functions of the ordering procedure are linked to the average values of the SU(2) and SU(1,1) generators.

II. TIME ORDERING AND LINK WITH THE BLOCH-TYPE DYNAMICS

In Refs. 3 and 4 we have shown that the evolution operator for a quantum system driven by a Hamiltonian of the type

$$\hat{H}(t)=\frac{\omega(t)}{2}\hat{F}_0+\Omega^*(t)\hat{F}_+-\Omega(t)\hat{F}_- \quad (\hbar=1), \tag{2.1}$$

can be written as follows:

$$\hat{U}(t,t_0)=\exp\left\{\left[h(t)-\frac{i}{2}\int_{t_0}^{t}\omega(\tau)d\tau\right]\hat{F}_0\right\}\cdot \times\exp\{g(t)\hat{F}_+\}\exp\{f(t)\hat{F}_-\}\hat{I}. \tag{2.2}$$

The operators $\hat{F}$ obey the following rules of commutation:

$$[\hat{F}_0,\hat{F}_\pm]=\pm 2\hat{F}_\pm, \quad [\hat{F}_+,\hat{F}_-]=-\delta\hat{F}_0, \tag{2.3}$$

and can be identified with the SU(2) or SU(1,1) generators according to whether $\delta=\pm 1$ (see Ref. 10 for further comments). The functions $\omega(t)$ and $\Omega(t)$ are nonsingular functions of time, real and complex, respectively. Furthermore, the $\omega(t)$, $\Omega(t)$, $\Omega^*(t)$, and h,g,f functions are linked by the system of differential equations

$$\begin{aligned}\dot{h}(t)&=\delta g(t)\dot{f}(t),\\ \dot{g}(t)&=-i\Omega^*(t)\exp\left\{-2h(t)+i\int_{t_0}^{t}\omega(\tau)d\tau\right\}\\ &\quad -\dot{h}(t)g(t),\\ \dot{f}(t)&=i\Omega(t)\exp\left\{2h(t)-i\int_{t_0}^{t}\omega(\tau)d\tau\right\}\end{aligned} \tag{2.4}$$

$$(h(t_0)=g(t_0)=f(t_0)=0).$$

[a)] ENEA guest.

 0022-2488/87/030618-04$02.50

It is well known that the above system can be solved once a single Riccati equation for $\dot{h}$ can be solved.[7] It is more convenient, for the present purposes, to introduce the new functions

$$\begin{aligned} &\mathscr{H}^* = e^{-h^*}, \quad \mathscr{H}^*(t_0) = 1, \quad \dot{\mathscr{H}}^*(t_0) = 0, \\ &\mathscr{F} = fe^{-h}, \quad \mathscr{F}(t_0) = 0, \quad \dot{\mathscr{F}}(t_0) = i\Omega(t_0), \end{aligned} \tag{2.5}$$

which, as immediately verified from (2.4), obey the following second-order differential equations:

$$\begin{aligned} &\ddot{\mathscr{H}}^* + [(-\dot{\Omega}^*/\Omega^*) - i\omega]\dot{\mathscr{H}}^* + \delta|\Omega|^2\mathscr{H}^* = 0, \\ &\ddot{\mathscr{F}} + [(-\dot{\Omega}/\Omega) + i\omega]\dot{\mathscr{F}} + \delta|\Omega|^2\mathscr{F} = 0. \end{aligned} \tag{2.6}$$

Equations (2.6) have a very familiar form, for $\delta = 1$ they reproduce indeed the well-known equations for the two-level atom amplitude probabilities,[8] while the $\delta = -1$ case is encountered in the analysis of two photon processes.[9] We can cast the motion equations of $\mathscr{F}$ and $\mathscr{H}^*$ in the form of a Bloch-type torque equation by embedding these functions as follows:

$$\begin{aligned} &W = |\mathscr{F}|^2 - \delta|\mathscr{H}|^2, \\ &U = \sqrt{\delta}[\mathscr{F}\mathscr{H} + \mathscr{F}^*\mathscr{H}^*], \\ &V = -i\sqrt{\delta}[\mathscr{F}\mathscr{H} - \mathscr{F}^*\mathscr{H}^*]. \end{aligned} \tag{2.7}$$

Identifying (U, V, W) as the components of a vector $\mathscr{M}_\delta$, we find

$$\dot{\mathscr{M}}_\delta = \Omega_\delta \times \mathscr{M}_\delta, \tag{2.8}$$

where

$$\Omega_\delta \equiv [2\delta^{3/2}\,\mathrm{Re}\,\Omega, 2\delta^{3/2}\,\mathrm{Im}\,\Omega, \omega] \tag{2.9}$$

(where Re Ω and Im Ω are, respectively, the real and imaginary part of Ω). A particularly interesting consequence is the following law of conservation:

$$|\mathscr{F}|^2 + \delta|\mathscr{H}|^2 = \delta. \tag{2.10}$$

The physical meaning of the vector $\mathscr{M}_\delta$ is clarified in the Appendix, where it is shown that

$$\begin{aligned} &W \propto -\langle \hat{F}_0 \rangle, \quad U \propto -(\langle \hat{F}_+ \rangle - \langle \hat{F}_- \rangle), \\ &V \propto i(\langle \hat{F}_+ \rangle - \langle \hat{F}_- \rangle). \end{aligned} \tag{2.11}$$

A few comments are now in order. When $\delta = 1$, Eqs. (2.8) are the ordinary Bloch equations and describe a rotation in Euclidean space. When $\delta = -1$, Eqs. (2.8) can be identified as the SU(1,1) Bloch equations introduced in Ref. 9 and can be understood as the rotation of the pseudovector $\mathscr{M}_{-1}$ in a Lobatchevsky space. [To be more precise when $\delta = -1$, the motion equations are relevant to be an O(2,1) space structure.] Furthermore, the relation (2.10) states the conservation of the "norm" of the vector $\mathscr{M}_\delta$. [It can be easily proved that the norm of $\mathscr{M}_\delta$ is linked to the average value of the Casimir invariants of both SU(2) and SU(1,1) groups.]

Let us now briefly discuss the relevance of the above results to derive time-ordering relations for higher-order groups. It has been shown that Bloch-type equations can be written for the SU(n) case too, under the form of a torque equation in a $(n^2 - 1)$-dimensional space.[1] It has also been proved that an ordering procedure of the type discussed in the paper can be exploited for the SU(3) group too.[6] We make therefore the following conjecture: a one to one correspondence may be found between the generalized Bloch vector components and a suitable combination of the $\mathscr{F}$ and $\mathscr{H}$ functions, entering the ordered form of the evolution operator for a SU(n) coherence preserving Hamiltonian. As a consequence, the SU(n) $\mathscr{H}$ and $\mathscr{F}$ functions can be cast in the form of a torque equation in a $(n^2 - 1)$-dimensional space. We have an indication that the conjecture is true for the SU(3) case. If proved true, in general the hypothesized correspondence may be a powerful tool in solving quite straightforwardly the time-ordering problems, or in general the disentangling problem, for the Hamiltonians linear combinations of SU(n) operators.

III. EXACT SOLUTIONS

We have already mentioned the possibility of obtaining what we have called a global exact solution for the problem under study.

To this aim we should specify: (a) under what conditions can Eqs. (2.6) be solved exactly, and (b) what is the form of the wave function of quantum states ruled by the Hamiltonian (2.1)?

In view of the fact that the Eqs. (2.6) are similar to those encountered in studying two-level systems, we can clarify the first point by generalizing the technique developed by Bambini and Berman.[2] The method consists in mapping Eqs. (2.6) onto the hypergeometric equation (i.e., $Z(1-Z)y'' + [\gamma - (\alpha + \beta + 1)Z]y' - \alpha\beta y = 0$), by means of a change of variables $Z = Z(t)$. Thus getting [e.g., for the second of Eqs. (2.6)]

$$\mathscr{F}'' + \frac{d/dt\ln(\dot{Z}/s) - i(\dot{\phi} - \omega)}{\dot{Z}}\mathscr{F}' + \frac{\delta s^2}{\dot{Z}^2}\mathscr{F} = 0. \tag{3.1}$$

(We assume a "chirped" pulse $\Omega = se^{i\phi(t)}$ and the prime means derivative with respect to Z.)

To map (2.1) onto the hypergeometric equation we must require that when t ranges from $-\infty$ to $+\infty$ the new variable Z ranges from 0 to 1, and furthermore

$$\begin{aligned} &\frac{\delta s^2}{\dot{Z}^2} = -\frac{\alpha\beta}{Z(1-Z)}, \\ &\dot{Z} = Z(1-Z)[i\omega - i\dot{\phi}]/[\gamma - \tfrac{1}{2} - (\alpha + \beta)Z], \end{aligned} \tag{3.2}$$

where α, β, and γ are the characteristic parameters of the hypergeometric equation.

Since we are free to fix both the form of the frequency and of the chirping, we impose (see also Ref. 11)

$$\begin{aligned} &\dot{Z} = Z(1-Z), \\ &i(\omega - \dot{\phi}) = \gamma - \tfrac{1}{2} - (\alpha + \beta)Z. \end{aligned} \tag{3.3}$$

Because ω and ϕ are real functions we must impose

$$\gamma = i\mu + \tfrac{1}{2}, \quad \alpha = i\lambda, \quad \beta = i\eta, \tag{3.4}$$

which amounts to

$$\omega - \dot{\phi} = \mu - (\lambda + \eta)Z. \tag{3.5}$$

The quantity $s^2/\dot{Z}^2$ is positive, therefore using the first equation of (3.2) and (3.4) two conditions follow: if $\delta = -1$ then $\lambda\eta < 0$ and if $\delta = 1$ then $\lambda\eta > 0$. In both cases, however, from (3.2) and (3.3) we obtain

$$Z=\tfrac{1}{4}\operatorname{sech}^2 t/2,\quad s=|\lambda\eta|^{1/2}\operatorname{sech}(t/2),$$
$$\omega-\dot{\phi}=\mu-(\lambda+\eta)[e^t/(1+e^t)]\,. \tag{3.6}$$

Finally the solution of (3.1) can be written as

$$\mathcal{F}_\delta = A\,{}_2F_1[i\lambda,i\mu,\tfrac{1}{2}+i\mu,e^t/(1+e^t)] + B\,[e^t/(1+e^t)]^{(1/2-i\mu)}\,{}_2F_1[i(\lambda-\mu) + \tfrac{1}{2},i(\eta-\mu)+1,\tfrac{3}{2}-i\mu,e^t/(1+e^t)]\,, \tag{3.7}$$

where A and B are constants and ${}_2F_1(\cdot)$ is the hypergeometric function. Note that a similar solution can also be obtained for $\mathcal{H}^*$. What is interesting in the above results is that the sech-type pulse and its generalization allow exact solutions even for the SU(1,1) case.

We now treat statement (b) and the main problem will be the search for a closed expression for the matrix elements of the evolution operator.

We discuss separately the SU(2) and SU(1,1) cases.

(a) SU(2): The wave function describing the evolution of states driven by a Hamiltonian of the type (2.1), where the $\hat{F}$ have been identified as the generators of the SU(2) group ($\hat{F}_0=2J_3$, $\hat{F}_+=\hat{J}_+$, $\hat{F}_-=-\hat{J}_-$), can be expressed as a linear superposition of angular momentum states

$$|\Psi(t)\rangle=\sum_{m=-J}^{J} C_m(t)|J,m\rangle. \tag{3.8}$$

The form of the time-dependent coefficients C_m depends on both the wave function initial values and on the "scattering matrix" $S(t,t_0)$ through the relation

$$C_m(t)=\sum_{m=-J}^{J} S_{m,n}(t,t_0)C_m(t_0), \tag{3.9}$$

and the matrix elements $S_{m,n}(t,t_0)$ are given by

$$S_{m,n}(t,t_0)=\langle J,n|\hat{U}(t,t_0)|J,m\rangle. \tag{3.10}$$

A straightforward but tedious application of the angular momentum operators properties, yields for $S_{m,n}$ the following expression ($t_0=0$):

$$S_{m,n}(t)=\left[\binom{J+n_>}{J+n_<}\binom{J-n_<}{J-n_>}\right]^{1/2}\exp\left\{-\mathrm{in}\int_0^t\omega(\tau)d\tau\right\} \times\mathcal{H}^{-(n+m)}[\operatorname{sgn}(n-m)|\mathcal{F}|]^{n_>-n_<} \times\exp\{i\chi(n-m)\}{}_2F_1[-J-n_<, J-n_<+1;n_>-n_<+1,|\mathcal{F}|^2] \quad[\chi=\arg(\mathcal{F}),\ n_>=\max(m,n),\ n_<=\min(m,n)]\,. \tag{3.11}$$

Therefore once $\mathcal{H}$ and $\mathcal{F}$ are analytically known the problem is completely solved.

(b) SU(1,1): In this case the procedure is almost similar to the previously described one. Once the $\hat{F}$ generators are recognized as those of the SU(1,1) group ($\hat{F}_0=2\hat{K}_0$, $\hat{F}_+=\hat{K}_+$, $\hat{F}_-=-\hat{K}_-$), the wave function describing the evolution of the quantum system ruled by the Hamiltonian (2.1) can be expanded as

$$|\Psi(t)\rangle=\sum_{n=0}^{\infty} C_n(t)|n,k\rangle, \tag{3.12}$$

where the states $|n,k\rangle$ diagonalize the compact generator $\hat{K}_0$ as $\hat{K}_0|n,k\rangle=(n+k)|n,k\rangle$ and furthermore, $\hat{K}_+|n,k\rangle=[(n+1)(n+2k)]^{1/2}|n+1,k\rangle$, $\hat{K}_-|n,k\rangle=[n(n+2k-1)]^{1/2}|n-1,k\rangle$. Finally, k is the Bergman index specifying the eigenvalue of the Casimir invariant.[10]

The explicit form of the coefficients $C_m(t)$ depends on the matrix element

$$S_{m,n}(t,t_0)=\langle n,k|\hat{U}(t,t_0)|m,k\rangle, \tag{3.13}$$

which for $t_0=0$ reads

$$S_{m,n}(t)=\left[\binom{n_>}{n_<}\binom{n_>+2k-1}{n_<+2k-1}\right]^{1/2}\exp\left\{-i(n+k)\int_0^t\omega(t)dt\right\} \times\mathcal{H}^{-(n+m+2k)}[\operatorname{sgn}(n-m)|\mathcal{F}|]^{n_>-n_<}\exp\{i\chi(n-m)\} \times{}_2F_1(-n_<,-n_<-2k+1;n_>-n_<+1;-|\mathcal{F}|^2). \tag{3.14}$$

This last relation completes our short analysis of the exact solutions for SU(2) and SU(1,1) problems.

Let us now summarize the results of the present paper.

(1) We have cast the characteristic equations of SU(2) and SU(1,1) time-ordering in a Bloch-type form.

(2) We have indicated the conditions under which the evolution operator can be calculated analytically.

(3) We have written the expression of the "scattering matrix" for both SU(2) and SU(1,1) state dynamics.

Similar considerations relevant to the SU(3) group will be published elsewhere.[6]

ACKNOWLEDGMENTS

The authors express their sincere appreciation to A. Bambini for enlightening discussions and for his kind interest and encouragements. It is also a pleasure to thank M. Matera for clarifying to us some experimental implications of the present work. Finally we owe our gratitude to A. Renieri and T. Hermsen for a number of useful discussions.

APPENDIX: EXPECTATION VALUES

In this appendix we sketch the derivation of Eqs. (2.11). The evolution operator in the interaction picture takes the form

$$\hat{U}_{\mathrm{int}}=\exp\{2h\hat{F}_0\}\exp\{g(t)\hat{F}_+\}\exp\{f(t)\hat{F}_-\}\hat{I}, \tag{A1}$$

and the functions, h, f, and g are defined by the system (2.4). The average value of the operators $\hat{F}_0$, $\hat{F}_+$, and $\hat{F}_-$ can be evaluated by means of the dot product

$$\langle|\hat{U}_{\mathrm{Int}}^{-1}\hat{F}_{0,\pm}\,\hat{U}_{\mathrm{Int}}|\rangle, \tag{A2}$$

where $\langle\,|$ denotes the initial state assumed for sake of simplicity to be an eigenstate of the operator $\hat{F}_0$.

The explicit expression of the evolution operator in the adopted representation and the commutation rules relevant

to the algebra involved provide us with the following relations between the components of the vector $\mathscr{M}_\delta$ and the average values of the $\hat{F}$ generators:

$$\begin{aligned}
&\langle \hat{F}_0(t)\rangle = (1-2\delta fg)\langle \hat{F}_0(0)\rangle,\\
&\langle \hat{F}_+(t)\rangle = -\delta f e^{-2h}\langle \hat{F}_0(0)\rangle,\\
&\langle \hat{F}_-(t)\rangle = \delta g e^{2h}(1-\delta fg)\langle \hat{F}_0(0)\rangle,
\end{aligned} \tag{A3}$$

$\langle \hat{F}_0(0)\rangle$ denoting the initial average value of the operator $\hat{F}_0$. By introducing the function

$$\mathscr{G} = g e^{h}, \tag{A4}$$

which by means of Eqs. (2.4) can be immediately recognized as the conjugate of the $\mathscr{F}$ one, defined by Eq. (2.5): $\mathscr{G} = \mathscr{F}^*$, we can recast the above relations as

$$\begin{aligned}
&\langle \hat{F}_0(t)\rangle = -\delta W\langle \hat{F}_0(0)\rangle,\\
&\langle \hat{F}_+\rangle + \langle \hat{F}_-\rangle = -i\sqrt{V}\langle \hat{F}_0(0)\rangle,\\
&\langle \hat{F}_+\rangle - \langle \hat{F}_-\rangle = -\sqrt{\delta U}\langle \hat{F}_0(0)\rangle,
\end{aligned} \tag{A5}$$

and thus the relation (2.11) is proved.

[1] J. N. Elgin, Phys. Lett. A **80**, 141 (1980); F. T. Hioe and J. H. Eberly, Phys. Rev. Lett. **47**, 838 (1981); G. Dattoli and R. Mignani, J. Math. Phys. **26**, 3200 (1985).
[2] A. Bambini and P. R. Berman, Phys. Rev. A **23**, 2496 (1981).
[3] G. Dattoli, J. C. Gallardo, and A. Torre, J. Math. Phys. **27**, 772 (1986).
[4] G. Dattoli, A. Torre, and R. Caloi, Phys. Rev. A **33**, 2789 (1986).
[5] G. Dattoli, F. Orsitto, and A. Torre, Phys. Rev. A **34**, 2466 (1986).
[6] G. Dattoli, P. Di Lazzaro, and A. Torre, Phys. Rev. A **34**, 2466 (1986).F. Ciocci, G. Dattoli, A. Renieri, and A. Torre, Phys. Rep. **141**, 1 (1986).
[7] J. Wei and E. Norman, J. Math. Phys. **4**, 575 (1963).
[8] L. Allen and J. H. Eberly *Optical Resonances and Two-Level Atoms* (Wiley, New York, 1975).
[9] G. Dattoli, A. Dipace, and A. Torre, Phys. Rev. A **33**, 4387 (1986).
[10] D. R. Truax, Phys. Rev. D **31**, 1988 (1985); C. C. Gerry, Phys. Rev. A **31**, 2721 (1985), and references therein.
[11] F. J. Hioe and C. E. Carroll, J. Opt. Soc. Am. B **3**, 497 (1985); J. Zakrzewski, Phys. Rev. A **32**, 3748 (1985).

Mathematical structures for long-range dynamics and symmetry breaking

G. Morchio
Dipartimento di Fisica dell'Università, Pisa, Italy

F. Strocchi
International School for Advanced Studies (ISAS), Trieste, Italy

(Received 10 June 1985; accepted for publication 15 October 1986)

The algebraic dynamics of systems with long-range (instantaneous) interactions requires an enlargement of the (quasi) local algebra which, in most relevant cases, includes variables at infinity that enter in an essential way in the time evolution of local variables. The mathematical structures emerging for the treatment of long-range dynamics are investigated also in connection with spontaneous symmetry breaking.

I. INTRODUCTION

The algebraic treatment of quantum dynamical systems with short-range or local interactions has been the subject of many investigations and basic structures, like time evolution of local field algebras, the definition of symmetries, and their spontaneous breaking, have been clarified also in connection with the thermodynamical limit.[1]

The situation is less under control in the case of long-range interactions and in general when the finite volume dynamics α^t_V does not converge in norm, as $V\to\infty$. The case in which α^t_V converges weakly and suitable conditions are satisfied in a *given* representation of the field algebra has been discussed in literature.[2] The emphasis is on the conditions of a certain (uniform) convergence of the correlation functions and the algebraic structure is somewhat lost. A general algebraic framework based on a family of physical states and the associated weak topology has been suggested and discussed later by Sewell.[3] From the point of view of the resulting general mathematical structure our Sec. II can be regarded as a development of a number of problems which are behind Sewell's structure. In particular, the relation between the algebraic dynamics and the symmetries of the finite volume (or infrared cutoff) dynamics has not been discussed in the previous approaches.[2,3] Also the role played by the algebra of essential localization and its connection with the enlargement of the family of relevant states (see Sec. VI) has not been discussed in the previous approaches.[2,3] These questions are the main motivation for this paper also in connection with the phenomenon of symmetry breaking without Goldstone's modes. (A short account of the basic structures which characterize long-range dynamics with emphasis on the phenomenon of energy gap generation has been given in Ref. 4. Here the attention is on the mathematical aspects.)

The result of our analysis is that the algebraic dynamics of systems with long-range (instantaneous) interactions requires an *enlargement of the (quasi) local algebra* by including limits with respect to a weak topology defined by a family of "relevant" states.

The nontrivial structure of such family, especially in the presence of symmetries, gives rise to an *algebra with a nontrivial center*, which enters in an essential way in the dynamics.

II. TIME EVOLUTION WITH LONG-RANGE INSTANTANEOUS INTERACTIONS

In this section we discuss a general algebraic framework for describing systems with long-range instantaneous interactions. The basic problem is to describe dynamical variables as elements of a suitable algebra $\mathscr{B}$ and states as positive linear functionals on such an algebra, in such a way that the time evolution is described by an automorphism group of $\mathscr{B}$. From a constructive point of view, one starts with a net of Von Neumann algebras $\mathscr{A}_V$ associated to the finite volumes V. Space translations α_x are generally defined as automorphisms of the "local algebra"

$$\mathscr{A}_0 \equiv \bigcup_V \mathscr{A}_V.$$

The dynamics is defined in terms of finite volume dynamics α^t_V which act as one-parameter groups of automorphisms of $\mathscr{A}_0$, or more generally of its norm closure $\mathscr{A}$. Typically α^t_V are generated by finite volume Hamiltonians H_V affiliated to $\mathscr{A}_0$; more generally α^t_V may describe the dynamics corresponding to an interaction with an infrared cutoff V. Eventually, one has to take the limit $V\to\infty$.

For interactions with finite propagation speed, for any $A\in\mathscr{A}_0$, $\alpha^t_V(A)$ becomes independent of V, for V large enough, and it defines the time evolution α^t as an automorphism of $\mathscr{A}_0$. More generally, for interactions with sufficiently short range, $\alpha^t_V(A)$ converges in norm[4] to an automorphism group of $\mathscr{A}$. For spin systems with two-body interaction, the potential must decay faster than $|\mathbf{x}|^{-3}$ (Ref. 5).

In the case of *long-range (instantaneous) interactions*, in particular for spin systems with potentials decaying slower than $|\mathbf{x}|^{-3}$, the finite volume dynamics α^t_V do not converge in norm and a weaker topology is needed. Physical considerations would suggest that the expectation values of $\alpha^t_V(A)$ converge. The convergence for any state over the algebra $\mathscr{A}$ coincides with the weak convergence with respect to the dual $\mathscr{A}'$, and it defines $\alpha^t(A)$ as an element of $\mathscr{A}''$, the universal Von Neumann algebra of $\mathscr{A}$. In all interesting cases, however, the time evolution for large V involves strongly delocalized variables, the expectation values of which converge only if the states are sufficiently regular at infinity. A detailed discussion of (physically relevant) models which exhibit such phenomena is deferred to subsequent papers. Typical examples are the BCS model,[6,7] the Kibble model,[4] and a large class of mean field spin models.[6]

As a matter of fact, both for gauge theories and for many body nonrelativistic systems, simple physical considerations indicate that the definition of the algebra itself (of dynamical variables) makes implicit reference to a class of states. In both cases it is therefore natural to associate to a system an

 0022-2488/87/030622-14$02.50

algebra of dynamical variables *and* a class of states F, which are at the basis of the physical interpretation of such an algebra. In this perspective, it is natural to require that α_V^t converges weakly with respect to the above class of states. To make the above framework more precise we consider a family F of continuous linear functionals over $\mathscr{A}$ with the following properties: (1) F is closed under linear combinations; (2) F is norm closed and separating, i.e., $\phi(A)=0$, $\forall\phi\in F$, implies $A=0$; and (3) F is "stable under local operations" in the sense that if $\phi\in F$, also $\phi_{AB}(\cdot)\equiv\phi(A\cdot B)$, with $A, B\in\mathscr{A}$, belongs to F. The positive part F^+ of F is thus a full folium as in Ref. 8. These states can be taken to be normal states when restricted to the Von Neumann algebras $\mathscr{A}_V$. We denote by τ_F the weak topology defined by F on $\mathscr{A}''$, the universal Von Neumann algebra of $\mathscr{A}$; in particular τ_F is a weak topology for $\mathscr{A}\subset\mathscr{A}''$.

The resulting structure is characterized by the following propositions.

Proposition 2.1: Let F be a family of linear functionals on A with properties (1), (2), (3). Then there exists a central projector E of the universal Von Neumann algebra $\mathscr{A}''$ such that the elements ϕ of F are characterized as

$$\phi(\cdot)=\psi(E\cdot),\quad \psi\in\mathscr{A}'.$$

The closure $\mathscr{M}$ of $\mathscr{A}$, with respect to the weak topology defined on $\mathscr{A}''$ by F, is a Von Neumann algebra isomorphic to the subalgebra $E\mathscr{A}''\subset\mathscr{A}''$. The weak topology τ_F defined by F on $\mathscr{M}$, coincides with the ultraweak topology defined on $\mathscr{M}$ by the subspace $\mathscr{H}_F=E\mathscr{H}$, $\mathscr{H}$ being the space of the universal representation of $\mathscr{A}$. It also coincides with the weak topology defined on $\mathscr{M}$ by F^+. As an abstract Von Neumann algebra $\mathscr{M}$ is characterized as the dual of the Banach space F and a linear functional on $\mathscr{M}$ belongs to F iff it is τ_F continuous.

The elements of F are also characterized as the linear functionals of $\mathscr{A}'$ which are τ_F continuous on $\mathscr{A}$.

Proposition 2.2: The limit $\lim_{V\to\infty}\alpha_V^t(A)$ exists in $\mathscr{M}$ for any $A\in\mathscr{A}$, in the τ_F topology and it is τ_F continuous on $\mathscr{A}$ iff the $\lim_{V\to\infty}\alpha_V^{t*}\phi$ exists for any $\phi\in F$, in the weak * topology induced by $\mathscr{A}$ on $\mathscr{A}'$, and it belongs to F.

Under the above conditions the mapping defined on $\mathscr{A}$ by

$$\alpha^t=\tau_F-\lim_{V\to\infty}\alpha_V^t \tag{2.1}$$

has a unique τ_F-continuous extension from $\mathscr{A}$ to $\mathscr{M}$, which preserves the sums, the multiplication by scalars, and the * operation. The extension is obtained by taking the transpose of the mapping $\phi\to\phi^t\equiv\lim_V\alpha_V^{t*}\phi$.

Proposition 2.3: The mapping α^t defined by Eq. (2.1) is a morphism of $\mathscr{M}$ iff α_V^t converges on $\mathscr{A}$ in the ultrastrong topology defined by F on $\mathscr{M}$.

Any of the following conditions guarantee that α^t satisfies the properties of Proposition 2.2 and that it is a group of automorphisms of $\mathscr{M}$: (i) α_V^t converges ultrastrongly on $\mathscr{A}$ and weakly on $\mathscr{M}$, (ii) α_V^{t*} converges in the norm topology on F, and (iii) there exists a C^* subalgebra $\mathscr{B}$ of $\mathscr{M}$, $\mathscr{B}\supset\mathscr{A}$, such that for any $B\in\mathscr{B}$, the weak limit of $\alpha_V^t(B)$ exists uniformly on the compact sets of F, defined by the weak * topology induced by $\mathscr{B}$ on F, and it is weakly continuous.

Proposition 2.4: Given α_V^t and $\mathscr{A}$ as above, the set of families of linear functionals F_α with properties (1), (2), (3) of Proposition 2.1 and (4) for all $t\in R$ the weak * limit of α_V^{t*} exists on F_α uniformly on the compact sets of F_α, defined by (the topology induced on F_α by) the weak closure of $\mathscr{A}$ with respect to F_α, maps F_α into itself and satisfies the group law in t, has one and only one maximal $F_{\max}$, which contains all the families F_α.

Condition (4) is equivalent to the convergence of $\alpha_V^t(A)$, $A\in\mathscr{A}$, in the ultrastrong topology defined by F_α, to a group of automorphisms of $\bar{\mathscr{A}}^{F_\alpha}$ (see Ref. 9, Theorem 5.7 and Proposition 2.3 above).

Propositions 2.1, 2.2, and 2.3 show that, under very general conditions, it is indeed possible to define an algebraic dynamics also in the presence of long-range interactions. The resulting picture is that, in this general case, the algebra of dynamical variables must be enlarged to include variables which are not localized. The relevant mathematical feature is that *in the presence of long-range instantaneous interactions the algebraic dynamics is defined on an algebra with a nontrivial center.* This reflects the fact that the time evolution of initially localized variables involves *infinitely delocalized variables*, which commute with $\mathscr{A}$. Since $\mathscr{A}_0$ is dense in $\mathscr{M}$, the center Z of $\mathscr{M}$ coincides with $\mathscr{M}\cap(\cap_V\mathscr{A}_V')$ in $\mathscr{H}_F$ and in this sense it consists of *variables at infinity.*

Proof of Proposition 2.1: The first statement is essentially Theorem 2.7, (iii), (Chap. III of Ref. 9) applied to the Von Neumann algebra $\mathscr{A}''$ and to its predual $\mathscr{A}'$.

To see that $\mathscr{M}$ is isomorphic to $E\mathscr{A}''$, we note that $\mathscr{A}$ with the τ_F topology is isomorphic to the subalgebra $E\mathscr{A}\subset\mathscr{A}''$; in fact the mapping $A\to EA$, $A\in\mathscr{A}$, preserves sums, products, * operation, and seminorms ($\forall\,\phi\in F$, $\phi(A)=\phi(EA)$) so that it is an injection since F separates points, and clearly it is surjective. Now if $\{A_\alpha\}$ is a τ_F convergent net of elements of $\mathscr{A}$ there will exist an element $A\in\mathscr{A}''$ (not necessarily unique since in general τ_F does not separate the points of $\mathscr{A}''$) such that $\forall\,\phi\in F$

$$\phi(A_\alpha)\to\phi(A).$$

By the characteristic property of F,

$$\phi(A_\alpha)\to\phi(A),\quad\forall\,\phi\in F,$$

iff

$$\chi(EA_\alpha)\to\chi(EA),\quad\forall\,\chi\in\mathscr{A}'.$$

Hence the τ_F closure of $\mathscr{A}$ can be identified with $E\mathscr{A}''$.

By Theorem 2.4, Chap. III, of Ref. 9, the weak topology τ_F defined by F on $E\mathscr{A}''\approx\mathscr{M}$ coincides with the ultraweak topology defined on $E\mathscr{A}''$ by $\mathscr{H}_F=E\mathscr{H}$, $\mathscr{H}$ being the universal representation of $\mathscr{A}$.

The dual of the Banach space F is contained in $\mathscr{A}''$ since any continuous linear functional on F can be extended to the whole $\mathscr{A}'$ by the Hahn–Banach theorem, because F is a linear closed subspace of $\mathscr{A}'$. On the other hand, $\mathscr{A}''$ and $\mathscr{M}$ coincide on F since $\phi(A)=\phi(EA)$, $\forall\,A\in\mathscr{A}''$, so that $\mathscr{M}$ is the dual of the Banach space F.

By Theorem 4.2, Chap. III of Ref. 9, every $\phi\in F$ is a complex linear combination of elements of F^+ so that $\tau_{F^+}=\tau_F$.

To prove the last statement of Proposition 2.1, we note that by definition of τ_F the elements of F are τ_F continuous.

Conversely, if $\phi \in \mathscr{A}'$ and it is τ_F continuous on $\mathscr{A}$, we consider a net $\{A_\alpha, A_\alpha \in \mathscr{A}\}$, such that, given $B \in \mathscr{A}$, $A_\alpha \overset{W}{\to} (1-E)B \in \mathscr{A}''$. Then $A_\alpha \to 0$ in the τ_F topology, since $\forall\, \chi \in F$, $\chi(A_\alpha) = \chi(EA_\alpha) \overset{W}{\to} \chi(E(1-E)B) = \chi(0) = 0$, and therefore $\phi(A_\alpha) \to 0$ since ϕ is τ_F continuous. Furthermore, as functionals of $\mathscr{A}'$, $\phi(A_\alpha) \to \phi((1-E)B)$, so that $\phi((1-E)B) = 0$, i.e., $\phi(\cdot) = \phi(E\cdot)$.

Proof of Proposition 2.2: If $\lim_{V\to\infty} \alpha_V^{t*}$ exists for any $\phi \in F$ in the weak * topology induced by $\mathscr{A}$ on $\mathscr{A}'$ and it belongs to F, then given any fixed element $A \in \mathscr{A}$

$$\phi(\alpha_V^t(A)) = (\alpha_V^{t*}\phi)(A) \underset{V\to\infty}{\to} \phi^t(A),$$

and since

$$|\phi^t(A)| \leqslant \|\phi\|\, \|A\|$$

[as a consequence of $|\phi(\alpha_V^t(A))| \leqslant \|\phi\|\, \|A\|$], $\phi^t(A)$ defines a continuous linear functional on F, i.e., an element $\alpha^t(A) \in \mathscr{M}$. Hence $\lim_{V\to\infty} \alpha_V^t(A)$ exists in $\mathscr{M}$ in the τ_F topology.

To see that $\alpha^t(A)$ is τ_F continuous in A, $A \in \mathscr{A}$ we note that $\phi(\alpha^t(A)) = \phi^t(A)$ with $\phi^t \in F$ and then $A_\alpha \to A$ in $\mathscr{A}$ implies $\alpha^t(A_\alpha) \to \alpha^t(A)$ in the τ_F topology.

Conversely, if $\lim_{V\to\infty} \alpha_V^t(A)$ exists in $\mathscr{M}$ for any $A \in \mathscr{A}$ in the τ_F topology and it is τ_F continuous on $\mathscr{A}$, then $\forall\, \phi \in F$, $\lim_{V\to\infty} (\alpha_V^{t*}\phi) \equiv \phi^t$ exists in $\mathscr{A}'$ in the topology induced by $\mathscr{A}$ on $\mathscr{A}'$ (weak * topology) since $\mathscr{A}'$ is complete with respect to the weak * topology (the dual of a Banach space is weakly complete).

Now by assumption α^t is τ_F continuous, so that $\forall\, \phi \in F$, ϕ^t is τ_F continuous on $\mathscr{A}$ and therefore by Proposition 2.1 $\phi^t \in F$.

To prove the last statement of Proposition 2.2 we note that $\mathscr{A}$ is τ_F dense in $\mathscr{M}$ and α^t is τ_F continuous on $\mathscr{A}$ so that there is a unique τ_F continuous extension of α^t from $\mathscr{A}$ to $\mathscr{M}$: in fact

$$A_\alpha \overset{\tau_F}{\to} A \in \mathscr{M}$$

implies

$$\phi(\alpha^t(A_\alpha)) = \phi^t(A_\alpha) \to \phi^t(A) \equiv \phi(\alpha^t(A)),$$

which defines $\alpha^t(A)$ as a continuous linear functional on F, i.e., an element of $\mathscr{M}$ (by Proposition 2.1). Since $\forall\, \phi \in F$, $\phi^t \in F$, α^t is τ_F continuous on $\mathscr{M}$. The extended mapping α^t preserves the sums, the multiplication by scalars, and the * operation because these operations are τ_F continuous.

Proof of Proposition 2.3: We start by proving the first statement. If α_V^t converges on $\mathscr{A}$ in the ultrastrong topology defined by F on $\mathscr{M}$ then, since the product is ultrastrongly continuous (Ref. 10, Chap. I, §3),

$$\alpha^t(AB) = \lim_V \alpha_V^t(AB) = \lim_V \alpha_V^t(A)\alpha_V^t(B)$$
$$= \alpha^t(A)\alpha^t(B), \quad \forall A,\ B \in \mathscr{A}.$$

By the τ_F continuity of the extension of α^t from $\mathscr{A}$ to $\mathscr{M}$ and the weak continuity of the product in each factor, separately, the above equation extends to any $A, B \in \mathscr{M}$: if $A_\alpha \to A \in \mathscr{M}$, $B_\alpha \to B \in \mathscr{M}$, $A_\alpha, B_\alpha \in \mathscr{A}$, then

$$\alpha^t(A)\alpha^t(B) = \lim_\alpha \alpha^t(A_\alpha)\alpha^t(B)$$
$$= \lim_\alpha \lim_\beta \alpha^t(A_\alpha)\alpha^t(B_\beta)$$
$$= \lim_\alpha \lim_\beta \alpha_V^t(A_\alpha\, B_\beta)$$
$$= \lim_\alpha \alpha^t(A_\alpha\, B) = \alpha^t(AB).$$

Thus α^t is a morphism of $\mathscr{M}$. Conversely, if α^t is a morphism of $\mathscr{M}$, $\alpha^t(A^*A) = (\alpha^t(A))^*\, \alpha^t(A)$ and therefore, for each $\psi \in \mathscr{H}_F$ we have

$$\|\alpha_V^t(A)\psi\|^2 = (\psi, \alpha_V^t(A^*)\alpha_V^t(A)\psi)$$
$$= (\psi, \alpha_V^t(A^*A)\psi) \to (\psi, \alpha^t(A^*A)\psi)$$
$$= (\psi, \alpha^t(A^*)\alpha^t(A)\psi) = \|\alpha^t(A)\psi\|^2.$$

On the other hand, $\alpha_V^t(A)\psi$ is a weakly convergent sequence of vectors and the convergence of the norms implies strong convergence. Since $\|\alpha_V^t(A)\| = \|A\|$, the strong convergence in $\mathscr{H}_F$ coincides with the ultrastrong convergence of $\alpha_V^t(A)$ (Ref. 10, Chap. I, §3).

We now prove that condition (i) implies that α_V^t satisfies the conditions of Proposition 2.2 and that α^t defined by Eq. (2.1) is a group of automorphisms of $\mathscr{M}$. Since α_V^t converges weakly on $\mathscr{M}$, $\forall\, \phi \in F$, $\alpha_V^{t*}\phi$ is a weakly convergent sequence in the predual of $\mathscr{M}$ and therefore by Corollary 5.2, Chap. III of Ref. 9, α_V^{t*} converges to an element $\alpha^t \in F$; thus the conditions of Proposition 2.2 are satisfied. The ultrastrong convergence of α_V^t on $\mathscr{A}$ together with the weak continuity of α^t on $\mathscr{M}$ implies that α^t is a morphism of $\mathscr{M}$. To prove the group law we note that a convergent sequence together with its limit defines a compact set with respect to the topology by which it converges, and therefore the weak convergence of α_V^t on $\mathscr{M}$ implies that the sequence $\{\alpha_V^{t*}\phi\}$ together with $\alpha^{t*}\phi$ defines a compact set of F with respect to weak * topology induced by $\mathscr{M}$ on F.

Now, $\forall\, A \in \mathscr{A}$, $\forall\, \phi \in F$

$$\phi(\alpha^t\alpha^s(A)) = \lim_V \lim_{V'} \phi(\alpha_{V'}^t\, \alpha_V^s(A))$$
$$= \lim_V \lim_{V'} (\alpha_{V'}^{t*}\phi)(\alpha_V^s(A)),$$

and since $\alpha_V^s(A)$ is ultrastrongly convergent, the limit is uniform on the compact set of F with respect to the weak * topology induced by on F, so that the above limit as $V \to \infty$ is uniform in V' and therefore it is equal to

$$\lim_V (\alpha_V^{t*}\phi)(\alpha_V^s(A)) = \lim_V \phi(\alpha_V^t\, \alpha_V^s(A))$$
$$= \lim_V \phi(\alpha_V^{t+s}(A)) = \phi(\alpha^{t+s}(A)).$$

Hence, since F separates points of $\mathscr{M}$, $\alpha^t\alpha^s = \alpha^{t+s}$ on $\mathscr{A}$. Furthermore, $\alpha^t\alpha^s$ and α^{t+s} are both defined and τ_F continuous on $\mathscr{M}$ (Proposition 2.2), and since they coincide on the τ_F dense subalgebra $\mathscr{A}$ they coincide everywhere on $\mathscr{M}$.

We now prove that condition (ii) guarantees that α^t is a group of automorphisms of $\mathcal{M}$. The properties of Proposition 2.2 are satisfied since the norm convergence of α_V^{t*} implies that the limit belongs to F. We now show that $\alpha^t\alpha^s = \alpha^{t+s}$. In fact

$$\begin{aligned}
&\|\alpha^{t*}\alpha^{s*}\phi - \alpha^{t+s*}\phi\| \\
&\quad = \|(\alpha^t - \alpha_V^t)^*\alpha^{s*}\phi + \alpha_V^{t*}(\alpha^s - \alpha_V^s)^*\phi \\
&\qquad + \alpha_V^{t*}\alpha_V^{s*}\phi - \alpha^{t+s*}\phi\| \leqslant \|(\alpha^t - \alpha_V^t)^*\alpha^{s*}\phi\| \\
&\qquad + \|\alpha_V^{t*}(\alpha^s - \alpha_V^s)^*\phi\| + \|(\alpha_V^{t+s*} - \alpha^{t+s*})\phi\| \\
&\quad = \|(\alpha^t - \alpha_V^t)^*\alpha^{s*}\phi\| \\
&\qquad + \|(\alpha^s - \alpha_V^s)^*\phi\| + \|(\alpha_V^{t+s} - \alpha^{t+s})^*\phi\|,
\end{aligned}$$

and the right-hand side converges to zero as $V \to \infty$, by assumption.

We now have to show that $\alpha^t(AB) = \alpha^t(A)\alpha^t(B)$. In fact,

$\forall\, A, B \in \mathcal{A}$,

$$\begin{aligned}
\phi(\alpha^t(A)B) &= \lim_V \phi(\alpha_V^t(A)B) \\
&= \lim_V (\alpha_V^{t*}\phi)(A\,\alpha_V^{-t}(B)) \\
&= \lim_V [(-\alpha_\phi^{t*} + \alpha_V^{t*}\phi)(A\,\alpha_V^{-t}(B)) \\
&\quad + (\alpha^{t*}\phi)(A\,\alpha_V^{-t}(B))].
\end{aligned}$$

On the other hand, $\|(\alpha_V^{t*} - \alpha^{t*})\phi\| \to 0$ as $V \to \infty$ by assumption and since $A\,\alpha_V^{-t}(B)$ is a bounded sequence ($\|A\,\alpha_V^{-t}(B)\| \leqslant \|A\|\,\|\alpha_V^{-t}(B)\| = \|A\|\,\|B\|$), the first term in square brackets goes to zero as $V \to \infty$ and we get

$$\phi(\alpha^t(A)B) = \phi(\alpha^t(A\,\alpha^{-t}(B)).$$

The above equation extends to the case $B \in \mathcal{M}$ by weak continuity, and therefore by taking $B = \alpha^t(C)$, $C \in \mathcal{M}$, and using the group law, we get

$$\alpha^t(A)\alpha^t(C) = \alpha^t(AC).$$

The extension to $A \in \mathcal{M}$ follows from τ_F continuity.

Convergence on $\mathcal{A}$ in the ultrastrong topology defined by F amounts to convergence of $\phi(\alpha_V^t(A))$, $A \in \mathcal{A}$ uniformly for ϕ in any subset of F compact with respect to the weak topology $\sigma(F,\mathcal{M})$ defined by $\mathcal{M}$ on F (Theorem 5.7, Chap. III of Ref. 9). By condition (iii), the limit is uniform on compact sets of F in the $\sigma(F,\mathcal{B})$ topology, with $\mathcal{B} \subset \mathcal{M}$. It is therefore enough to show that every compact set with respect to the $\sigma(F,\mathcal{M})$ topology is also $\sigma(F,\mathcal{B})$ compact, i.e., that if any $\sigma(F,\mathcal{M})$ open covering of a set K has a finite subcovering, then the same occurs also for any $\sigma(F,\mathcal{B})$ open covering of F. This follows because $\mathcal{M} \supset \mathcal{B}$ implies that $\sigma(F,\mathcal{M})$ is finer than $\sigma(F,\mathcal{B})$ and therefore an open set with respect to $\sigma(F,\mathcal{B})$ is also $\sigma(F,\mathcal{M})$ open; therefore a $\sigma(F,\mathcal{B})$ open covering is also a $\sigma(F,\mathcal{M})$ open covering.

To prove the group law as for condition (i) we show that $\forall\, A \in \mathcal{A}$, $\lim_V \phi(\alpha_V^t\,\alpha_V^s(A))$ is uniform with respect to V'. In fact $\alpha_{V'}^{t*}$ together with its limit ϕ^t define a compact set with respect to the topology induced by $\mathcal{B}$ on F, since by assumption $\alpha_{V'}^{t*}(B)$ converges $\forall\, B \in \mathcal{B}$. Hence $\forall\, A \in \mathcal{A}$

$$\lim_V \phi(\alpha_{V'}^t\,\alpha_V^s(A)) = \lim_V (\alpha_{V'}^{t*}\phi)(\alpha_V^s(A))$$

converges uniformly with respect to V'. The argument is essentially the same as for condition (i): since α^t is assumed to be τ_F continuous on $\mathcal{B} \supset \mathcal{A}$, it has a unique τ_F continuous extension to $\mathcal{M}$ (Proposition 2.2) and

$$\begin{aligned}
\phi(\alpha^t\alpha^s(A)) &= \lim_V \phi(\alpha^t\,\alpha_V^s(A)) \\
&= \lim_V \lim_{V'} \phi(\alpha_{V'}^t\,\alpha_V^s(A)) \\
&= \lim_V \phi(\alpha_V^t\,\alpha_V^s(A)) = \phi(\alpha^{t+s}(A)),\ A \in \mathcal{A},
\end{aligned}$$

where in the last but one equality we have used the uniformity of the limit with respect to V'. The extension of the above equation to $\mathcal{M}$ follows from weak continuity.

Proof of Proposition 2.4: Let us consider a totally ordered chain $F_{\alpha_1} \subset F_{\alpha_2} \subset \ldots$, then there exists a majorant element

$$F = \overline{\bigcup_{\alpha_j} F_{\alpha_j}} \equiv \overline{F}_0,$$

where the bar denotes the norm closure, which satisfies properties (1)–(4). In fact F_0 is obviously closed under linear combinations and multiplications by scalars are norm continuous, property (1) follows. In a similar way one proves property (3) since the operation $\phi \to \phi_{AB} = \phi(A \cdot B)$, $A, B \in \mathcal{A}$ is norm continuous. Property (2) is obvious. Property (4) for $\overline{F}_0$ is equivalent to the strong convergence of $\alpha_V^t(A)\Psi$, $\forall\, \Psi \in \mathcal{H}_{\overline{F}_0}$, $\forall\, A \in \mathcal{A}$ with $\mathcal{H}_{\overline{F}_0}$ = the subspace of the space $\mathcal{H}_{V\cdot N}$ of the universal representation of $\mathcal{A}$, defined by Proposition 2.1. Now, property (4) for each F_{α_j} implies strong convergence of $\alpha_V^t(A)\Psi$, $\forall\, \Psi$ in a dense subspace $D = \cup_{\alpha_j} \mathcal{H}_{F_{\alpha_j}}$ of $\mathcal{H}_{\overline{F}_0}$ (the density of D in $\mathcal{H}_{\overline{F}_0}$ follows because if $\phi = \|\ \| - \lim \phi_n$, $\phi_n \in F_{\alpha_n}$, then its representative vector Φ in $\mathcal{H}_{\overline{F}_0}$ cannot be orthogonal to all $\mathcal{H}_{F_{\alpha_n}}$) and therefore $\forall\ \Psi \in \mathcal{H}_{\overline{F}_0}$ by norm boundedness of $\alpha_V^t(A)$.

Thus, by Zorn's Lemma, there exist (several) maximal elements of the form F discussed above. We now show that given two such elements F_1, F_2 there exists a G which contains both and therefore there is only one maximal element. In fact, let us put

$$G = \overline{F_1 + F_2}$$

(i.e., the norm closure of elements of the form $\phi = \phi_1 + \phi_2$, $\phi_1 \in F_i$, $i = 1,2$). Then $G \supset F_i$, $i = 1,2$ and properties (1)–(3) hold. To show that property (4) holds let $\mathcal{H}_1$, $\mathcal{H}_2$ be the subspaces of the space $\mathcal{H}$ of the universal representation of $\mathcal{A}$, corresponding to F_1 and F_2. Clearly the subspace $\mathcal{H}_G$ corresponding to G contains $\mathcal{H}_1$ and $\mathcal{H}_2$ and the smallest subspace $\hat{\mathcal{H}}$ containing $\mathcal{H}_1$, $\mathcal{H}_2$ represents all the states of the form $\phi_1 + \phi_2$, $\phi_i \in F_i$, $i = 1,2$, and therefore their norm limits. It follows that $\mathcal{H}_G$ is the Hilbert space generated by $\mathcal{H}_1$ and $\mathcal{H}_2$. Every vector x of $\mathcal{H}_G$ can be written (in a nonunique way) in the form $x = x_1 + x_2$, $x_i \in \mathcal{H}_i$, $i = 1,2$, and therefore the strong convergence in $\mathcal{H}_1$, $\mathcal{H}_2$ implies the strong convergence in $\mathcal{H}_G$.

As stressed before, the construction of algebraic dynamics makes reference to a class of "relevant states," which in the nonlocal case are implicitly at the basis of the definition of the problem, from the beginning. By Proposition 2.4 the class of relevant states may be taken to be maximal and in particular stable under the symmetries of the finite volume dynamics (see Sec. III).

In conclusion, a *dynamical system* should in general be defined as a triple $(\mathcal{M}, F, \alpha^t)$ with $\mathcal{M}$ a Von Neumann algebra with predual F and α^t a one-parameter group of automorphisms of $\mathcal{M}$; this structure is naturally constructed in terms of a quasilocal algebra $\mathcal{A}$, a set of relevant states F, and a family of finite volume (or infrared cutoff) dynamics α^t_V.

As we shall see in the following sections, the above framework is rich enough to allow the *algebraic discussion of spontaneous symmetry breaking* in the presence of long-range instantaneous interactions.

This structure can be seen as a generalization of Kadison's definition of a dynamical system[11]; the essential difference is, however, that here α^{t*} is not required to be continuous on F in the weak * topology defined by $\mathcal{A}$. This property is in fact equivalent (Ref. 11 and Sec. IV) to the stability of $\mathcal{A}$ under time evolution and its failure plays a crucial role in the explanation of *energy gap associated to spontaneous breaking of continuous symmetries.*

III. SYMMETRIES OF NONLOCAL ALGEBRAIC DYNAMICS

For a large class of systems in quantum field theory and in many-body theory, one is interested in symmetries which commute with space-time translations, sometimes called internal symmetries. As a matter of fact it is for this class of symmetries that Goldstone's Theorem constrains the implications of spontaneous symmetry breaking. We are thus led to consider the analog of such symmetries in the case of nonlocal dynamics.

Given an automorphism α of the algebra $\mathcal{A}$ of localized variables the commutativity with the space translation automorphism α_x

$$\alpha\,\alpha_x = \alpha_x\,\alpha \tag{3.1}$$

does not present problems since in general α_x is a well-defined automorphism of $\mathcal{A}$. The situation is quite different for time translations since, as discussed in the previous section, in the case of nonlocal algebraic dynamics, α^t does not leave $\mathcal{A}$ stable. To define the commutation relation

$$\alpha\,\alpha^t = \alpha^t\,\alpha, \tag{3.2}$$

one needs an algebra stable under α and under time translations. One must therefore extend α to the algebra $\mathcal{M}$ stable under α^t.

It is worthwhile to remark that two different topological structures are naturally associated to $\mathcal{M}$. As weak closure of $\mathcal{A}$ with respect to the family of relevant states F, $\mathcal{M}$ is a Von Neumann algebra. On the other hand, regarded as a C^*-algebra, $\mathcal{M}$ identifies a set of states (the dual of $\mathcal{M}$ as C^*-algebra) which not only properly contains F, but also contains states which are not identified by their values on $\mathcal{A}$. This structure would then lead to a significant enlargement of the original problem (see Sec. II). It is an important fact that if an automorphism α of $\mathcal{A}$ can be extended to an automorphism of $\mathcal{M}$, as a C^* algebra, then it is automatically weakly continuous.[10] Therefore the extension is completely defined in the structure $(\mathcal{M}, F)$.

Proposition 3.1: Given an automorphism α of $\mathcal{A}$, α can be extended to an automorphism of $\mathcal{M}$ if and only if the family of states F is stable under α^* and $(\alpha^{-1})^*$.

Since automorphisms of Von Neumann algebras are weakly continuous and $\mathcal{A}$ is weakly dense in $\mathcal{M}$, the extension is uniquely determined by the action of α on $\mathcal{A}$.

Proof: Here α^* is defined on the dual $\mathcal{A}'$ of $\mathcal{A}$ and it is norm preserving since α is an automorphism of $\mathcal{A}$. If α^* leaves family F stable, then $(\alpha^*\phi)(A), A \in \mathcal{M}$, defines a continuous linear functional $f_A(\phi)$ on F, since

$$|f_A(\phi)| < \|\alpha^*\phi\|\,\|A\| = \|\phi\|\,\|A\|.$$

Therefore $f_A(\phi)$ defines an element of the dual of F, i.e., an element of $\mathcal{M}$

$$f_A(\phi) = \phi(B) \equiv \phi(\alpha(A)).$$

This provides an extension of α from $\mathcal{A}$ to $\mathcal{M}$, which is weakly continuous since the weak topology is defined by the elements of F and F is α^* stable. Since $\mathcal{A}$ is weakly dense in $\mathcal{M}$, the extended α preserves the sums, the multiplication by scalars, and the * operation, because these operators are weakly continuous. Moreover, since the product is separately weakly continuous it follows that

$$\alpha(A)\alpha(B) = \alpha(AB),$$

$\forall\, A \in \mathcal{A}, B \in \mathcal{M}$ and therefore also for $A, B \in \mathcal{M}$. Finally α is invertible on $\mathcal{M}$ since α^{-1} can be extended to $\mathcal{M}$ and $\alpha^*\alpha^{-1*} = 1$ on F.

Conversely, if α is an automorphism of $\mathcal{M}$, α is weakly continuous (see Ref. 10) and therefore $\phi(\alpha(A))$ defines a linear functional $f_\phi(A)$ on $\mathcal{M}$ which is weakly continuous since α and ϕ are also. Hence this identifies an element of F, since the functionals of F are the only ones which are weakly continuous on $\mathcal{M}$. Finally, since for $A \in \mathcal{A}$, $f_\phi(A) = (\alpha^*\phi)(A)$, and the functionals of F are determined by their values on $\mathcal{A}$, one has

$$f_\phi = \alpha^*\phi,$$

and α^* maps F into F. The same argument applies to α^{-1}.

In the following, we shall always consider automorphisms α of $\mathcal{A}$ with the property that α^*, α^{-1*} leave the family F stable, so that they can be extended to automorphisms of the Von Neumann algebra $\mathcal{M}$. Proposition 3.2 below shows that this property can always be guaranteed if the algebraic dynamics α^t is defined as a weak limit of approximate (finite volume) automorphisms α^t_V of $\mathcal{A}$, which are α symmetric

$$\alpha\,\alpha^t_V = \alpha^t_V\,\alpha. \tag{3.3}$$

This equation holds if α^t_V is generated by a finite volume Hamiltonian H_V (affiliated to some localized subalgebra $\mathcal{A}_{V'}$), which is invariant under α: $\alpha(H_V) = H_V$. Proposition 3.3 below shows that, in the framework discussed in Sec. II, property (3.3) constrains the algebraic dynamics α^t to be α symmetric

$$\alpha\,\alpha^t = \alpha^t\,\alpha, \quad \text{on } \mathcal{M}. \tag{3.4}$$

It is important to remark that for this result the strict invariance of H_V is required since, in the presence of long-range interactions, boundary terms arising from the transformation of H_V under α may give rise to persistent effects in the infinite volume limit.

Proposition 3.2: If the finite volume dynamics are α symmetric

$$\alpha\,\alpha^t_V = \alpha^t_V\,\alpha, \tag{3.5}$$

and α^t_V defines an algebraic dynamics α^t as a weak limit with respect to a family of states F, then α^t_V also defines an algebraic dynamics with respect to a family $G \supset F$, which is stable under α^*, α^{-1*}.

Proof: As stated in Sec. II, given α^t_V there is a unique maximal (see Proposition 2.4) set of states F_M with respect to which α^t_V defines an algebraic dynamics α^t. On the other side, by assumption (see Proposition 2.3, first part) $\forall A \in \mathcal{A}$, $(\alpha^{t*}_V\phi)(A)$ converges to $\phi_t(A)$, $\phi_t \in F$, uniformly on the weakly compact sets of F. Therefore

$$(\alpha^{t*}_V\,\alpha^*\phi)(A) = (\alpha^*\,\alpha^{t*}_V\phi)(A) = (\alpha^{t*}_V\phi)(\alpha(A))$$

converges to $\phi_t(\alpha(A)) = (\alpha^*\phi_t)(A)$, $\alpha^*\phi_t \in \alpha^*F$, since α is an automorphism of $\mathcal{A}$, i.e., $(\alpha^{t*}_V\chi)(A)$ converges in α^*F, whenever $\chi \in \alpha^*F$.

Since α is invertible, α^{**} defines an isomorphism of the Von Neumann algebras $\overline{\mathcal{A}}^F$ and $\overline{\mathcal{A}}^{\alpha^*F}$ and therefore α^* defines a one-to-one map between weakly compact sets of F and weakly compact sets of α^*F. In conclusion $(\alpha^{t*}_V\chi)(A)$ converges uniformly on the weakly compact sets of α^*F and α^t_V defines an algebraic dynamics also with respect to α^*F. Thus F_M contains α^*F_M; similarly F_M contains $\alpha^{-1*}F_M$ and therefore

$$\alpha^*F_M = F_M. \tag{3.6}$$

Proposition 3.3: If the finite volume Hamiltonians H_V are α symmetric,

$$\alpha\,\alpha^t_V = \alpha^t_V\,\alpha, \tag{3.7}$$

the algebraic dynamics α^t defined by α^*-stable family of states F is symmetric,

$$\alpha\,\alpha^t = \alpha^t\,\alpha. \tag{3.8}$$

Proof: Since F is α^* stable, α is weakly continuous on $\mathcal{M} \equiv \overline{\mathcal{A}}^F$ (see Proposition 3.1). Since α_t is the weak limit of α^t_V, for any $A \in \mathcal{A}$ one has

$$\alpha\,\alpha^t(A) = \alpha\ \text{w-}\lim_V \alpha^t_V(A) = \text{w-}\lim_V \alpha\,\alpha^t_V(A)$$

$$= \text{w-}\lim_V \alpha^t_V\,\alpha(A) = \alpha^t\alpha(A).$$

Now, both α and α^t are defined and weakly continuous on $\mathcal{M}$ so that by weak continuity the above equation holds on $\mathcal{M}$.

The space translations can be treated in a similar way. We assume that the space translations define a group of automorphisms $\alpha_{\mathbf{x}}$ of $\mathcal{A}$ and that the finite volume dynamics α^t_V is covariant under space translations

$$\alpha_{\mathbf{x}}\,\alpha^t_V\,\alpha_{\mathbf{x}}^{-1} = \alpha^t_{V+\mathbf{x}}. \tag{3.9}$$

The infinite volume dynamics α^t_V will turn out to be symmetric under space translations if the limit of $\alpha^t_{V_n}$ is independent of the sequence of volumes V_n, $V_n \to \infty$, within a class of sequences which is stable under space translations. Under this assumption the family F of relevant states can be chosen to be stable under space translations (see Proposition 3.2) and therefore $\alpha_{\mathbf{x}}$ has a unique extension to a group of automorphisms of $\mathcal{M}$ which commute with α^t (see Proposition 3.3),

$$\alpha_{\mathbf{x}}\,\alpha^t = \alpha^t\,\alpha_{\mathbf{x}}. \tag{3.10}$$

An automorphism α of the C^*-algebra $\mathcal{A}$ is said to be *broken* in the representation π if the representation $\pi \circ \alpha$ is not equivalent to π. When α can be extended to $\mathcal{M}$, see Proposition 4.1, α is broken in π iff it does not leave the corresponding central projection $E_\pi \in \mathcal{M}$ stable. Equivalently α is broken in π iff α is not continuous with respect to the weak topology defined by the states of π (this follows from the last part of Proposition 3.1 with $F^+ =$ states of π).

IV. SYMMETRIES GENERATED BY "LOCAL" CHARGES

A. Local approximation of symmetries and nonlocal algebraic dynamics

In this section we shall consider symmetries β of the algebra $\mathcal{A}$, which can be approximated by "localized" automorphism β_R of $\mathcal{A}_0$ of the form

$$\beta_R(A) = U_R\,A\,U_R^{-1}, \tag{4.1}$$

with U_R unitary and belonging to $\mathcal{A}_0$, such that

$$\beta(A) = \lim_{R\to\infty} \beta_R(A). \tag{4.2}$$

The existence of the limit on $\mathcal{A}_0$ in the weak topology defined by F is enough. Actually, by norm continuity Eq. (4.2) extends to $\mathcal{A}$ and weak convergence on $\mathcal{A}_0$ implies weak convergence on $\mathcal{A}$ and therefore ultrastrong convergence on $\mathcal{A}$, since β is an automorphism of $\mathcal{A}$.

For concreteness, in most examples $\beta_R = \beta$ on the algebras localized within spheres of radius R. In this case, since $\mathcal{A}$ = norm closure of $\cup_V \mathcal{A}_V$, β_R is actually norm converging on $\mathcal{A}$.

In the case of a continuous one-parameter group of symmetries β^λ, $\lambda \in R$, in most cases one may construct a localized automorphism β^λ_R approximating β^λ, by taking

$$U^\lambda_R = \exp(iQ_R\,\lambda), \tag{4.3}$$

with $Q_R = Q^*_R$ affiliated to some $\mathcal{A}_V$ (briefly affiliated to $\mathcal{A}_0$). In this case one has on a norm dense set of vectors in $\mathcal{H}_F$ (see Proposition 2.1)

$$-i\frac{d}{d\lambda}\langle\Psi, U^\lambda_R\,A(U^\lambda_R)^{-1}\Phi\rangle\Big|_{\lambda=0}$$
$$= \langle Q_R\,\Psi, A\Phi\rangle - \langle A\Psi, Q_R\,\Phi\rangle. \tag{4.4}$$

With an abuse of notation in the following we shall write the rhs of Eq. (4.4) as $\langle\Psi, [Q_R, A]\Phi\rangle$.

Definition 4.1: A one-parameter continuous group of symmetries β^λ, $\lambda \in R$ is generated by local charges Q_R, affiliated to $\mathcal{A}_0$, on a weakly dense algebra $\mathcal{A}_l \subset \mathcal{M}$, in the state ϕ, if $\forall A \in \mathcal{A}_l$,

$$\frac{d}{d\lambda}\phi(\beta^\lambda(A))\Big|_{\lambda=0} = i \lim_{R\to\infty} \phi([Q_R, A]). \tag{4.5}$$

A crucial hypothesis for the proof of the (standard) Goldstone Theorem about spontaneously broken symmetries in a given representation is the validity of Eq. (4.5) on the ground state for an *algebra $\mathscr{A}$, stable under time translations*. In the standard case of strictly local dynamics, it is enough to have β^λ generated by Q_R on the local algebra $\mathscr{A}_0$, since $\mathscr{A}_0$ is stable under time evolution. For systems with nonlocal algebraic dynamics it is by no means guaranteed that β^λ is generated by Q_R on an algebra stable under time evolution; actually in most cases this property cannot hold as the following arguments show.

Proposition 4.2: If the one-parameter group of automorphisms $\beta^\lambda, \lambda \in R$ is spontaneously broken in some representation of the family F, then one cannot have

$$\beta^\lambda(A) = \text{w-}\lim_{R\to\infty} e^{iQ_R\lambda} A e^{-iQ_R\lambda}, \tag{4.6}$$

with Q_R affiliated to $\mathscr{M}$, for all $A \in \mathscr{M}$. More generally, given a subalgebra $\mathscr{B} \subset \mathscr{M}$, stable under β^λ, with a center Z which is not pointwise invariant under β^λ, Eq. (4.6) cannot hold on $\mathscr{B}$ and β^λ cannot be generated (Definition 4.1) by local charges on $\mathscr{B}$, in any state in which $\phi(\beta^\lambda(z)) \neq \phi(z)$, for some $z \in Z$ and some $\lambda \in R$.

Proof: Clearly Eq. (4.6) implies $\beta^\lambda(z) = z$, $\forall z$ in the center of $\mathscr{M}$, which implies β^λ unbroken. The same is true for $\mathscr{B}$: if β^λ is generated by local charges on $\mathscr{B}$, then

$$\frac{d}{d\lambda}\phi(\beta^\lambda(z)) = 0, \quad \forall \lambda,$$

since $\beta^\lambda(Z) \subset Z$.

Remark: It may be useful to note that if $\beta_R(A)$ in Eq. (4.1) converges as $R \to \infty$ on a weakly dense subalgebra $\mathscr{C} \subset \mathscr{M}$, to a weakly continuous automorphism γ of $\mathscr{C}$, then (see the proof of Proposition 4.1) γ has a unique extension to $\mathscr{M}$ and, if $\mathscr{C} \supset \mathscr{A}_0$, γ coincides with β on $\mathscr{M}$, since $\gamma = \beta$ on $\mathscr{A}_0$ (Proposition 3.1). In particular, if Eq. (4.1) is used to define an automorphism of $\mathscr{M}$, this is completely determined by Eq. (4.1) on $\mathscr{A}_0$.

The above Proposition rules out the possibility of approximating β^λ by (local) charges [Eq. (4.6)] on $\mathscr{M}$, the obviously stable algebra under time evolution. Actually, when variables at infinity get involved in the time evolution of elements of $\mathscr{A}_0$, it is impossible to have an algebra $\mathscr{B} \subset \mathscr{M}$ stable under time evolution and such that the (unique) weakly continuous extension of β^λ from $\mathscr{A}$ to $\mathscr{B}$ is generated by local charges, except, of course, the uninteresting case in which β^λ is unbroken. Thus the nonlocality of the algebraic dynamics, in the precise sense of Sec. II, provides a natural mechanism for evading the existence of Goldstone modes in the presence of spontaneous symmetry breaking: essentially the equation

$$\delta(A_t) = \lim_R [Q_R, A_t] \tag{4.7}$$

does not hold.

A generalization of Proposition 4.2, which exploits the topological aspects of the phenomenon, shows that mass gap generation in the presence of spontaneous symmetry breaking can be seen as the consequence of rather simple and general algebraic structures. In particular, it will become clear that assumptions like existence of a "local" conserved current generating the symmetry, validity of the infinitesimal form on suitable domains and even the existence of local charges, in the sense of Eq. (4.5), are not the relevant points for the phenomenon. As emphasized before, the important issue is the proper definition of α^t. (In particular one cannot discuss the assumptions at the basis of the rigorous proof of Goldstone's Theorem, in more general cases than the strictly local one, without facing this problem.)

An essential difference between local and nonlocal algebraic dynamics is given by the continuity properties of α^{t*}.

Proposition 4.3: Given the structure $(\mathscr{A}, F, \mathscr{M})$ (see Sec. III) α^{t*} is continuous on F with respect to the weak * topology induced on F by $\mathscr{A}$, iff α^t leaves $\mathscr{A}$ stable. More generally, α^{t*} is continuous on F in the weak * topology $\tau_{\mathscr{B}}$ defined on F by a subalgebra $\mathscr{B}$ of $\mathscr{M}$ iff α^t leaves $\mathscr{B}$ stable.

Proof: It suffices to prove the second part. Clearly if α^t leaves $\mathscr{B}$ stable then α^t maps weak * seminorms p_B on F, defined by elements B of $\mathscr{B}$, into themselves:

$$p_B(\alpha^{t*}\phi) \equiv |(\alpha^{t*}\phi)(B)| = |\phi(\alpha^t(B))| = p_{\alpha^t(B)}(\phi). \tag{4.8}$$

Thus α^{t*} is $\tau_{\mathscr{B}}$ continuous.

Conversely, if α^{t*} is $\tau_{\mathscr{B}}$ continuous, then $\forall B \in \mathscr{B}$,

$$f_{B,t}(\phi) \equiv \phi(\alpha^t(B)) = (\alpha^{t*}\phi)(B)$$

as a composition of continuous functions defines a $\tau_{\mathscr{B}}$ continuous linear functional on F. Therefore given $B \in \mathscr{B}$, there exist $B_1, \ldots, B_n \in \mathscr{B}$ such that

$$|f_{B,t}(\phi)| < \sup_i |\phi(B_i)|, \quad \forall \phi \in F.$$

By a standard argument,[12] this implies that $f_{B,t}(\phi)$ is of the form $\phi(\sum_{i=1}^n c_i B_i)$. In fact the above bound implies $f_{B,t}(\phi) = 0$ if $\phi(B_i) = 0$, $i = 1,\ldots,n$. Hence $\phi_1(B_i) = \phi_2(B_i)$, $\forall i$ implies $f_{B,t}(\phi_1) = f_{B,t}(\phi_2)$, and therefore there exists a linear functional g: $\mathbb{C}^n \to \mathbb{C}$ such that

$$f_{B,t}(\phi) = g(\phi(B_1), \ldots, \phi(B_n)) = \sum_i c_i \phi(B_i) = \phi\left(\sum_i c_i B_i\right). \tag{4.9}$$

In conclusion

$$\phi(\alpha^t(B)) = \phi\left(\sum_i c_i B_i\right), \quad \forall \phi \in F,$$

and since F is separating for $\mathscr{M}$

$$\alpha^t(B) = \sum_i c_i B_i,$$

i.e., α^t leaves $\mathscr{B}$ stable.

Remark: The first part of Proposition 4.3 also follows from Kadison's Theorem 4.5 and Corollary 4.7 in Ref. 11, since F^+ is a full family of states in the sense of Kadison. It may be useful to remark that all the complications in Kadison's proof come from the relation between a dense set of (positive, normalized) states and the space of all continuous

linear functionals on $\mathscr{A}$ with respect to the weak * topology (Lemmas 4.1, 4.2, and 4.3 of Ref. 11).

The above discontinuity properties of α^{t*} explains why in general $\phi(\beta_R^\lambda \alpha^t(A))$ does not converge to $\phi(\beta^\lambda \alpha^t(A))$. All that is needed is that in the state ϕ the symmetry β^λ is approximated on $\mathscr{A}$ by automorphisms β_R^λ of $\mathscr{A}$. [It is enough to assume convergence of $\phi(\beta_R^\lambda(A))$ to $\phi(\beta^\lambda(A))$ for all A in a norm-dense subalgebra of $\mathscr{A}$.] This is equivalent to the weak * convergence of $\beta_R^{\lambda *}\phi$ to $\beta^{\lambda *}\phi$. In general, one cannot expect weak convergence of $\beta_R^{\lambda *}\phi$, i.e., with respect to the weak topology induced by $\mathscr{M}$; weak convergence of $\beta_R^{\lambda *}\phi$ is in fact excluded if β_R^λ is unbroken in the representation defined by ϕ and β^λ is broken (see Proposition 4.2). On the other hand, convergence of $\phi(\beta_R^\lambda \alpha^t(A))$, for all $A \in \mathscr{A}$, is equivalent to weak * convergence of $\alpha^{t*}\beta_R^{\lambda *}\phi$ and α^{t*} is weakly continuous but not weak * continuous, whenever $\mathscr{A}$ is not α^t stable. Then, $\alpha^{t*}\beta_R^\lambda \phi$ does not converge to $\alpha^{t*}\beta^{\lambda *}\phi$, in general. This argument shows that there is a general topological property which prevents the combination of time evolution and the approximation of β^λ by β_R^λ in the way required for Goldstone's Theorem.

The constraints on the energy spectrum, following from spontaneous breaking of continuous symmetries, can be sharply characterized in terms of the time evolution of large bubbles, as $R \to \infty$. To this purpose we consider a one-parameter continuous group of symmetries $\beta^\lambda, \lambda \in R$, approximated by localized automorphisms β_R^λ in the sense of Eqs. (4.1)–(4.3), and $\phi = \langle \Psi, \cdot \Psi \rangle$, a state invariant under space and time translations, with Ψ in the domain of all the operators Q_R. Then, since automorphisms of the form (4.1), with $U_R \in \mathscr{M}$, clearly extended to automorphisms of the Von Neumann algebra $\mathscr{M}$, still given by Eq. (4.1),

$$\phi([Q_R, \alpha^t(A)]) = -i\frac{d}{d\lambda}\phi(\beta_R^\lambda \alpha^t(A))\Big|_{\lambda=0} = -i\frac{d}{d\lambda}(\alpha^{t*}\beta_R^{\lambda *}\phi)(A)\Big|_{\lambda=0}, \quad (4.10)$$

for any A in $\mathscr{A}$, or more generally in $\mathscr{M}$.

The limit $R \to \infty$ of the first term on the left-hand side will be discussed in Sec. VI and related to the low momentum behavior of the energy spectrum. On the right-hand side the limit $R \to \infty$ can be interchanged with $d/d\lambda$, for A in a subalgebra of $\mathscr{M}$, under general technical conditions [see Secs. IV and VI]. Thus the information on the energy spectrum is provided by the function

$$\lim_{R\to\infty} (\alpha^{t*}\beta_R^{\lambda *}\phi)(A). \quad (4.11)$$

For $A \in \mathscr{A}$, the limit (4.11) is given by the weak * limit of the time evolution of large "bubbles" of radius R,[4] as $R \to \infty$. In the case of local algebraic dynamics such limit exists and it is independent of time. By Eq. (4.10) this implies the appearance of $\omega = 0$ in the point spectrum of the energy of excitations at $\mathbf{k} = 0$. When the algebraic dynamics is nonlocal, i.e., $\mathscr{A}$ is not α^t stable, the limit (4.11) is not in general given by $\beta^*\phi$, since α^{t*} is not weak * continuous. The time dependence of the limit (4.11) (when it exists) is responsible for an energy gap at low momenta.

The above discussion shows that the generalization of Goldstone's Theorem to the nonlocal case requires the control of the limit (4.11).

B. Effective localization of the dynamics and local approximation of symmetries

The occurrence of variables at infinity in the time evolution of local variables precludes the possibility of a local approximation of symmetries on an algebra stable under time evolution. However, since in each factorial representation π of $\mathscr{A}$, stable under time evolution, the variables at infinity become time independent c-numbers, it is reasonable to expect that the representation π defines a reduced algebraic dynamics α_π^t which leaves stable an "essentially local" algebra, i.e., a subalgebra $\mathscr{A}_l$ of $\mathscr{M}$ that does not contain infinitely delocalized variables. Such effective localization of the dynamics can be formalized in the following way. First we introduce the following.

Definition 4.4: We shall say that the *algebraic dynamics* is *essentially local* if there exists a subalgebra $\mathscr{A}_l \subset \mathscr{M}$, called *algebra of essential localization* (without loss of generality $\mathscr{A}_l$ can be taken closed in the norm topology since α^t is continuous in such topology) with the following properties:

(a) $\mathscr{A}_l$ has a trivial center,

(b) $\mathscr{A}_l$ is weakly dense in $\mathscr{M}$,

(c) $\mathscr{A}_l$ is stable under α^t.

Clearly the above definition tries to abstract the relevant structure properties associated to local dynamics: in fact, in the constructive approach discussed in Sec. III, if the interaction is local the algebra (e.g., of local canonical variables) $\mathscr{A}_0$ is stable under α^t and one can take $\mathscr{A}_l = \mathscr{A}_0$.

Definition 4.5: The algebraic dynamics is said to be *essentially nonlocal* if every subalgebra of $\mathscr{M}$ which is stable under α^t and weakly dense in $\mathscr{M}$ has a nontrivial center. This is always the case if the dynamics involves variables at infinity.

The above definitions can be reformulated in terms of continuity properties of α^{t*}. Since $\mathscr{A}_l \subset \mathscr{M}$, $\mathscr{A}_l$ is faithfully represented by F and therefore by Proposition 4.3 $\mathscr{A}_l$ is α^t stable iff α^{t*} is continuous with respect to the weak * topology defined by $\mathscr{A}_l$ on F. Moreover, one can show that $\mathscr{A}_l$ is weakly dense in $\mathscr{M}$ iff the weak * topology defined by $\mathscr{A}_l$ on F separates the points. Hence the dynamics is essentially local iff α^{t*} is continuous with respect to some weak Hausdorff topology defined by a subalgebra $\mathscr{A}_l \subset \mathscr{M}$, with trivial center. Thus, the "essential nonlocality" of the dynamics corresponds to the discontinuity of α^{t*} with respect to all weak Hausdorff topologies defined by subalgebras of $\mathscr{M}$ with trivial center.

Definition 4.6: A factorial representation π of the family F, stable under α^{t*} leads to an *effective localization of the dynamics* if there exists a subalgebra $\mathscr{A}_l \subset \mathscr{M}$, called *algebra of effective localization* with the following properties: (i) $\mathscr{A}_l$ is faithfully represented by π, (ii) $\mathscr{A}_l$ is weakly dense in $\mathscr{M}$, and (iii) there exists an automorphism α_π^t of $\mathscr{A}_l$, which coincides with α^t in the representation π, namely, for any state $\phi \in \pi$,

$$\phi(\alpha^t(A)) = \phi(\alpha_\pi^t(A)), \quad \forall A \in \mathscr{A}_l. \quad (4.12)$$

Since $\mathscr{A}_l$ is weakly dense in $\mathscr{M}$, $\pi(\mathscr{A}_l)$ is weakly dense in $\pi(\mathscr{M})$ and therefore the center of $\pi(\mathscr{A}_l)$ is contained in the center of $\pi(\mathscr{M})$ which is trivial because π is factorial. [If $z \in \pi(\mathscr{A}_l) \cap \pi(\mathscr{A}_l)'$, clearly $z \in \pi(\mathscr{M})$, and moreover z commutes with $\pi(\mathscr{M})$ since it commutes with a weakly dense subalgebra.] Since $\mathscr{A}_l$ is faithfully represented by $\pi(\mathscr{A}_l)$, it follows that $\mathscr{A}_l$ has a trivial center. Thus condition (i) is the analog of property (a), in Definition 4.4, and in particular $\mathscr{A}_l$ does not contain any infinitely delocalized variable.

Furthermore, since π yields a faithful representation of $\mathscr{A}_l$, the automorphism α^t_π is unique, if it exists.

The physical meaning of this structure is that the essential nonlocal effects of the algebraic dynamics are due to the involvement of the variables at ∞; once such variables are frozen to c-numbers, as happens with the choice of a factorial representation π, then one obtains a dynamics which maps a complete set $\mathscr{A}_l$ of "essentially local variables" into themselves.

More generally a family of factorial representations $\{\pi_\alpha\}$, each stable under time evolution, effectively localizes the dynamics on $\mathscr{A}_l$ if $\mathscr{A}_l$ is weakly dense in $\mathscr{M}$, it is faithfully represented by each π_α, and for any π_α there exists an automorphism $\alpha^t_{\pi_\alpha}$ that satisfies condition (iii).

The connection with the essentially local case is given by the following.

Proposition 4.7: If the algebraic dynamics is essentially local, with localization algebra $\mathscr{A}_l$, then any factorial representation π, stable under α^{t*}, which faithfully represents $\mathscr{A}_l$, effectively localizes the dynamics.

The above algebraic structure offers a convenient mathematical framework for the generalization of the Goldstone Theorem: the effective localization algebra $\mathscr{A}_l$ is in fact a natural algebra on which symmetries may be locally approximated (see Proposition 4.2) and furthermore it is stable under a reduced dynamics, α^t_π, which coincides with α_t in π.

Proposition 4.8: Let ϕ be a state of a factorial representation π, which effectively localizes the dynamics on an algebra $\mathscr{A}_l \subset \mathscr{M}$ (see Definition 4.6), and β an automorphism of $\mathscr{M}$ which is weakly approximated on $\mathscr{A}_l$ by automorphisms β_R of $\mathscr{A}_l$, which are not broken in π. Then for $A \in \mathscr{A}_l$,

$$\lim_{R\to\infty} \phi(\beta_R \alpha^t(A)) = \phi(\beta \alpha^t_\pi(A)). \tag{4.13}$$

Proof: Since β_R is not broken in π, the state $\beta^*_R \phi$ still belongs to π and therefore by Eq. (4.12)

$$\begin{aligned}\phi(\beta_R \alpha^t(A)) &= (\beta^*_R \phi)(\alpha^t(A)) \\ &= (\beta^*_R \phi)(\alpha^t_\pi(A)) = \phi(\beta_R\, \alpha^t_\pi(A)).\end{aligned}$$

Now, since $\mathscr{A}_l$ is α^t_π stable and β_R converges weakly to β on $\mathscr{A}_l$, one has

$$\lim_{R\to\infty} \phi(\beta_R \alpha^t_\pi(A)) = \phi(\beta \alpha^t_\pi(A)).$$

Clearly all that is needed for the proof is that β_R converges weakly to β on $\mathscr{A}_l$ in the representation π. Equation (4.13) can also be written in the form

$$\lim_{R\to\infty} \alpha^{t*} \beta^*_R \phi = \alpha^{t*}_\pi \beta^* \phi \tag{4.14}$$

as an equation for states on the algebra $\mathscr{A}_l$, and the limit is taken in the weak * topology defined by $\mathscr{A}_l$. In the logic of the discussion in Sec. IV, (ii), the point is that α^{t*}_π is continuous in the weak * topology in which $\beta^*_R \phi$ converges to $\beta^* \phi$, whereas α^{t*} is not. By exploiting the effective localization of the dynamics in the representation π, one can therefore compute the weak * limit of the time evolution of bubbles as $R \to \infty$. Since α^t_π is in general different from α^t, it does not in general commute with β. As a result, (1) the time evolution of large bubbles depends on the representation defined by the state ϕ, typically by the behavior at infinity of the states of π; (2) a nontrivial time evolution may persist in the limit $R \to \infty$ and it can be computed in terms of α^t_π; and (3) the relation between α^t_π and β becomes a relevant step for the generalization of Goldstone's Theorem.

V. EFFECTIVE LOCALIZATION OF THE DYNAMICS AND SYMMETRIES

In this section we discuss the relation between α^t, α^t_π and the automorphisms of $\mathscr{M}$. In the following, we always consider an algebraic dynamics α^t, a (factorial) representation π which effectively localizes the dynamics on $\mathscr{A}_l$ (Definition 4.6) and a one-parameter group of automorphisms β^λ of $\mathscr{A}_l$ with the following properties:

(I) $\beta^\lambda \alpha^t = \alpha^t \beta^\lambda$,

(II) β^λ leaves $\mathscr{A}_l$ stable.

We start by listing some characteristic features of the *effective dynamics* α^t_π.

Proposition 5.1: If the representation π localizes the dynamics on $\mathscr{A}_l$ with α^t_π the corresponding time automorphism, then also the representations π_λ defined by

$$\pi_\lambda(\cdot) = \pi(\beta^\lambda(\cdot))$$

localize the dynamics on $\mathscr{A}_l$ with time automorphism given by

$$\alpha^t_{\pi_\lambda} = (\beta^\lambda)^{-1} \alpha^t_\pi \beta^\lambda.$$

Proof: Since β^λ is an automorphism of $\mathscr{A}_l$, if π is a faithful representation of $\mathscr{A}_l$, so is π_λ and by Eq. (I), π_λ is stable under α^{t*}. By assumption $\mathscr{A}_l$ is β^λ and α^t_π stable and therefore it is also $\alpha^t_{\pi_\lambda}$ stable. Furthermore, in the representation π_λ, $\alpha^t_{\pi_\lambda}$ reduces to α^t since $\forall\, A \in \mathscr{A}_l$,

$$\begin{aligned}\pi_\lambda(\alpha^t(A)) &= \pi(\beta^\lambda \alpha^t(A)) \\ &= \pi(\alpha^t \beta^\lambda(A)) = \pi(\alpha^t_\pi\, \beta^\lambda(A)) \\ &= \pi_\lambda((\beta^\lambda)^{-1} \alpha^t_\pi\, \beta^\lambda(A)),\end{aligned}$$

where in the second equality we have used the commutation $\alpha^t \beta^\lambda = \beta^\lambda \alpha^t$, and in the third equality we have used the stability of $\mathscr{A}_l$ under β^λ.

As we have already discussed, the crucial property for the phenomenon of mass generation associated to spontaneous symmetry breaking is that the effective dynamics is not invariant under β^λ, i.e., on $\mathscr{A}_l$, $\beta^\lambda\, \alpha^t_\pi \neq \alpha^t_\pi\, \beta^\lambda$. Clearly, this is possible only if $\alpha^t_\pi \neq \alpha^t$. The case in which $\alpha^t = \alpha^t_\pi$ can actually be considered as equivalent to the standard case and the usual approach to Goldstone's Theorem can be used in this case.

Therefore a necessary condition for a departure from the standard Goldstone Theorem is that $\alpha^t_\pi \neq \alpha^t$.

Proposition 5.2: The stability of $\mathscr{A}_l$ under α^t is a necessary and sufficient condition for $\alpha^t_\pi = \alpha^t$.

Proof: Obviously, α^t and α^t_π are both defined as maps from $\mathscr{A}_l$ to $\mathscr{M}$; since α^t_π leaves $\mathscr{A}_l$ stable, $\alpha^t = \alpha^t_\pi$ implies α^t stability of $\mathscr{A}_l$. Conversely if $\mathscr{A}_l$ is α^t stable, α^t and α^t_π leave $\mathscr{A}_l$ stable and they coincide in a representation π which is a faithful representation of $\mathscr{A}_l$; hence they coincide on $\mathscr{A}_l$.

Remark: Clearly, the condition $\alpha^t_\pi \neq \alpha^t$ is β^λ covariant in the sense that $\alpha^t_\pi \neq \alpha^t$ implies

$$\alpha^t_{\pi_\lambda} \neq \alpha^t$$

for any other representation π_λ (see Proposition 5.2). In fact, if for some λ, $\alpha^t_{\pi_\lambda} = \alpha^t$, then, by using the definition of $\alpha^t_{\pi_\lambda}$ and the commutation $\beta^\lambda \alpha^t = \alpha^t \beta^\lambda$, one also gets $\alpha^t_\pi = \alpha^t$.

The following proposition deals with the relation between $\beta^\lambda \alpha^t_\pi \neq \alpha^t_\pi \beta_\lambda$ and the spontaneous breaking of β^λ in the representation π.

Proposition 5.3: If β^λ is unbroken in a representation π which localizes the dynamics on $\mathscr{A}_l$, then

$$\beta^\lambda \alpha^t_\pi = \alpha^t_\pi \beta^\lambda;$$

as automorphisms of $\mathscr{A}_l$, equivalently, if

$$\beta^\lambda \alpha^t_\pi \neq \alpha^t_\pi \beta^\lambda,$$

then β^λ is broken in π.

Proof: Unbroken β^λ means that $\pi \circ \beta^\lambda$ is quasiequivalent to π and therefore, for all states ϕ of $\pi \circ \beta^\lambda$,

$$\phi(\alpha^t(A)) = \phi(\alpha^t_\pi(A)), \quad \forall\, A \in \mathscr{A}_l.$$

Then, $\forall\, A \in \mathscr{A}_l$,

$$\begin{aligned}\pi(\alpha^t_\pi \beta^\lambda(A)) &= \pi(\alpha^t \beta^\lambda(A)) \\ &= \pi(\beta^\lambda \alpha^t(A)) = \pi(\beta^\lambda \alpha^t_\pi(A)),\end{aligned}$$

and since $\mathscr{A}_l$ is faithfully represented by π,

$$\alpha^t_\pi \beta^\lambda = \beta^\lambda \alpha^t_\pi \quad \text{on} \quad \mathscr{A}_l.$$

Corollary 5.4: If the space translations $\alpha_{\mathbf{x}}$ define automorphisms of $\mathscr{M}$ which commute with α^t, and $\mathscr{A}_l$ is $\alpha_{\mathbf{x}}$ stable, then if $\alpha_{\mathbf{x}}$ is unbroken

$$\alpha^t_\pi \alpha_{\mathbf{x}} = \alpha_{\mathbf{x}} \alpha^t_\pi.$$

In conclusion we have shown that the condition

$$\beta^\lambda \alpha^t_\pi \neq \alpha^t_\pi \beta^\lambda,$$

which will be at the basis of the discussion in the next sections, implies (α) β^λ is spontaneously broken, and (β) $\alpha^t_\pi \neq \alpha^t$ or equivalently $\mathscr{A}_l$ is not α^t stable. The discussion can be made sharper by introducing the following.

Definition 5.4: Given a dynamical system $(\mathscr{M}, \alpha^t)$, *the algebraic dynamics is essentially local,* with localization algebra $\mathscr{A}_l$, with respect to the family of states $S \subset F$, which is norm closed and stable under $\mathscr{A}$, if there is a weakly dense subalgebra $\mathscr{A}_l \subset \mathscr{M}$, with trivial center and stable under α^t in the representation space $\mathscr{H}_S = P_S \mathscr{H}$ (where P_S is the central projection corresponding to the states S, see Proposition 2.1), i.e., α^t leaves $P_S \mathscr{A}_l$ stable.

Proposition 5.5: Given a dynamical system $(\mathscr{M}, \alpha^t)$ and a representation π which effectively localizes the dynamics on $\mathscr{A}_l$ (localization algebra), let S denote the family of states of the form $\pi_\lambda = \pi \circ \beta^\lambda$ and P_S the corresponding projection, then the following conditions are equivalent:

(1) $\beta^\lambda \alpha^t_\pi \neq \alpha^t_\pi \beta^\lambda$, for some λ,

(2) $P_S\, \alpha^t\, \mathscr{A}_l \neq P_S\, \alpha^t_\pi\, \mathscr{A}_l$,

(3) the algebraic dynamics is not essentially local, with localization algebra $\mathscr{A}_l$, with respect to the family of states S.

Proof: We now show that (1) is equivalent to (2). In fact, if $\beta^\lambda \alpha^t_\pi = \alpha^t_\pi \beta^\lambda$, for any π_λ we get

$$\begin{aligned}\pi_\lambda(\alpha^t_\pi(\mathscr{A}_l)) &= \pi_\lambda(\beta^{-\lambda} \alpha^t_\pi \beta^\lambda(\mathscr{A}_l)) \\ &= \pi(\alpha^t_\pi \beta^\lambda(\mathscr{A}_l)) = \pi(\alpha^t \beta_\lambda(\mathscr{A}_l)) \\ &= \pi(\beta^\lambda \alpha^t(\mathscr{A}_l)) = \pi_\lambda(\alpha^t(\mathscr{A}_l)).\end{aligned}$$

Thus

$$P_S\, \alpha^t_\pi\, \mathscr{A}_l = P_S\, \alpha^t\, \mathscr{A}_l,$$

i.e., (2) implies (1). On the other hand, if (1) holds, since both β^λ and α^t_π leave $\mathscr{A}_l$ stable and π yields a faithful representation of $\mathscr{A}_l$, we have

$$\begin{aligned}&\pi(\beta^\lambda \alpha^t_\pi \beta^{-\lambda}(\mathscr{A}_l)) \\ &\quad \neq \pi(\alpha^t_\pi(\mathscr{A}_l)) = \pi(\alpha^t(\mathscr{A}_l)),\end{aligned}$$

and therefore

$$\begin{aligned}&\pi_\lambda(\alpha^t_\pi \beta^{-\lambda}(\mathscr{A}_l)) \\ &\quad \neq \pi_\lambda(\beta^{-\lambda} \alpha^t(\mathscr{A}_l)) = \pi_\lambda(\alpha^t \beta^{-\lambda}(\mathscr{A}_l)),\end{aligned}$$

i.e.,

$$\alpha^t \neq \alpha^t_\pi \quad \text{on} \quad P_{\pi_\lambda} \mathscr{A}_l,$$

for any λ such that (1) holds. Hence (1) implies (2). To complete the proof it is enough to show that (2) is equivalent to (3). In fact, if (3) does not hold, $P_S \mathscr{A}_l$ is α^t stable (Definition 5.4); moreover, $P_S \mathscr{A}_l$ is faithfully represented in π because $\mathscr{A}_l \to P_S \mathscr{A}_l \to P_\pi \mathscr{A}_l$ are morphisms and $\mathscr{A}_l \to P_\pi \mathscr{A}_l$ is an isomorphism. Hence

$$\begin{aligned}P_\pi \alpha^t P_S\, \mathscr{A}_l &= \alpha^t P_\pi\, \mathscr{A}_l \\ &= \alpha^t_\pi P_\pi\, \mathscr{A}_l = P_\pi \alpha^t_\pi\, \mathscr{A}_l\end{aligned}$$

implies

$$\alpha^t P_S\, \mathscr{A}_l = \alpha^t_\pi\, \mathscr{A}_l,$$

and, (2) $\to$ (3). Conversely, if (2) does not hold, i.e., $P_S\, \alpha^t\, \mathscr{A}_l = P_S\, \alpha^t_\pi\, \mathscr{A}_l$ and therefore

$$\begin{aligned}\alpha^t P_S\, \mathscr{A}_l &= P_S\, \alpha^t\, \mathscr{A}_l \\ &= P_S \alpha^t_\pi \mathscr{A}_l \subset P_S\, \mathscr{A}_l,\end{aligned}$$

and (3) cannot hold.

VI. CHARGE COMMUTATORS AND ENERGY SPECTRUM AT LOW MOMENTA

In view of our interest in clarifying the conditions under which spontaneous symmetry breaking occurs together with an energy gap above the ground state, it is convenient to reexamine the relation between charge commutators and energy spectrum at zero momentum.

In the rigorous proofs of Goldstone's Theorem a crucial role is played by the local structure of the algebra of fields. The two main points in which locality enters in an essential way is the stability under time evolution of the local algebra *and* the R independence of $\langle [Q_R, A^t] \rangle$ for sufficiently large R.

For systems with long-range interactions, in general one loses locality and the discussions in the literature have been done with the philosophy of giving appropriate substitutes of locality (typically assumptions about the decay of the correlation functions). Unfortunately, this approach prevents from the beginning extensions to the case of an energy gap associated to spontaneous symmetry breaking. The aim of this section is to discuss the relation between energy spectrum and charge commutators independently from any locality assumption, in a way that applies to the energy gap generation.

The relevant condition[4] at the basis of our discussion is the integrability of the charge as a commutator. As a special case, when the vacuum expectation value of the charge commutator is time independent, as follows from locality or from conditions which guarantee a "sufficiently local" structure,[13,14] we obtain the existence of the isolated point $\omega = 0$ in the energy spectrum of excitations at $\mathbf{k} = 0$, without special assumptions on the Fourier transform of charge commutators.[13,15]

We consider a representation π of an algebra $\mathcal{A}_l$ with space and time translations described by strongly continuous unitary operators $U(\mathbf{a})$, $U(t)$ and with a unique space-time translational invariant (vacuum) vector Ψ_0. As in the standard discussion of Goldstone's Theorem we are led to consider the limit $R \to \infty$ of the commutators $[Q_R, A^t]$ where A is an Hermitian operator, $A^t \equiv U(t) A U(t)^{-1}$ and the "local charge" Q_R is essentially the integral of a charge density $j_0(\mathbf{x})$ over a region of size R. In general a time smearing is needed and $j_0(\mathbf{x},t)$ is assumed to be a Hermitian operator-valued distribution of $\mathcal{S}'(R^4)$, which transforms covariantly under space-time translations and has Ψ_0 in its domain.

The relevant condition[4] [condition (A)] is that the charge density commutator

$$i\langle \Psi_0, [j_0(\mathbf{x}), A^t]\Psi_0\rangle = i\langle \Psi_0, [j_0(\mathbf{x}, -t), A]\Psi_0\rangle \equiv J(\mathbf{x},t) \tag{6.1}$$

is a finite measure in the $\mathbf{x}$ variable, i.e., there is a Schwartz seminorm $\| \ \|_{\mathcal{S}}$ such that for $g(t) \in \mathcal{S}(R^1)$, $f(\mathbf{x}) \in \mathcal{S}(R^3)$,

$$|J[fg]| < C \|g\|_{\mathcal{S}} \sup_{\mathbf{x}} |f(\mathbf{x})|.$$

As a consequence the integral of $J(\mathbf{x},t)$ over $\mathbf{x}$ defines a tempered distribution $J(t)$. Actually, as we shall see, without loss of generality one can assume that the above (time smeared) commutator is a C^∞ integrable function. In general, for any sequence $\{f_R\}$, with $0 \leqslant f_R(\mathbf{x}) \leqslant 1$ and $f_R(\mathbf{x}) = 1$ for $|\mathbf{x}| < R$, we have

$$J(t) = \lim_{R\to\infty} \langle [j_0(f_R), A^t]\rangle_0 \tag{6.2}$$

in the sense of distributions [since for any finite measure $\int f_R(x) d\mu(x) \to \int d\mu(x)$].

The crucial role of condition (A) for a rigorous derivation of the nonrelativistic Goldstone Theorem, with a careful handling of the distributional and measure theoretical delicate points, does not seem to have been realized in the vast previous literature.[13,15,16] As will be clear from the following Proposition 6.1 and Corollary 6.2, the derivation of the Goldstone energy spectrum at $k = 0$ (especially in the case of point spectrum) is made possible by this condition. It should be stressed that this condition involves the behavior of the commutator, not of the two-point function.

Proposition 6.1: Within the above framework, if the charge density commutator

$$i\langle \Psi_0, [j_0(\mathbf{x}), A^t]\Psi_0\rangle \equiv J(\mathbf{x},t) \tag{6.3}$$

is a finite measure in $\mathbf{x}$, after smearing in t [condition (A)], then as a tempered distribution in ω

$$i\langle \Psi_0, j_0(f) dE_\omega \, dE_k \, A\Psi_0\rangle - i\,\overline{\langle \Psi_0, j_0(f) dE_{-\omega} dE_{-k} A\Psi_0\rangle} \equiv \tilde{J}^f(k,\omega) \tag{6.4}$$

is a continuous function of k, with

$$\lim_{k\to 0} (2\pi)^3 \tilde{J}^f(k,\omega) = \tilde{J}(\omega), \tag{6.5}$$

where f is any real test function $\in \mathcal{S}(R^3)$, with

$$\int f(\mathbf{x}) d^3x = 1, \tag{6.6}$$

dE_ω, dE_k are the spectral measures associated to the generators of time and space translations and $\tilde{J}(\omega)$ is the Fourier transform of

$$J(t) = i \lim_{R\to\infty} \langle [j_0(f_R), A^t]\rangle.$$

Equations (6.4) and (6.5) imply that for any real $g(t)$

$$J[g] = -2(2\pi)^3 \operatorname{Im}\langle \Psi_0, j_0(f) dE\, [\tilde{\bar{g}}] dE_{k=0} A\Psi_0\rangle, \tag{6.5'}$$

and therefore as distributions on real symmetric test functions $\tilde{g}(\omega)$

$$\tilde{J}(\omega) = -2(2\pi)^3 \operatorname{Im}\langle \Psi_0, j_0(f) dE_\omega \, dE_{k=0} A\Psi_0\rangle. \tag{6.7}$$

Furthermore, if Ψ_0 is the lowest energy state, i.e., $dE_\omega = 0$ for $\omega < 0$ (spectral condition), then on test functions with support in $\omega \geqslant 0$,

$$\tilde{J}(\omega) = i(2\pi)^3 \langle \Psi_0, j_0(f) dE_\omega \, dE_{k=0} A\Psi_0\rangle. \tag{6.8}$$

Proof: We consider the expectation value

$$i\langle j_0(f_{\mathbf{x}})\Psi_0, A^t\Psi_0\rangle = i\langle j_0(f)\Psi_0, U(\mathbf{x}) A^t \Psi_0\rangle \equiv F(\mathbf{x},t),$$

with $f \in \mathcal{S}$, satisfying Eq. (6.6). After time smearing, it is a C^∞ bounded function of $\mathbf{x}$, because f is in $\mathcal{S}$ and $U(\mathbf{x})$ is a unitary operator. By condition (A) the commutator

$$i\langle [j_0(f_{\mathbf{x}}), A^t]\rangle_0 \equiv J^f(\mathbf{x},t)$$

is a C^∞ integrable function of $\mathbf{x}$ and

$$\int d^3x \, J^f(\mathbf{x},t) = \int d^3x \, d^3y \, f(\mathbf{x}+\mathbf{y}) J(\mathbf{y},t) = J(t). \tag{6.9}$$

The Fourier transform of F is therefore a complex measure given by

$$\tilde{F}(k,t) = i\langle j_0(f)\Psi_0, dE_k \, A^t \Psi_0\rangle,$$

and

$$\tilde{F}(k,t) + \overline{\tilde{F}(-k,t)} = \tilde{J}^f(k,t) \tag{6.10}$$

is a continuous function of k, after smearing in t.

Since the spectral resolutions of energy and momentum commute we also have

$$\tilde{J}^f(k,t) = i\int_\omega e^{i\omega t}\left[\langle j_0(f)dE_\omega\, dE_k\, A\rangle - \overline{\langle j_0(f)dE_{-\omega}\, dE_{-k}\, A\rangle}\right],$$

and $\tilde{J}^f(k,\omega)$ is a finite measure in k, and a tempered distribution in ω, and it is continuous in k, when smeared with a $\tilde{g}(\omega)\in\mathcal{S}$. By the estimate in (A), $\tilde{J}^f(k,\omega)$ is actually continuous in k as a distribution in ω and by Eq. (6.9)

$$(2\pi)^3\tilde{J}^f(0,\omega) = \tilde{J}(\omega).$$

By definition $J(t)$ is real and $\tilde{J}(\omega) = \overline{\tilde{J}}(-\omega)$, so that $\tilde{J}(\omega)$ vanishes on the test functions $\tilde{g}(\omega)$ with the property $\tilde{g}(\omega) = -\overline{\tilde{g}}(-\omega)$. On the test functions $\tilde{g}(\omega) = \overline{\tilde{g}}(-\omega)$, Eq. (6.5′) follows from Eq. (6.5). If $\tilde{g}$ is real, Eq. (6.7) follows from Eq. (6.5). Equation (6.8) is an immediate consequence of Eq. (6.5) and the spectral condition.

Corollary 6.2 (Goldstone's Theorem): Under the assumptions of Proposition 6.1, if $J(t)$ is a constant, as it happens if on an algebra stable under time evolution, the charge Q_R generates an automorphism which commutes with time translations, one gets that on real symmetric test functions $\tilde{g}(\omega)$ the imaginary part of the expectation value of the spectral measure dE_ω at $k=0$ is concentrated in $\omega=0$. Such an isolated point in the energy spectrum only arises from contributions of states orthogonal to the ground state.

Proof: It follows easily from Eq. (6.7). A smearing of Eq. (6.4) by real symmetric test functions $h(k)$ and $g(\omega)$ shows that the ground state cannot contribute to $J^f(k,\omega)$:

$$i\langle\Psi_0, j_0(f)dE(h)\, dE(g)\Psi_0\rangle\langle\Psi_0, A\Psi_0\rangle - i\langle\Psi_0, j_0(f)\, dE(h)\, dE(g)\Psi_0\rangle\langle\Psi_0, A\Psi_0\rangle = 0.$$

Remark: When the spectrum of $J(t)$ is discrete, Proposition 6.1 implies that the corresponding points ω_i in the energy spectrum describe (elementary) excitations with a lifetime which becomes infinite in the limit $\mathbf{k}\to 0$.

In the following we will consider automorphisms β^λ, $\lambda\in R$, which are generated on $\mathcal{A}_l$ by local charges Q_R

$$\frac{d}{d\lambda}\phi_0(\beta^\lambda(A))|_{\lambda=0} = i\lim_{R\to\infty}\phi_0([Q_R, A]), \tag{6.11}$$

on a space-time translationally invariant primary state ϕ_0 and Q_R satisfies condition (A).

The following propositions characterize the time dependence of $J(t)$ in terms of symmetry properties of the time evolution of elements of $\mathcal{A}_l$ in the representation π. By Definition 4.6 the effective time evolution, mapping $\mathcal{A}_l$ into $\mathcal{A}_l$, is given by α^t_π and we have the following.

Proposition 6.3: Under the above assumptions, $\forall A\in\mathcal{A}_l$,

$$J(t) = i\lim_{R\to\infty}<[Q_R,\alpha^t(A)]>_{\phi_0} = \frac{d}{d\lambda}<\beta^\lambda\alpha^t_\pi(A)>\Big|_{\lambda=0}. \tag{6.12}$$

Proof: Since in the representation π, $\forall A\in\mathcal{A}_l$,

$$\pi(\alpha^t(A)) = \pi(\alpha^t_\pi(A)) = U(t)\pi(A)U(t)^{-1}, \tag{6.13}$$

we have

$$<[Q_R,\alpha^t(A)]> = <[Q_R,\alpha^t_\pi(A)]>.$$

On the other hand, $\alpha^t_\pi(A)\in\mathcal{A}_l$ and therefore by Eq. (6.11) the limit as $R\to\infty$ exists; by Eq. (6.13) it yields $J(t)$ defined as in Proposition 6.1 and by Eq. (6.11) it is equal to the rhs of Eq. (6.12).

Proposition 6.4: Given $A\in\mathcal{A}_l$, the corresponding $J(t)$ is independent of t (standard Goldstone's Theorem) if and only if

$$\frac{d}{d\lambda}<[\beta^\lambda,\alpha^t_\pi](A)>\Big|_{\lambda=0} = 0. \tag{6.14}$$

Proof: Clearly if Eq. (6.14) holds, then

$$J(t) = \frac{d}{d\lambda}<\beta^\lambda\alpha^t_\pi(A)>\Big|_{\lambda=0} = \frac{d}{d\lambda}<\alpha^t_\pi\beta^\lambda(A)>\Big|_{\lambda=0} = \frac{d}{d\lambda}<\beta^\lambda(A)>\Big|_{\lambda=0}.$$

Conversely, if $J(t)=J(0)$, $\forall t$, then

$$J(t) = \frac{d}{d\lambda}<\beta^\lambda(A)>\Big|_{\lambda=0} = \frac{d}{d\lambda}<\alpha^t_\pi\beta^\lambda(A)>\Big|_{\lambda=0},$$

i.e., Eq. (6.14) holds.

We denote by S the minimal set of (primary) states on $\mathcal{A}_l$ containing ϕ_0 and stable under α^{t*}_π and $\beta^{\lambda *}$. We then consider the representation Π_S of $\mathcal{A}_l$ given by the direct sum of all representations defined by states of S and we denote by Z_S the center of $\Pi_S(\mathcal{A}_l)''$. We further assume that $\mathcal{A}_l$ is a weakly asymptotically Abelian with respect to space translations, in the representation π, i.e., $\forall A, B\in\mathcal{A}_l$

$$\text{w-lim}_{|x|\to\infty}\,\pi[A_x,B] = 0. \tag{6.15}$$

Proposition 6.5: Under the above assumptions (1) $\forall A\in\mathcal{A}_l$, the ergodic limit

$$\lim_{V\to\infty}\frac{1}{V}\int_V\alpha_{\mathbf{x}}(A)d^3x \equiv \lim_V A_V \equiv A_\infty \tag{6.16}$$

exists in the weak topology of Π_S and it belongs to Z_S; (2) the unique extensions of β^λ and α^t_π to $\Pi_S(\mathcal{A}_l)''$ map Z_S into Z_S; and (3) given $A\in\mathcal{A}_l$ and $J(t)$ as in Proposition 6.3,

$$J(t) = \frac{d}{d\lambda}\phi_0(\beta^\lambda\alpha^t_\pi(A_\infty))\Big|_{\lambda=0} \equiv \phi_0(\delta^\lambda\alpha^t_\pi(A_\infty)). \tag{6.17}$$

Proof: To prove (1) we remark that since the space translations are implemented by continuous unitary operators, the ergodic limit

$$\text{w-lim}_{v\to\infty}\frac{1}{V}\int d^3x\,\pi(A_{\mathbf{x}})$$

exists for any $A\in\mathcal{A}_l$ and it commutes with $\pi(\mathcal{A}_l)$ (Ref. 17).

To prove (2) we note that by construction Π_S is stable under $\beta^{\lambda *}$ and α^{t*}_π and therefore, by Proposition 3.1, β^λ and α^t_π have a unique extension to $\Pi_S(\mathcal{A}_l)''$, which is weakly continuous. As automorphisms of $\Pi_S(\mathcal{A}_l)''$ they leave the center Z_S invariant.

To prove Eq. (6.17) it suffices to note that since β^λ and α_π^t are weakly continuous and commute with $\alpha_{\mathbf{x}}$,

$$\begin{aligned}
&\phi_0(\beta^\lambda\alpha_\pi^t(A_\infty)) \\
&\quad = \phi_0(\beta^\lambda\alpha_\pi^t(\lim_V A_V)) \\
&\quad = \lim_V \phi_0(\beta^\lambda\alpha_\pi^t(A_V)) \\
&\quad = \lim_V \frac{1}{V}\int_V d^3x\,\phi_0(\alpha_{\mathbf{x}}\beta^\lambda\alpha_\pi^t(A)) = \phi_0(\beta^\lambda\alpha_\pi^t(A)).
\end{aligned} \tag{6.18}$$

Equation (6.17) reduces the study of the energy spectrum at $\mathbf{k}\to 0$ to the time evolution of the ergodic limit A_∞, when the initial value is infinitesimally close to that of the representation π, in the direction of $\beta^{\lambda *}$. The characterization of the energy spectrum at $\mathbf{k}=0$ simplifies if [condition (B)] given a fixed $A\in\mathscr{A}_l$, and the set $\mathscr{B}=\{\alpha_\pi^\tau(A),\tau\in R\}$, the automorphisms $\beta^\lambda(t)\equiv\alpha_\pi^{-t}\beta^\lambda\alpha_\pi^t$ of $\mathscr{A}_l$ are generated by a finite number of charges on $\mathscr{B}$ in the state ϕ_0 (see Definition 4.1):

$$\begin{aligned}
&\frac{d}{d\lambda}\phi_0(\beta^\lambda(t)\alpha_\pi^\tau(A))\Big|_{\lambda=0} \\
&\quad = i\sum_{i=1}^N a_i(t)\lim_{R\to\infty}\phi_0([Q_R^i,\alpha_\pi^\tau(A)]).
\end{aligned} \tag{6.19}$$

This means that a finite number of expectation values $\lim_{R\to\infty}\phi_0([Q_R(t),\alpha_\pi^\tau(A)])$ are independent as functions of $\tau\in R$. Without loss of generality one can choose $Q_R^i=Q_R(t_i)$, for suitable times $t_1,\ldots,t_N$. We also consider Hermitian A and we then have the following.

Proposition 6.6: Within the framework specified so far, the spectrum of $J(\omega)$ is discrete iff condition (B) holds.

Proof: With the choice $Q_R^i=Q_R(t_i)$, and putting $\beta_i^\lambda\equiv\beta^\lambda(t_i)$, condition (B) gives

$$\frac{d}{d\lambda}\phi_0(\beta_i^\lambda(t)(A))\Big|_{\lambda=0} = \sum_k c_{ik}(t)\frac{d}{d\lambda}\phi_0(\beta_k^\lambda(A))\Big|_{\lambda=0}$$

with

$$c_{ik}(t)=a_k(t_i+t)$$

real, for Hermitian A. Moreover,

$$\begin{aligned}
&\sum_k c_{ik}(t+s)\frac{d}{d\lambda}\phi_0(\beta_k^\lambda(A))\Big|_{\lambda=0} \\
&\quad = \frac{d}{d\lambda}\phi_0(\beta_i^\lambda(s)\alpha_\pi^t(A))\Big|_{\lambda=0} \\
&\quad = \sum_j c_{ij}(s)\frac{d}{d\lambda}\phi_0(\beta_j^\lambda(\alpha_\pi^t(A)))\Big|_{\lambda=0} \\
&\quad = \sum_{j,k} c_{ij}(s)c_{jk}(t)\frac{d}{d\lambda}\phi_0(\beta_k^\lambda(A))\Big|_{\lambda=0},
\end{aligned} \tag{6.20}$$

i.e., $c_{ik}(t)$ is a one-parameter group of real matrices. One can then write $c_{ik}(t)=(\exp(Kt))_{ik}$ with K a real matrix. The spectrum is then discrete and it has no algebraic multiplicities if $J(t)$ is uniformly bounded in t.

Conversely, if the spectrum is discrete then

$$\begin{aligned}
J(\tau-t) &= \lim_{R\to\infty} i\phi_0([Q_R(t),\alpha_\pi^\tau(A)]) \\
&= \sum_k P_k(\tau-t)\exp(-i\omega_k t)\exp[i\omega_k\tau],
\end{aligned} \tag{6.21}$$

with $P_k(\tau-t)$ polynomials and ω_k real. Then as t varies only a finite number of functions of τ, $J(\tau-t)$, are independent. This means that only a finite number of expectation values $\lim\phi_0([Q_R(t),\alpha_\pi^\tau(A)])$ are independent as functions of τ as t varies, i.e., (B) holds.

Remark: In terms of classical motion of variables at infinity, by Proposition 6.5, condition (B) is equivalent to

$$\begin{aligned}
&\frac{d}{d\lambda}\phi_0(\beta^\lambda(t)(\alpha_\pi^\tau(A)))\Big|_{\lambda=0} \\
&\quad = \frac{d}{d\lambda}\phi_0(\beta^\lambda(t)\alpha_\pi^\tau(A_\infty)) \\
&\quad \equiv \phi_0(\delta^\lambda(t)\alpha_\pi^\tau(A_\infty)) = \sum_{i=1}^N a_i(t)\phi_0(\delta^\lambda{}_{(i)}\alpha_\pi^\tau(A_\infty)).
\end{aligned} \tag{6.22}$$

Proposition 6.7: Under the same assumption as in Proposition 6.6 with $J(t)$ uniformly bounded, if the number n of independent charges, for which Eq. (6.19) holds for any $A\in\mathscr{A}_l$ is minimal, then the spectrum of the generator of the group $c_{ik}(t)$ has no multiplicity. If n is even, then $\omega_i\neq 0$ for any i; if n is odd then, for any choice of the basis $Q_R^i\equiv Q_R(t_i)$, the charge

$$\mathscr{Q}_R\equiv\sum_{i=1}^n a_i Q_R(t_i), \tag{6.23}$$

with

$$a_i\equiv\lim_{T\to\infty}\frac{1}{T}\int_0^T c_{1i}(t)dt$$

defines an automorphism α_0^λ of $\mathscr{A}_l$, which is broken on ϕ_0 and which satisfies

$$\phi_0(\alpha_0^\lambda\alpha_\pi^t(A))=\phi_0(\alpha_0^\lambda(A)),\quad \forall A\in\mathscr{A}_l$$

(standard Goldstone case). The charge

$$\tilde{Q}_R\equiv Q_R-\sum_{i=1}^n a_i Q_R(t_i) \tag{6.24}$$

generates an automorphism β^λ which has an energy gap associated to its spontaneous symmetry breaking.

Proof of Proposition 6.7: Given a minimal set of charges, for which Eq. (6.19) holds, by Proposition (6.6) and the uniform boundedness of $J(t)$ one can find n linear combinations Q^p of the previous charges such that

$$\lim_R\phi_0([Q_R(t),A])=\sum_{p=1}^n b_p e^{i\omega_p t}\lim_R\phi_0([Q_R^p,A]). \tag{6.25}$$

Now, if two frequencies, say ω_1,ω_2, are equal, then the rhs of Eq. (6.25) can also be written as

$$\begin{aligned}
&\sum_{p=3}^n b_p e^{i\omega_p t}\lim_R\phi_0([Q_R^p,A]) \\
&\quad + e^{i\omega_1 t}\lim_R\phi_0([b_1Q_R^1+b_2Q_R^2,A]),
\end{aligned}$$

that is Eq. (6.19) holds with $n-1$ charges.

From the symmetry of the spectrum of K, $\omega = 0$ is present iff n is odd. In this case, Eq. (6.23) defines a charge which satisfies

$$\lim_R \phi_0([Q_R(t), A]) = \sum_i \lim_{T\to\infty} \frac{1}{T}\int_0^T dt'\, c_{1i}(t') \lim_R \phi_0([Q_R(t+t_i), A]) = \sum_i \lim_T \frac{1}{T}\int_0^T dt'\, c_{1i}(t')c_{ik}(t) \lim_R \phi_0([Q_R(t_k), A]),$$

and the rhs is independent to t, by the group properties of the matrices c_{ik}.

Finally the last statement of Proposition 6.7 follows from the vanishing of the (time) ergodic mean of $\phi_0([Q_R(t), A])$ by the definition of the a_i's.

ACKNOWLEDGMENT

This paper was supported in part by Istituto Nazionale di Física Nucleare (INFN), Sezione di Pisa.

[1]O. Brattelli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics*, edited by D. Kastler (Springer, Berlin, 1979), Vols. I and II; *C* Algebras and Their Applications to Statistical Mechanics and Quantum Field Theory, Proceedings of the International School "E. Fermi,"* Varenna, 1973 (Soc. Ital. Fis., Bologna, 1976).
[2]D. A. Dubin and G. L. Sewell, J. Math. Phys. **11**, 2990 (1970); M. B. Ruskai, Commun. Math. Phys. **20**, 193 (1971); G. L. Sewell, *ibid.* **33**, 43 (1973).
[3]G. Sewell, Lett. Math. Phys. **6**, 209 (1982).
[4]G. Morchio and F. Strocchi, Commun. Math. Phys. **99**, 153 (1985).
[5]R. Haag, N. M. Hugenholz, and M. Winnink, Commun. Math. Phys. **5**, 215 (1967); D. W. Robinson, *ibid.* **7**, 337 (1968).
[6]R. Haag, Nuovo Cimento **25**, 1078 (1962).
[7]W. Thirring, Commun. Math. Phys. **7**, 181 (1968); *The Many-Body Problem, International School of Physics*, Mallorca, 1969 (Plenum, New York, 1969).
[8]R. Haag, R. V. Kadison, and D. Kastler, Commun. Math. Phys. **16**, 81 (1970).
[9]M. Takesaki, *Theory of Operator Algebras* (Springer, Berlin, 1979), Vol. I.
[10]J. Dixmier, *Les algèbres d'opérateurs dans l'espace hilbertien* (Gauthier–Villars, Paris, 1957), Chap. I, § 4.3.
[11]R. V. Kadison, Topology **3**, 177 (1965).
[12]K. Yoshida, *Functional Analysis* (Springer, Berlin, 1966), 2nd printing, Theorem 8.1, Chap. IV.
[13]J. A. Swieca, Commun. Math. Phys. **4**, 1 (1967).
[14]D. Kastler, *Proceedings of the 1967 International Conference on Particles and Fields*, edited by C. R. Hagen, G. Guralnik, and V. A. Mathur (Interscience, New York, 1967).
[15]J. A. Swieca, Cargèse Lectures, Vol. 4, 1970.
[16]M. Requardt, J. Phys. A: Math. Gen. **13**, 1769 (1980); M. Requardt and W. F. Wreszinski, *ibid.* **18**, 705 (1985).
[17]O. Brattelli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics I* (Springer, Berlin, 1979), pp. 396 and 397.

Transmission coefficients in anharmonic symmetrical potentials

J. A. Caballero Carretero and A. Martín Sánchez
Departamento de Termodinámica, Facultad de Ciencias, Universidad de Extremadura, 06071-Badajoz, Spain

(Received 20 June 1986; accepted for publication 12 November 1986)

Barrier transmission in potentials of the type $V(x) = Ax^2 + Bx^4$ is studied using the phase integral method, the same as the JWKB approximation in lower orders. Elliptic functions are used for the classical solutions. The transmission coefficient is calculated for all signs and values of A and B that give a potential barrier.

I. INTRODUCTION

Formal expressions for the transmission coefficients of potential barriers have been widely studied in the literature (for example; Refs. 1–10). Using the phase integral method, Fröman and Fröman[1] obtain approximate expressions for the transmission coefficients in both the sub-barrier and super-barrier cases.[2] The same treatment is applied by Fröman and Dammert to a system of two real potential barriers. Cramer and Nix[4] study the penetrability through a double peaked fission barrier, taken to be two parabolic peaks connected smoothly with a third parabola forming the intermediate well, and they applied their study to a double asymmetric barrier used to describe direct-reaction fission data; their results are calculated exactly and they use the Jeffreys–Wentzel–Kramers–Brillouin (JWKB) method for comparison. In the textbook of Rapp,[5] several types of barriers are described in order to apply the JWKB approximation to some simple cases; that study includes single and double potential barriers, and the inverted parabolic barrier is used as a particular example of an exactly calculable transmission coefficient. McLaughlin[6] obtains JWKB formulas for the barrier penetration from a path integral, using complex time in his expressions. The same subject was treated by Holstein and Swift,[7] and it was applied to the barrier penetration problem[8] including the semiclassical treatment of above barrier scattering[9] (super-barrier case). Finally, Dammert[10] applied the phase integral method to the transmission through a system of potential barriers. A number of recent very relevant references about the subject can be found in Refs. 11–14.

In this paper we are interested in the determination of the transmission coefficient for potentials of the type

$$V(x) = Ax^2 + Bx^4, \tag{1}$$

when A and B are both negative (simple barrier), with the special cases $A = 0$ (quartic barrier) or $B = 0$ (inverted parabolic barrier); or when A is positive and B is negative (double barrier case). The double-well potential has $A < 0$ and $B > 0$; it presents a barrier between two wells. The calculation of energy levels including tunneling through the intermediate barrier has been reported elsewhere.[15] The case $A > 0$ and $B > 0$ is a well, not a barrier: its energy levels were also studied in Ref. 15.

II. SIMPLE BARRIERS

A. Sub-barrier transmission

When the total energy of the incident particle wave lies below the potential maximum, the transmission coefficient is, in the lowest order of approximation of the JWKB method,

$$T = \exp(-2K_{II}), \tag{2}$$

with

$$K_{II} = \hbar^{-1}\int_{-a}^{a} |p_1(x)|dx \tag{3}$$

and

$$p_1(x) = \{2\mu[V(x) - E]\}^{1/2}, \tag{4}$$

where μ is the mass of the incident particle, a and $-a$ the turning points [real roots of the equation: $V(x) = E$], and E is the total energy.

It is well known that these expressions do not conserve the unitarity relation between the reflection and transmission coefficients. Fröman and Fröman[1] demonstrated that refinements in the approximation using the phase integral method lead to a transmission coefficient

$$T' = [1 + \exp(2K_{II})]^{-1}. \tag{5}$$

This formula conserves the unitarity relation, and gives the same value as T in Eq. (2) when the value of E is negligible compared with the potential maximum.

Let us now apply the preceding relations to calculating transmission in the one-dimensional potential of Fig. 1 with A and B both less than zero. In order to calculate the transmission coefficient, we must first have the classical solutions. Figure 1 shows these for the different regions of the potential, following Díaz Bejarano *et al.*[15–17] in the region below the maximum of the barrier for $E > V(x)$, and Bradbury[18] above the maximum. For the solution under barrier $[V(x) > E]$ we used two properties of our potential and of the elliptic functions: the change $V(x)$ to $-V(x)$ transforms potential barriers into potential wells and vice versa, and because the elliptic functions have real and imaginary periods, it follows that they are the "imaginary classical solutions" in the region under the barrier if we change the parameter m into $m_1 = 1 - m$ and change the sign of the energy.[15] The distinctive mark of this solution is the imagi-

 0022-2488/87/030636-07$02.50

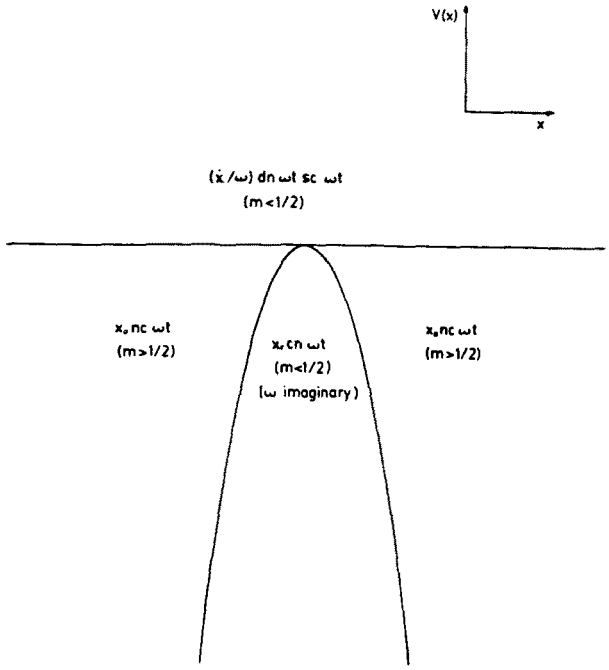


FIG. 1. The generic potential $V(x) = Ax^2 + Bx^4$, with A and B both less than zero, showing classical solutions of the equation of motion, in terms of Jacobi elliptic functions of parameter m, in the distinct regions for different initial conditions. For $E<0$ the initial conditions are $x(0) = x_0$, $\dot{x}(0) = 0$, for $E>0$, they are $x(0) = 0$, $\dot{x}(0) = \dot{x}_0$. There are other solutions in all regions for different types of initial conditions, but only the solutions used in the calculations are given.

nary character of ω as also used by Díaz Bejarano *et al.* in connection with the anharmonic asymmetrical oscillators.[19] All the solutions of this figure are Jacobi elliptic functions[20] with parameter m and complementary parameter m_1 or, alternatively,[21] with parameters $k^2 = m$ and $k'^2 = m_1$. These functions are periodic with generally two complex periods.

In our case (Fig. 1) the solution beneath the barrier (region where we must integrate) is

$$x = x_0 \,\mathrm{cn}\,\omega t \tag{6}$$

with imaginary ω. First ω is calculated as a function of the coefficients A and B, making use of relations between the energy, the coefficients, and the parameter of the elliptic functions[15–17]

$$\omega^2 = 2A/[\mu(1-2m)], \quad x_0^2 = m\mu\omega^2/2B, \qquad E = m_1\mu\omega^2 x_0^2/2. \tag{7}$$

In the quantum treatment, we change Eq. (3) to the more convenient t variable

$$K_{II} = \mu\hbar^{-1}\int_0^{T/2} [\dot{x}(t)]^2\, dt. \tag{8}$$

Using the solution Eq. (6), and by substitution in Eq. (8) we have

$$K_{II} = \hbar^{-1}\mu\omega^2 x_0^2 \int_0^K \mathrm{sn}^2\,\omega t\, \mathrm{dn}^2\,\omega t\, dt, \tag{9}$$

where K is the complete elliptic integral of the first kind.[20] The integral Eq. (9) and similar integrals are tabulated in the Handbook of Byrd and Friedman.[21] After simple algebra, we obtain

$$K_{II} = (-2E/3mm_1\hbar\omega)[(2m_1-1)E' + mK'], \tag{10}$$

where K' and E' are the complete elliptic complementary integrals of the first and second kind, respectively. To obtain Eq. (10) we have used the properties of the potential and the elliptic functions mentioned above.

With Eq. (7) we can transform Eq. (10) into

$$K_{II} = (1/3\hbar)(4\mu^2|E|^3/|B|m^3m_1^3)^{1/4} \times [(2m_1-1)E' + mK'], \tag{11}$$

i.e., K_{II} as a function of m, E, and B, or

$$K_{II} = (|E|/3mm_1\hbar)[2\mu(2m-1)/|A|]^{1/2} \times [(2m_1-1)E' + mK'], \tag{12}$$

i.e., K_{II} as function of m, E, and A.

In this way, the transmission coefficient is totally determined in the sub-barrier case: It depends only on the incident energy (E) the elliptic function parameter (m) and the coefficients (A,B) of the potential.

We now study two particular cases of simple barriers.

First, we consider the quartic potential barrier. Using Eq. (7) and taking into account that $m = m_1 = \frac{1}{2}$ for this case, we obtain from Eq. (11):

$$K_{II} = (2/3\hbar)(|E|^3\mu^2/|B|)^{1/4}K(\tfrac{1}{2}). \tag{13}$$

The second particular case is the inverted parabolic barrier. In this case the parameter of the elliptic functions is $m = 0$, and the solution in the region under study is[20]

$$x = x_0 \cos\omega t \tag{14}$$

but with ω imaginary. In a similar way we obtain a transmission coefficient of the form

$$T' = \{1 + \exp[(2\pi|E|/\hbar)(\mu/2|A|)^{1/2}]\}^{-1}. \tag{15}$$

This is the same as the exact expression given by Rapp.[5] The reason is that our method is a second-order approximation where, for the special case of the inverted parabolic barrier, the treatment is exact.

B. Super-barrier transmission

For energies above the barrier the turning points are complex. A complete detailed study of this case is given by Fröman and Fröman,[1,2] Fröman and Dammert,[3] and Holstein,[8] among others. In the present case of simple barriers, the transmission coefficient is

$$T' = [1 + \exp(2|K|)]^{-1} \tag{16}$$

and

$$K = i\sigma = i\int_{x'-iy'}^{x'+iy'} q(z)dz \tag{17}$$

and

$$q(z) = \{2\mu[E - V(x)]\}^{1/2}, \tag{18}$$

where $x' \pm iy'$ are the turning points in the complex plane. We have used the notation of Fröman and Fröman.[1] These expressions are very similar to the expressions for sub-barrier penetration, the only difference being the complex turning points.

The solution in this region is given by (see Fig. 1)

$$x = (\dot{x}_0/\omega)\,\mathrm{dn}\,\omega t\,\mathrm{sc}\,\omega t, \quad m < \tfrac{1}{2}, \tag{19}$$

where we have taken the initial conditions to be velocity nonzero ($\dot{x}_0$) and position zero. We will use H for the total energy in order to avoid confusion with the complete elliptic integral of the second kind. A straightforward calculation gives the following connections between ω, the coefficients, the energy, and the parameter of the elliptic functions:

$$\begin{aligned} &\omega^2 = A/[\mu(2m-1)], \\ &\dot{x}_0^2 = -\mu\omega^4/2B, \quad H = \mu\dot{x}_0^2/2. \end{aligned} \tag{20}$$

Our first problem is to determine the K integral (or equivalently σ). For simplicity let us express Eq. (17) as

$$\sigma = \int_{x_1}^{x_2} q(z)\,dz, \tag{21}$$

where x_1 and x_2 are the conjugate complex turning points. Transforming to the t variable we have

$$\sigma = \mu\hbar^{-1}\int_{t_1}^{t_2} [\dot{x}(t)]^2\,dt. \tag{22}$$

Using Byrd and Friedman's tables,[21] the integral of Eq. (22) with the functions of Eq. (19) can be expressed in the form

$$\begin{aligned} \sigma = \frac{\mu\dot{x}_0^2}{\hbar}\Bigg\{ &\frac{4m_1 t}{3} + \frac{1}{3\omega} \\ &\times [4(2m-1)E(\omega t) + m\,\mathrm{sn}\,\omega t\,\mathrm{cn}\,\omega t\,\mathrm{dn}\,\omega t \\ &+ (2 - 4m + m_1 nc^2\omega t)\,\mathrm{sc}\,\omega t\,\mathrm{dn}\,\omega t\,]\Bigg\}_{t_1}^{t_2}. \end{aligned} \tag{23}$$

We must now determine the limits of integration (t_1, t_2). In the previous case (sub-barrier), integration between the turning points was equal to integration of the corresponding elliptic function between zero and the real half-period. All the operations were made in the real plane. The present case (super-barrier) is the same and the integration is carried out using the complex half-periods of the elliptic solutions of Eq. (19). The limits of integration in $\underline{x}$ and in $\underline{t}$ are given in Appendix A. For t we have

$$\omega t_2 = 3(K + iK')/2, \quad \omega t_1 = (K + 3iK')/2, \tag{24}$$

where K is the complete elliptic integral of the first kind and K' is the complementary integral.

The different elliptic functions which appear in Eq. (23) can be found from the tables of Byrd and Friedman.[21] However, the determination of the noncomplete elliptic integral of the second kind $E(\omega t)$ is more difficult (see Appendix B). The resulting expressions for the given limits are

$$\begin{aligned} E\left[\frac{3(K+iK')}{2};k\right] &= \left[\frac{3E}{2} + \frac{k'-1}{2} - \frac{k'(1-k')}{1+k-k'}\right] \\ &\quad - i\left[\frac{3E'}{2} + \frac{k-1}{2}\right. \\ &\quad \left. - \frac{3}{2}\left(\frac{EK'}{K} + E' - \frac{\pi}{2K}\right) + \frac{kk'}{1+k-k'}\right] \end{aligned} \tag{25}$$

and

$$\begin{aligned} E\left[\frac{(K+3iK')}{2};k\right] &= \left[\frac{E}{2} + \frac{1-k'}{2} + \frac{k'(1-k')}{1+k-k'}\right] \\ &\quad - i\left[\frac{3E'}{2} + \frac{k-1}{2}\right. \\ &\quad \left. - \frac{3}{2}\left(\frac{EK'}{K} + E' - \frac{\pi}{2K}\right) + \frac{kk'}{1+k-k'}\right], \end{aligned} \tag{26}$$

where E, E' are the complete elliptic integrals of the second kind and its complementary integral, and k, k' are the parameters of the elliptic function.

Using Eq. (20) and (23)–(26) the transmission coefficient for energies above a simple barrier is totally determined as a function of known data: incident energy and coefficients of the potential. Let us consider as an example the inverted parabolic barrier in super-barrier transmission. In this case, the elliptic solution transforms into a hyperbolic function due to the value of the parameter being unity[20]

$$x = (\dot{x}_0/\omega)\sinh\omega t. \tag{27}$$

Following a similar process to the general case we now obtain

$$T' = \{1 + \exp[-(\pi H/\hbar)(2\mu/|A|)^{1/2}]\}^{-1}. \tag{28}$$

This expression is that obtained by Holstein[8] using the same method as ours but without the use of the t variable.

Several examples are given in Fig. 2. We have plotted the value of T for different cases, the T' coefficients are expanded into a series in K_{II} and second-order terms are rejected. The figure shows the continuity of T', which has the value $\frac{1}{2}$ when

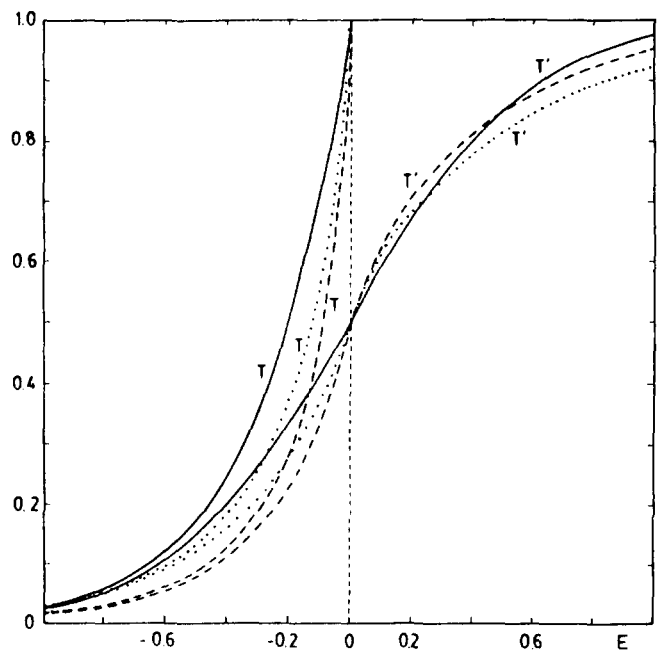


FIG. 2. Transmission coefficients for the inverted parabolic barrier $V(x) = Ax^2$ with $A = -0.8$ (full line), for the quartic barrier $V(x) = Bx^4$ with $B = -0.5$ (broken line), and for the potential $V(x) = Ax^2 + Bx^4$ with $A = -0.1$ and $B = 1.0$ (dotted line). Sub- and super-barrier transmission are plotted for each potential. The curves of T in the figure represents the first term of a series expansion of T'. The vertical broken line shows the value of the potential maximum (in our case, $V_{max} = 0$).

the incident particle energy equals the potential maximum (zero energy in the present case); this means an equal probability for reflection or transmission. Also for $E \ll 0$, $T = T'$, as was observed at the beginning of this section, but for $E \simeq 0$, T does not conserve unitarity (the reflection coefficient is equal to one for all energies in the lower order of JWKB approximation) whereas T' does (the reflection coefficient in this case is exactly $R' = 1 - T'$).

III. DOUBLE BARRIERS

In this section we consider the barriers given by the potential of Eq. (1) with $A > 0$ and $B < 0$. There are three interesting regions. The first is when $H < 0$, and in this case the tunneling can be understood as through a simple barrier, because there are only two real turning points. The second region is when the intermediate potential well is actually seen; the value of the energy in this case is $0 < H < V_{\max}$. When $H > V_{\max}$ we have the super-barrier transmission. All the cases are given in Fig. 3 in the same way as the simple barrier case was represented in Fig. 1. Restrictions on parameter m are indicated where necessary; we have also indicated the zones where ω is imaginary [when $V(x) > H$].

The relations between the different magnitudes required are for $H < 0$,

$$\omega^2 = \frac{2A}{\mu(1-2m)}, \quad x_0^2 = \frac{-m_1\mu\omega^2}{2B}, \quad H = -m\mu\omega^2 x_0^2/2; \tag{29}$$

for $0 < H < V_{\max}$ and the solution in the well,

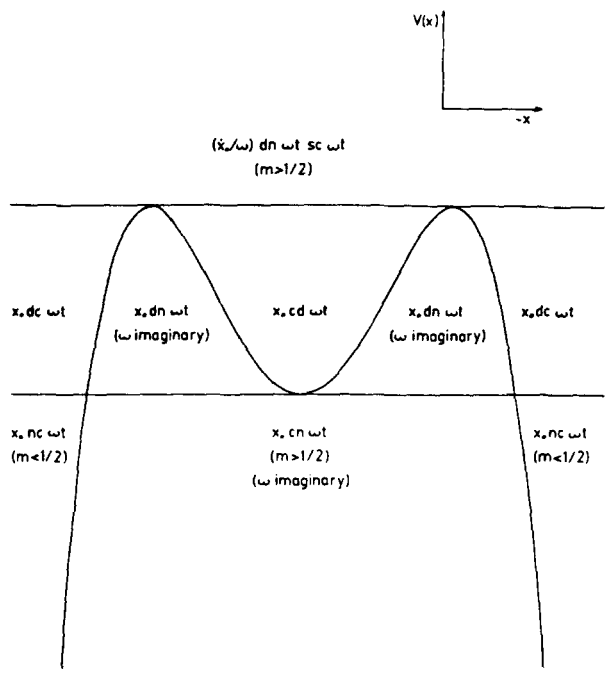


FIG. 3. The generic potential $V(x) = Ax^2 + Bx^4$, with A greater than zero and B less than zero, showing the classical solutions of the equation of motion, in terms of Jacobi elliptic functions with parameter m, in the different regions and for the following initial conditions: for $E < V_{\max}$ the initial conditions are $x(0) = x_0$, $\dot{x}(0) = 0$, for $E > V_{\max}$ they are $x(0) = 0$, $\dot{x}(0) = \dot{x}_0$. There are other solutions in all regions for the different types of initial conditions, but only the solutions used in our calculations are given.

$$\omega^2 = \frac{2A}{\mu(1+m)}, \quad x_0^2 = \frac{-m\mu\omega^2}{2B}, \quad H = \mu\omega^2 x_0^2/2; \tag{30}$$

and outside the barriers,

$$\omega^2 = \frac{2A}{\mu(1+m)}, \quad x_0^2 = \frac{-\mu\omega^2}{2B}, \quad H = \frac{m\mu\omega^2 x_0^2}{2}. \tag{31}$$

For $H > V_{\max}$ the relations are the same as for the simple barrier case, Eq. (20), because the solution is of the same form, the distinctive feature being the value of the parameter m.

A. Sub-barrier transmission

For $H < 0$ the calculations are the same as in the simple barrier case.

In the region $0 < H < V_{\max}$ the expression for the coefficient is given in many textbooks on quantum mechanics.[5,22,23] In the JWKB approximation it is

$$T = \exp[-2(K_{\rm II} + K_{\rm IV})]/4\cos^2 L_{\rm III} \tag{32}$$

with

$$K_{\rm II} = \hbar^{-1}\int_a^b p_1(x')dx', \quad K_{\rm IV} = \hbar^{-1}\int_c^d p_1(x')dx', \quad L_{\rm III} = \hbar^{-1}\int_b^c p_2(x')dx', \tag{33}$$

and

$$p_1(x) = \{2\mu[V(x) - E]\}^{1/2}, \quad p_2(x) = ip_1(x), \tag{34}$$

where a, b, c, and d are the four real turning points. In our case $K_{\rm II} = K_{\rm IV}$ because of the symmetry of the potential. Using the corresponding elliptic functions shown in Fig. 3, Eqs. (30) and (33), and following a process similar to the case of the simple barrier in sub-barrier transmission, we obtain

$$K_{\rm II} = K_{\rm IV} = (1/3\hbar)(4\mu^2 H^3/m^3|B|)^{1/4}[(1+m)E' - 2mK'] \tag{35}$$

and

$$L_{\rm III} = (2/3\hbar)(4\mu^2 m/H^3|B|)^{1/4}[(1+m)E - m_1 K], \tag{36}$$

or

$$K_{\rm II} = K_{\rm IV} = -(H/3m\hbar)[2\mu(1+m)/A]^{1/2} \times [(1+m)E' - 2mK'] \tag{37}$$

and

$$L_{\rm III} = (2H/3m\hbar)[2\mu(1+m)/A]^{1/2}[(1+m)E - m_1 K]. \tag{38}$$

In this way, we now have two equivalent expressions for the transmission coefficient as a function of known data: parameter of the elliptic functions, coefficients of the potential, and incident energy.

B. Super-barrier case

As we can see in Fig. 3, for this case the elliptic function solution is Eq. (19), but with $m > \frac{1}{2}$. The expression for the

transmission coefficient can be found in Ref. 3. Then we have

$$T = \exp[-2(K_1 + K_2)] \times [(s-1)^2 + 4s\cos^2\alpha]^{-1} \tag{39}$$

with

$$K_1 = -\left|\int_{x_1}^{x_2} q(z)dz\right|, \quad K_2 = -\left|\int_{x_3}^{x_4} q(z)dz\right|, \tag{40}$$

and

$$s = [1 + \exp(-2K_1)]^{1/2}[1 + \exp(-2K_2)]^{1/2}. \tag{41}$$

The α term can be changed to

$$\alpha = L - (\sigma_1 + \sigma_2), \tag{42}$$

where

$$L = \left|\mathrm{Re}\int_{x_2}^{x_3} q(z)dz\right| \tag{43}$$

and x_1, x_2, x_3, and x_4 are the four complex turning points; q(z) is as defined in the simple super-barrier transmission case. The quantities σ_1 and σ_2 are real and can be approximated by[3]

$$\sigma_1 = \sigma(K_1/\pi), \quad \sigma_2 = \sigma(K_2/\pi), \tag{44}$$

where

$$\sigma\left(\frac{x}{\pi}\right) = \frac{1}{2}\left\{\frac{x}{\pi}\ln\left|\frac{x}{\pi}\right| - \frac{x}{\pi} + \arg\Gamma\left[\frac{1}{2} - i\left(\frac{x}{\pi}\right)\right]\right\}, \tag{45}$$

Ford *et al.*[24] have calculated the quantum effects near to the potential maximum:

$$\frac{1}{2}\arg\Gamma\left[\frac{1}{4} - i\left(\frac{x}{\pi}\right)\right] = -\frac{x}{2\pi}\left[\ln\left(\frac{x}{\pi e}\right)^2 + \left(\frac{1}{4}\gamma\right)^2\right]^{1/4} \tag{46}$$

with $\gamma = 1.781\,07...$ If we join all the preceding results, we then have Connor's expression[25] for the transmission coefficient.

If we approximate α by L, we get Ponomarev's formula.[26] This only gives a good approximation for energy values far from the potential maxima, because the contribution of σ_1 and σ_2 in the zone near the maxima can be important.

In our case, due to the symmetry of the potential we have $K_1 = K_2$. We must also evaluate the integral for L, Eq. (43), where the limits are x_2 and x_3. We change to the more convenient t variable and find the integration limits as in Appendix A. The resultant expressions are similar to the simple super-barrier case except that now $m > \frac{1}{2}$.

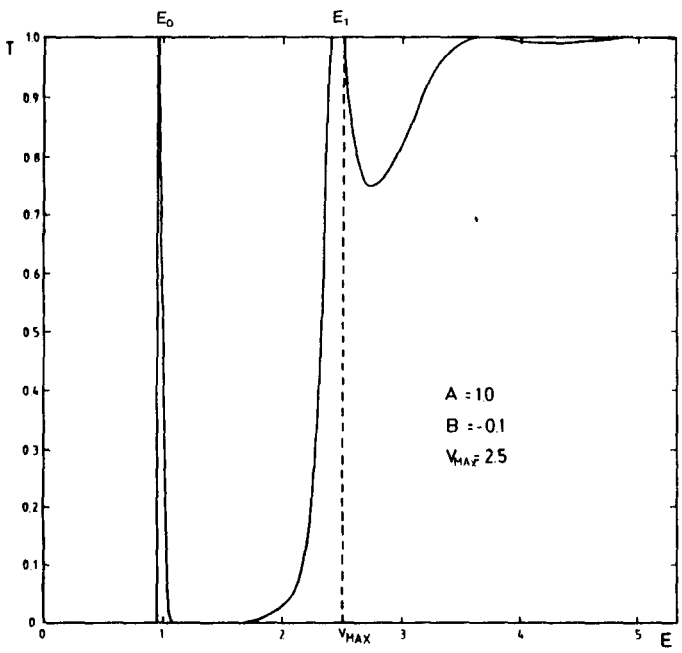


FIG. 4. Transmission coefficient versus total energy for the potential $V(x) = x^2 - 0.1x^4$. The broken vertical line shows the position of potential maximum.

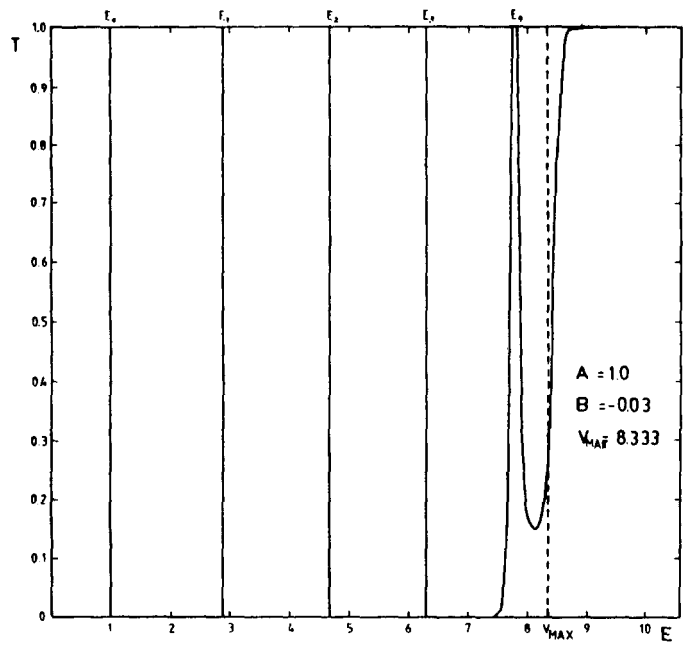


FIG. 5. The same as Fig. 4 for the potential $V(x) = x^2 - 0.03x^4$.

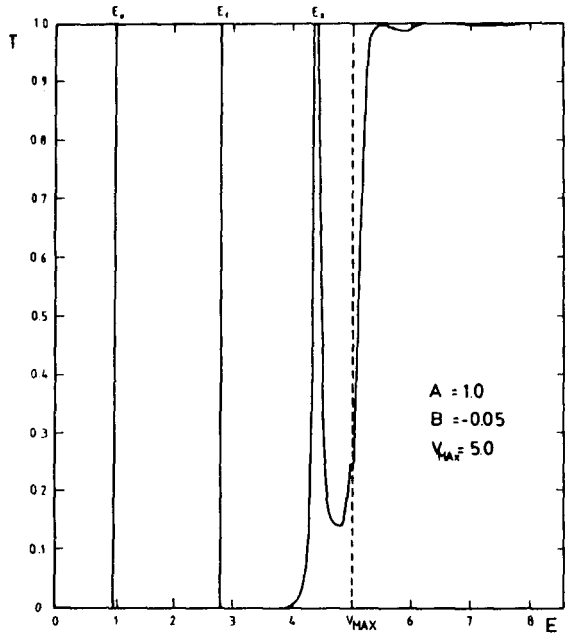


FIG. 6. The same as Fig. 4 for the potential $V(x) = x^2 - 0.05x^4$.

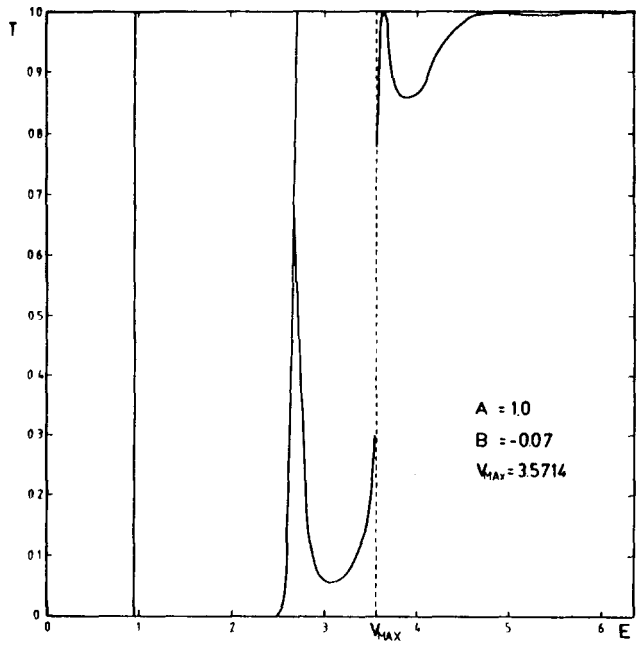


FIG. 7. The same as Fig. 4 for the potential $V(x) = x^2 - 0.07x^4$.

Some examples are shown in Figs. 4–8. In Figs. 4–7 we have plotted T versus energy E with $E > 0$. For given A and B, T is zero when $E < 0$. In Fig. 8 we can compare the Ponomarev[26] and the Connor[25] formulas. These figures show the discontinuity of T near the minimum and maxima because of the nonvalidity of the approximation used in these regions. In all the figures one observes the existence of resonances below the maxima, they appear in the places calculated by Díaz *et al.*[15] using a JWKB calculation for the energy levels in a potential well. Moreover, the value of T oscillates somewhat when $E > V_{\max}$ before approaching unity.

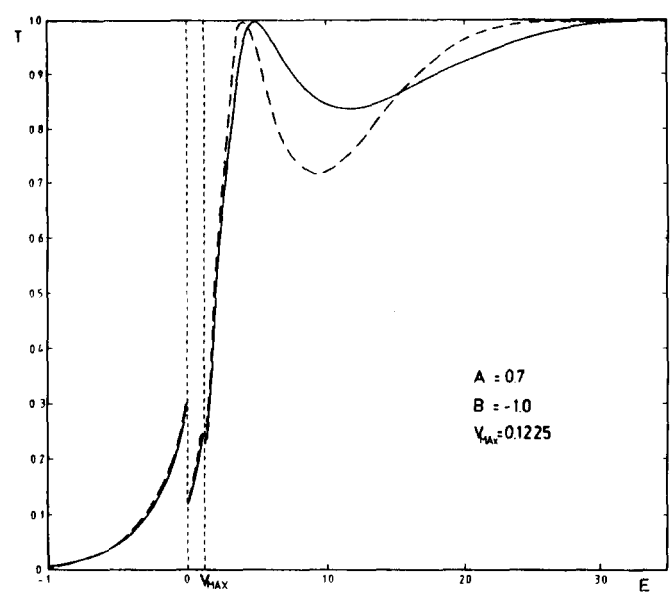


FIG. 8. T versus E for the potential $V(x) = 0.7x^2 - x^4$ calculated in the Ponomarev approximation (full line) and the Connor approximation (broken line).

ACKNOWLEDGMENT

The authors thank the Comisión Asesora de Investigación Científica y Técnica (CAICYT), Spain, for financial support.

APPENDIX A: TURNING POINTS AND INTEGRATION LIMITS IN ABOVE-BARRIER TRANSMISSION

The turning points are the solutions of $V(x) = H$. Using the results of Eq. (20), after some algebraic manipulation, we obtain

$$x^2 = (\dot{x}_0/\omega)^2(2k^2 - 1 \pm 2ikk')$$

with $k^2 = m$ and $k' = 1 - m = m_1$. This can be written as

$$x = (\dot{x}_0/\omega)(\pm k \pm ik')$$

so that the four turning points are pairwise symmetrical complex conjugates.

The calculation of the integration limits is the same for all cases. We give only one of them, $x_3 = (\dot{x}_0/\omega)(k + ik')$. From this one obtains

$$x_3 = (\dot{x}_0/\omega)(\mathrm{sn}\,\omega t_3\,\mathrm{dn}\,\omega t_3/\mathrm{cn}\,\omega t_3).$$

If we take the square of this expression and use the properties of the elliptic functions,[20] then

$$k^2\,\mathrm{sn}^4\,\omega t_3 - [1 + (k + ik')^2]\mathrm{sn}^2\,\omega t_3 + (k + ik')^2 = 0$$

with the solution

$$\mathrm{sn}^2\,\omega t_3 = (k + ik')/k$$

or[20,21]

$$\omega t_3 = 3(K + iK')/2.$$

APPENDIX B: NONCOMPLETE ELLIPTIC INTEGRALS OF THE SECOND KIND CALCULATED AT THE TURNING POINTS

We preferred not to make a series expansion for the calculation of these integrals, and instead to look for an analytic expression. The definition of Jacobi's zeta function is

$$Z(u + iv,k) = E(u + iv,k) - (E/K)(u + iv). \quad \text{(B1)}$$

One property of this function is [21]

$$\begin{aligned} Z(u + iv,k) = \{&Z(u,k) \\ &+ [k^2\,\mathrm{sn}(u,k)\mathrm{cn}(u,k)\mathrm{dn}(u,k)\mathrm{sn}^2(v,k')] \\ &\times [1 - \mathrm{sn}^2(v,k')\mathrm{dn}^2(u,k)]^{-1}\} \\ &- i\{Z(v,k') + v\pi/2KK' \\ &- [\mathrm{dn}^2(u,k)\mathrm{cn}(v,k')\mathrm{sn}(v,k')\mathrm{dn}(v,k')] \\ &\times [1 - \mathrm{sn}^2(v,k')\mathrm{dn}^2(u,k)]^{-1}\}. \end{aligned}$$

Finding the value of $E(u + iv,k)$ from (B1) and using

$$Z(u,k) = E(u,k) - (E/K)u,$$

$$Z(v,k') = E(v,k') - (E'/K')v$$

we obtain

$$\begin{aligned} &E(u + iv,k) \\ &\quad = \{E(u,k) + [k^2\,\mathrm{sn}(u,k)\mathrm{cn}(u,k)\mathrm{dn}(u,k)\mathrm{sn}^2(v,k')] \\ &\qquad \times [1 - \mathrm{sn}^2(v,k')\mathrm{dn}^2(u,k)]^{-1}\} \end{aligned}$$

$$- i\{E(v,k') - (E/K + E'/K')v + v\pi/2KK'$$
$$- [\mathrm{dn}(u,k)\mathrm{cn}(v,k')\mathrm{sn}(v,k')\mathrm{dn}(v,k')]$$
$$\times [1 - \mathrm{sn}^2(v,k')\mathrm{dn}^2(u,k)]^{-1}\}.$$

Let us look at the calculation in detail for the turning point used in Appendix A. We must use the following relations for elliptic functions[21] of argument $u + iv = 3(K + iK')/2$:

$$\mathrm{sn}(3K/2,k) = \mathrm{sn}(K/2,k) = (1 + k')^{-1/2},$$
$$\mathrm{cn}(3K/2,k) = -[k'/(1 + k')]^{1/2},$$
$$\mathrm{dn}(3K/2,k) = (k')^{1/2},$$
$$\mathrm{sn}(3K'/2,k') = (1 + k)^{-1/2},$$
$$\mathrm{cn}(3K'/2,k') = -[k/(1 + k)]^{1/2},$$
$$\mathrm{dn}(3K'/2,k') = (k)^{1/2}.$$

Then

$$E\left[\frac{3(K + iK')}{2},k\right] = \left\{E\left(\frac{3K}{2},k\right) - \left[\frac{k'(1 - k')}{(1 + k - k')}\right]\right\} - i\left[E\left(\frac{3K'}{2},k'\right) - \frac{3(EK'/K + E' - \pi/2K)}{2} + \frac{kk'}{(1 + k - k')}\right]. \quad \text{(B2)}$$

The problem now is to calculate $E(3K/2,k)$ and $E(3K'/2,k')$. We use Eq. (B1) in the form

$$Z(u + v,k) = E(u + v,k) - (E/K)(u + v)$$

and[21]

$$Z(u + v,k) = Z(u,k) + Z(v,k) - k^2\,\mathrm{sn}(u,k)\mathrm{sn}(v,k)\mathrm{sn}(u + v,k).$$

Then

$$E\left(\frac{3K}{2},k\right) = Z\left(\frac{K}{2},k\right) + Z(K,k) - k^2\,\mathrm{sn}\left(\frac{K}{2},k\right) \times \mathrm{sn}(K,k)\mathrm{sn}\left(\frac{K}{2} + K,k\right) + \frac{E}{2} + E.$$

Using Eq. (B1) again, taking into consideration the known values of the elliptic functions, and with[21]

$$E(K/2,k) = [E + (1 - k')]/2$$

we obtain

$$E(3K/2,k) = [3E - (1 - k')]/2.$$

The same procedure can be used to find

$$E(3K'/2,k') = [3E' + (k - 1)]/2.$$

These results substituted into Eq. (B2) give Eq. (25).

For the other three turning points the same method is followed.

[1] N. Fröman and P. O. Fröman, JWKB *Approximation. Contribution to the Theory* (North-Holland, Amsterdam, 1965).
[2] N. Fröman and P. O. Fröman, Nucl. Phys. A **147**, 606 (1970).
[3] N. Fröman and O. Dammert, Nucl. Phys. A **147**, 627 (1970).
[4] J. D. Cramer and J. R. Nix, Phys. Rev. C **2**, 1048 (1970).
[5] D. Rapp, *Quantum Mechanics* (Holt, Rinehart, and Winston, New York, 1971).
[6] D. W. McLaughlin, J. Math. Phys. **13**, 1099 (1972).
[7] B. R. Holstein and A. R. Swift, Am. J. Phys. **50**, 829 (1982).
[8] B. R. Holstein and A. R. Swift, Am. J. Phys. **50**, 833 (1982).
[9] B. R. Holstein, Am. J. Phys. **52**, 321 (1984).
[10] O. Dammert, J. Math. Phys. **24**, 2163 (1984).
[11] R. D. Carlitz and D. A. Nicole, Ann. Phys. (NY) **164**, 411 (1985).
[12] G. Barton, Ann. Phys. (NY) **166**, 322 (1986).
[13] A. Radosz., J. Phys. C **18**, L189 (1985).
[14] B. R. Holstein, J. Phys. C **19**, L279 (1986).
[15] A. Martín Sánchez and J. Díaz Bejarano, J. Phys. A: Math. Gen. **19**, 887 (1986).
[16] J. Díaz Bejarano, A. Martín Sánchez, and C. Miró Rodríguez, An. Fis. Ser. A **78**, 159 (1982).
[17] J. Díaz Bejarano and A. Martín Sánchez, An. Fis. Ser. A **79**, 8 (1983).
[18] T. C. Bradbury, *Theoretical Mechanics* (Wiley, New York, 1968).
[19] J. Díaz Bejarano, A. Martín Sánchez, and C. Miró Rodríguez, J. Chem. Phys. **85**, 5128 (1986).
[20] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1972).
[21] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientists* (Springer, Berlin, 1971).
[22] D. Bohm, *Quantum Theory* (Prentice–Hall, New York, 1951).
[23] A. Galindo and P. Pascual, *Mecánica Cuántica* (Alhambra, Madrid, 1978).
[24] K. W. Ford, D. L. Hill, M. Wakano, and J. A. Wheeler, Ann. Phys. (NY) 7, 239 (1959).
[25] J. N. L. Connor, Mol. Phys. **12**, 401 (1967).
[26] L. I. Ponomarev, Lectures on quasiclassics ITF-67-53. Institute for Theoretical Physics, Acad. Sc. Ukr. SSR Kiev, 1968.

Analytical continuation of the Faddeev equation for local potentials

A. Delfino[a)]
Departamento de Física, Universidade Federal de Pernambuco, 50.000 Recife-PE, Brazil

(Received 21 April 1986; accepted for publication 1 October 1986)

It is shown how to analytically continue the Faddeev equation in the second sheet of the complex energy plane when one has a local two-body interaction.

I. INTRODUCTION

In scattering theory, virtual states and resonances are associated with poles of the on-the-energy-shell S matrix on the unphysical sheet of the complex energy plane. If such poles are close to the positive real axis (physical scattering region), then scattering observables like phase shifts are very strongly influenced. In two-body problems the situation is very clear: in general one only has to obtain Jost functions and to look for zeros corresponding to poles of the on-shell S matrix.[1] Virtual states correspond to poles at real negative energies, whereas resonances correspond to pairs of complex conjugate poles on the second sheet. For short-ranged potentials the above statement is at most an easy numerical exercise. Three-body systems, however, are essentially more complicated. For such a system few proposals have been made in order to calculate resonances and virtual states. Here we discuss briefly the ideas of Girard and Fuda[2] (GF), Fonseca, Tomio, and Adhikari[3] (FTA), Pearce and Afnan[4] (PA), and Glöckles[5] (G). The GF method uses partial wave dispersion relations and numerically continues the approximate solution of the partial wave N/D equations for the neutron/deuteron problem to the relevant unphysical sheet. The second proposal (FTA) analytically continues the approximate solution of the Faddeev equation (calculated in the first sheet of energy) with known analytic properties onto the unphysical sheet associated with the lowest scattering threshold. They applied this method to study Efimov virtual states in the three-boson Amado model. The (PA) method reduces the finding of resonances to the solution of an auxiliary eigenvalue Faddeev equation for complex energies. They applied this method to πd elastic scattering with coupling to the $N\Delta$ channel. Glöckle's method, however, analytically continues Faddeev's equation (before solution) into the unphysical sheet of energy. It was first applied to study the S-matrix pole trajectory in a three-neutron model[5] and more recently to calculate the virtual state of the triton where a two-pion exchange three-nucleon force was present.[6]

To solve the Faddeev equation (on the first or the second sheet of the complex energy plane) for a system interacting via local two-body potentials, we must solve in general a coupled set of two-dimensional integral equations. To avoid the complications of two-dimensional integral equations, the above methods use separable two-nucleon interactions. This reduces the problem to a coupled set of one-dimensional integral equations. Calculations of virtual states and resonances for a three-nucleon system interacting via a local two-nucleon potential do not exist to the best of our knowledge. The theoretical investigation of the Faddeev's equation for such a problem is by itself an interesting subject of scattering theory and would have several applications in physics.

In this paper, using the G method, we show how to analytically continue Faddeev's equation for a system of three bosons interacting via local two-body potentials. It is done conveniently using some aspects of the formulation of Karlsson and Zeiger[7] [hereafter (KZ)] of Faddeev's equation. In reality we just use the idea of KZ formulation concerning the two-body t matrix written in its half-off-the-energy-shell form. It has the advantage that the three-body energy, instead of appearing in the off-energy-shell t-matrix (as in the usual form of the Faddeev equation), appears in the resolvent operator of the equation. This fact turns out to be important in our approach in order to perform the previously mentioned analytical continuation.

In order to make the paper more consistent we present in Sec. II the G method to a two-body system. In Sec. III we present how to continue the Faddeev equation onto the second sheet of energy associated with the lowest scattering threshold if we have a local potential.

II. TWO-NUCLEON SYSTEM

We start with the S-wave Schrödinger equation in momentum space ($h = 2m = 1$),

$$\psi(p) = \int_0^\infty \frac{p'^2\, dp'\, V(p,p')\psi(p')}{E - p'^2}, \tag{1}$$

where p is the relative momentum between two nucleons and m is the mass of each nucleon. The integration limits in Eq. (1) and in the rest of the paper extend from 0 to ∞. For a local interaction $V(p,p')$ is given by

$$V(p,p') = \frac{2}{\pi pp'} \int_0^\infty dr \sin(pr) \sin(p'r) V(r). \tag{2}$$

For instance, if we put the Reid $^1S_0$ potential[8] in (1), the system does not support a bound state. At this point we are on the first sheet of energy. First the G method generalizes (1) to the following form:

$$\eta(E)\psi(p) = \int \frac{p'^2\, dp'\, V(p,p')\psi(p')}{E - p'^2}. \tag{3}$$

Since the kernel of the equation is compact, provided $V(r)$ satisfies certain bounds, there exists an infinite number of discrete eigenvalues η. Equation (3) is Eq. (1) with potential V/η. Now for some $\eta < 1$ Eq. (2) has a nontrivial solution. In a classical work[9] Weinberg showed that $\eta(E)$ is a

[a)] Present address: Departamento de Física, Universidade Federal Fluminense, 24.000 Niterói-RJ, Brazil.

 0022-2488/87/030643-04$02.50

monotomic function of energy ($E<0$). If η increases, then E increases, too. The idea of the analytical continuation is shown in Fig. 1. Sections I and II refer to first (physical) and second (unphysical) sheets of the complex energy plane. Equation (3) has a square-root unitarity cut along the real positive energy axis, shown in Fig. 1. Our aim is to analytically continue Eq. (3) through the square-root unitarity cut and look for the pole in the second sheet. The pole in the first sheet corresponds to some $\eta<1$. As η increases E_b follows the indicated arrow. If we approach the upper rim of the cut ($\eta\to 1$) we encounter a pole singularity in $p'=p_0=\sqrt{E}$. Once the analyticity of $V(p,p')$ and $\psi(p')$ in the neighborhood of p_0 is established we can deform the path of integration in p' away from the real axis near $p'=\sqrt{E}$. It allows us to move with E onto the upper rim of the cut and even across the cut onto the lower half-plane of the second sheet. Then the path of integration is shifted back to the real axis. Thereby one sweeps over the pole $P_0=-i\sqrt{|E|}$ and picks up a residue term. It means that, in place of Eq. (3), we have

$$\eta^{\mathrm{II}}(E)\psi^{\mathrm{II}}(p)=\int_0^\infty \frac{p'^2\,dp'\,V(p,p')\psi^{\mathrm{II}}(p')}{p_0^2-p'^2} - i\pi p_0 V(p,p_0)\,\psi^{\mathrm{II}}(p_0)\,. \tag{4a}$$

The superscript II indicates that this equation is valid in the second sheet of energy. The analyticity of $\eta(E)$ is extensively discussed in Ref. 10. In order to obtain a closed set of integral equations we define

$$\eta^{\mathrm{II}}(E)\psi^{\mathrm{II}}(p_0)=\int_0^\infty \frac{p'^2\,dp'\,V(p_0,p')\psi^{\mathrm{II}}(p')}{p_0^2-p'^2} - i\pi p_0 V(p_0,p_0)\,\psi^{\mathrm{II}}(p_0)\,. \tag{4b}$$

By solving set (4) we calculate the virtual state (E_v) of two nucleon in singlet state interacting via the Reid soft core potential. The condition

$$\eta^{\mathrm{II}}(E_v)=1 \tag{5}$$

gives us

$$E_v=-0.1218\ \mathrm{MeV}\,.$$

Our value leads credence to the result obtained in Ref. 3.

III. THREE-BODY SYSTEM

To simplify the discussion we take a system of three identical bosons with mass m ($h=m=1$) interacting through a local potential. The Faddeev equation for bound state (s wave) in the momentum space reads[9]

$$\psi(pq)=\frac{1}{E-p^2-\frac{3}{4}q^2}\int_0^\infty q'^2\,dq' \times\int_{-1}^{1} dx\,t(p,\pi_1,E-\tfrac{3}{4}q^2)\,\psi(\pi_2,q')\,, \tag{6}$$

where

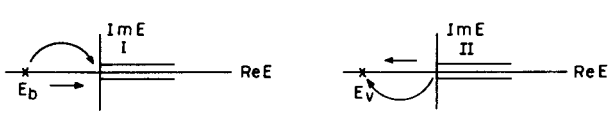


FIG. 1. The first and the second sheet of the complex energy plane. The positions of the ground state and virtual states are indicated.

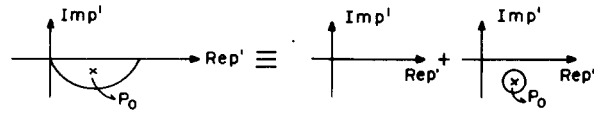


FIG. 2. Deformation of the path of integration necessary for continuation to the second sheet.

$$\pi_1=\sqrt{\tfrac{1}{4}q^2+q'^2+qq'x}\quad\text{and}\quad \pi_2=\sqrt{q^2+\tfrac{1}{4}q'^2+qq'x}\,. \tag{7}$$

In Eq. (6) p refers to the Jacobi relative momentum of two particles, q denotes the relative momentum of the center of mass of the pair and the remaining particle, t is the two-body off-the-energy-shell scattering amplitude in the three-body space, and E is the three-body energy. At this point it is interesting to see that the reduction of (6) when the interaction is separable[5] is given by

$$F(q)=\int_0^\infty \frac{q'^2\,dq'\,Z(q,q')F(q')}{E-\frac{3}{4}q'^2-E_2}\,, \tag{8}$$

where Z (the "effective potential") is related to form factors and E_2 is the two-body binding energy. The amplitude F satisfying (8) has two unitarity cuts starting at the elastic scattering and breakup thresholds. The former corresponds to a square-root cut. Equation (8) is formally equivalent to Eq. (3). In Refs. 5 and 6 the analytical continuation is done as in Sec. II.

In contrast to Eq. (8), Eq. (6) without modifications needs a more detailed study in order to be analytically continued into the second sheet of energy. Note that the cut arising from the two-body t matrix in the three-body space is hidden in the variables $(E-\frac{3}{4}q^2)$ and π_1. This fact makes it difficult to perform a direct contour deformation in the q' plane to the second sheet of energy. In order to obtain a formal equivalence between Eq. (6) and Eq. (8) we use the idea of Karlsson and Zeiger.[7] It consists of substituting the t matrix into Eq. (6) in the form

$$t(Z)=V+VG(Z)V, \tag{9}$$

with the complete propagator $G(Z)$ in its spectral representation

$$G(z)=\sum_b \frac{|\phi_b\rangle\langle\phi_b|}{Z-E_b}+\int_0^\infty \frac{K^2\,dK\,|\phi_K^+\rangle\langle\phi_K^+|}{Z-E_K}\,, \tag{10}$$

where $Z=E-\frac{3}{4}q^2$, $|\phi_b\rangle$, E_b, $|\phi_K^+\rangle$, and E_K refers to two-body bound-state wave function, binding energy, outgoing scattering state, and kinetic energy, respectively. To avoid unnecessary subscripts we restrict ourselves to the case where just one bound state exists, namely $E_b=E_2$. With such consideration the t matrix of Eq. (6) can be written as

$$t(p,\pi_1,E-\tfrac{3}{4}q^2) = V(p,\pi_1)+\frac{\langle p|V|\phi_2\rangle\langle\phi_2|V|\pi_1\rangle}{E-\frac{3}{4}q^2-E_2} + \int_0^\infty K^2\,dK\,\frac{\langle p|V|\psi_K^+\rangle\langle\psi_K^+|V|\pi_1\rangle}{E-\frac{3}{4}q^2-E_K}\,. \tag{11}$$

This equation relates the off-the-energy-shell t matrix with its half-energy-shell form. Note that

$$\langle p|V|\phi_K^+\rangle\langle\phi_K^+|V|\pi_1\rangle = t^+(p,K)t(K,\pi_1)\,. \tag{12}$$

By substituting Eq. (11) into Eq. (6) we have

$$\psi(p,q) = \frac{1}{E-p^2-\frac{3}{4}q^2}\int_0^\infty q'^2\,dq'\int_{-1}^1 dx \times\left\{V(p,\pi_1)+\int_0^\infty \frac{K^2\,dK\,t^+(p,K)t(K,\pi_1)}{E-\frac{3}{4}q^2-E_K} + \frac{\langle p|V|\phi_2\rangle\langle\phi_2|V|\pi_1\rangle}{E-\frac{3}{4}q^2-E_2}\right\}\psi(\pi_2,q')\,. \tag{13}$$

Our aim now is to put (13) into the same structure as (8) where we know how to analytically continue the equation in the second sheet by the G method. In order to get it we introduce the following definitions:

$$\begin{aligned} F(q) &= \int_0^\infty q'^2\,dq'\int_{-1}^1 dx\langle\phi_2|V|\pi_1\rangle\,\psi(\pi_2,q')\,,\\ F_0(p,q) &= \int_0^\infty q'^2\,dq'\int_{-1}^1 dx\,V(p,\pi_1)\,\psi(\pi_2,q')\,,\\ F_K(q) &= \int_0^\infty q'^2\,dq'\int_{-1}^1 dx\,t(K,\pi_1)\,\psi(\pi_2,q')\,. \end{aligned} \tag{14}$$

With them, (13) becomes

$$\psi(p,q) = \frac{1}{E-p^2-\frac{3}{4}q^2}\left\{F_0(p,q)+\frac{\langle p|V|\psi_2\rangle}{E-\frac{3}{4}q^2-E_2}F(q) + \int_0^\infty \frac{K^2\,dK\,t^+(p,K)F_K(q)}{E-\frac{3}{4}q^2-E_K}\right\}. \tag{15}$$

Now, using (15) the set of F's amplitudes satisfy the following integral equations:

$$F(q) = \left\{\int q'^2\,dq'\int_{-1}^1 dx\left[\frac{A(q,q',x)}{N(q,q',x)}F'(q') + \int K^2\,dK\,\frac{A_K(q,q',x)}{N_K(q,q',x)}F_K(q') + \frac{B_0(q,q',x)}{N_0(q,q',x)}F_0(\pi_2,q')\right]\right\}, \tag{16a}$$

$$F_K(q) = \left\{\int q'^2\,dq'\int_{-1}^1 dx\left[\frac{A_K(q,q',x)}{N(q,q',x)}F(q') + \int K'^2\frac{dK'\,A_{KK'}(q,q',x)}{N_{K'}(q,q',x)}F_{K'}(q') + \frac{B_K(q,q',x)}{N_0(q,q',x)}F_0(\pi_2,q')\right]\right\}, \tag{16b}$$

$$F_0(p,q) = \left\{\int q'^2\,dq'\int_{-1}^1 dx\left[\frac{C(q,q',x)}{N(q,q',x)}F(q') + \int\frac{K^2\,dK\,C_K(q,q',x)}{N_K(q,q',x)}F_K(q') + \frac{C_0(q,q',x)}{N_0(q,q',x)}F_0(\pi_2,q')\right]\right\}, \tag{16c}$$

where

$$\begin{aligned} A(q,q',x) &= \langle\phi_2|V|\pi_1\rangle\langle\pi_2|V|\phi_2\rangle\,,\\ A_K(q,q',x) &= \langle\phi_2|V|\pi_1\rangle\,t^+(\pi_2,K)\,,\\ B_0(q,q',x) &= \langle\phi_2|V|\pi_1\rangle\,,\\ A_{KK'}(q,q',x) &= t^+(\pi_1,K)t(K,\pi_2)\,,\\ B_K(q,q',x) &= t(K,\pi_1)\,,\\ C(q,q',x) &= V(p,\pi_1)\langle\pi_2|V|\phi_2\rangle\,,\\ C_K(q,q',x) &= V(p,\pi_1)t^+(\pi_2,K)\,,\\ C_0(q,q',x) &= V(p,\pi_1)\,,\\ N_0(q,q',x) &= E-\pi_2^2-\tfrac{3}{4}q'^2 = E-q^2-q'^2-qq'x\,,\\ N(q,q',x) &= (E-\tfrac{3}{4}q'^2-E_2)N_0(q,q',x)\,,\\ N_K(q,q',x) &= (E-\tfrac{3}{4}q'^2-E_K)\,N_0(q,q',x)\,. \end{aligned} \tag{17}$$

As we can see, the set of integral equations (16) has now through $N(q,q',x)$ the same explicit cut structure of (8). In other words, $N(q,q',x)$ generates the elastic scattering cut as shown in Fig. 3. The three-fragment channel cuts starting at $E=0$ and generated by the free propagators $N_0(q,q',x)$ are also indicated in Fig. 3. Before we start with the analytical continuation of set (16) we have to analyze the analyticity of F's and the functions defined in (17). It was made in detail in Ref. 11. Here we should mention that such analyticity depends on the form of the local potential V. For example, if V has a Yukawa form we may study the cut structure of the problem. In this case, a cut arising from $V(p,\pi)$ starts at $\pi = \pm i\mu/2$ [or equivalently at

$$q' = -2qx \pm 2i\sqrt{q^2(1-x^2)+\mu^2}$$

and

$$q' = -qx/2 \pm (i/2)\sqrt{q^2(1-x^2)+4\mu^2}$$

for $|x|<1$ and $0<q<\infty$].

Therefore to obtain the singularity lines we need a numerical study to find the domain where the F's are analytical. Similar studies are necessary for each of the numerators of (16). Hence, we assume that the branch cuts arising from the numerators of (16) are far from the region where we deform our path of integration. Now let us use the same recipe (G method) used in Sec. II and Refs. 5 and 6. If we approach the upper rim of the cut, as indicated by the arrow in Fig. 3, we encounter a pole singularity in $N(q,q',x)$. In the neighborhood of the real q' axis, we can deform the path of integration in q' *away* from the real axis near $q'=q_0=\sqrt{\frac{4}{3}(E-E_2)}$. With such prescription and generalization (as in Sec. II) $F=KF\to\eta F=KF$; we obtain the analytical continuation for Eqs. (16),

$$\begin{aligned} &\eta^{II}(E)F^{II}(q) = \{\cdots\} - \frac{4\pi}{3}iq_0\int_{-1}^1\frac{dx\,A(q,q_0,x)}{N_0(q,q_0,x)}F^{II}(q_0)\,,\\ &\eta^{II}(E)F_K^{II}(q) = \{\cdots\} - \frac{4\pi}{3}iq_0\int_{-1}^1\frac{dx\,A_K(q,q_0,x)}{N_0(q,q_0,x)}F^{II}(q_0)\,,\\ &\eta^{II}(E)F_0^{II}(p,q) = \{\cdots\} - \frac{4\pi}{3}iq_0\int_{-1}^1\frac{dx\,C(q,q_0,x)}{N_0(q,q_0,x)}F^{II}(q_0)\,, \end{aligned} \tag{18a}$$

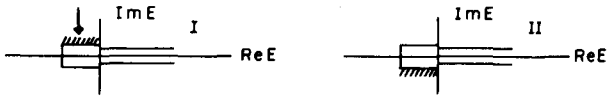


FIG. 3. The first and the second sheet of the complex energy plane. The structure of the cuts are explained in text.

where $\{\cdots\}$ are the corresponding right-hand sides of Eqs. (16). The superscript II indicates the second sheet of energy in the complex plane. In order to get a closed set, this equation has to be complemented by the following set of equations:

$$\eta^{II}(E)F^{II}(q_0) = \{\cdots\}_{q=q_0} - \frac{4\pi}{3} iq_0 \int_{-1}^{1} \frac{dx\, A(q_0,q_0,x)}{N_0(q_0,q_0,x)} F^{II}(q_0) ,$$

$$\eta^{II}(E)F_K^{II}(q_0) = \{\cdots\}_{q=q_0} - \frac{4\pi}{3} iq_0 \int_{-1}^{1} \frac{dx\, A_K(q_0,q_0,x)}{N_0(q_0,q_0,x)} F^{II}(q_0) , \tag{18b}$$

$$\eta^{II}(E)F_0^{II}(p,q_0) = \{\cdots\}_{q=q_0} - \frac{4\pi}{3} iq_0 \int_{-1}^{1} \frac{dx\, C(q_0,q_0,x)}{N_0(q_0,q_0,x)} F^{II}(q_0) .$$

The set of equations (18a) and (18b) constitutes the desired formulation of the eigenvalue problem for the virtual states of three particle interacting via a local potential. Its energy E_V is determined through $\eta(E_V) = 1$.

IV. SUMMARY

We have shown how to analytically continue the Faddeev equation for a system of three particles interacting via a local potential. From the structure of the set (18) we see that such a problem is much more complicated than the case when a separable interaction is present. Although the equations were presented for the simple case $L = S = 0$ (S being the total spin), a generalization seems to be quite trivial.

If we compare sets (18a) and (18b) with those obtained in Ref. 6 for separable potentials, calculations with local potentials in the second sheet of energy seem to be quite involved. In spite of that our study sheds light on the nature of such a calculation.

ACKNOWLEDGMENTS

The author thanks W. Glöckle very much for the suggestion to use the method of Karlsson and Zeiger to obtain Eq. (18). He also thanks the warm hospitality in the Ruhr–Universität Bochum where part of this work was performed.

This paper was partially supported by the Conselho Nacional de Pesquisas (CNPq) of Brazil.

[1] R. G. Newton, *Scattering Theory of Waves and Particles* (Springer, New York, 1982).
[2] B. A. Girard and M. G. Fuda, Phys. Rev. C **19**, 579 (1979).
[3] S. K. Adhikari, A. C. Fonseca, and L. Tomio, Phys. Rev. C **26**, 77 (1982).
[4] R. C. Pearce and I. R. Afnan, Phys. Rev. C **30**, 2022 (1984).
[5] W. Glöckle, Phys. Rev. C **18**, 564 (1978).
[6] A. Delfino and W. Glöckle, Phys. Rev. C **30**, 376 (1984).
[7] B. Karlsson and M. Zeiger, Phys. Rev. D **11**, 939 (1975).
[8] R. V. Reid, Ann. Phys. (NY) **50**, 411 (1968).
[9] S. Weinberg, Phys. Rev. **131**, 440 (1963); see also Ref. 8 where this point is discussed.
[10] W. Glöckle, *The Quantum Mechanical Few Body Problem* (Springer, New York, 1983).
[11] A. Delfino, Ph. D. thesis, Ruhr-Universität-Bochum and Universidade Federal de Pernambuco, 1984.

Geometric space-time perturbation. I. Multiparameter perturbations

M. D. Maia
Universidade de Brasilia, Departamento de Matematica, 70910, Brasilia, D. F. Brazil[a] *and Department of Physics, FM-15 University of Washington, Seattle, Washington, 98195*

(Received 26 August 1985; accepted for publication 8 October 1986)

The standard definition of space-time perturbation is reexamined. It is seen that the noninvariance of the metric under identification gauge transformations is a consequence of the adopted zero signature in the fifth dimension of the space of space-times. An n-parameter extension of that definition is proposed, with a $(4+n)$-dimensional flat space of space-times with a nonsingular metric. It is shown that in the vicinity of a point in the background space-time there is a geometrically defined family of perturbations, which are solutions of the Einstein–Yang–Mills equations.

I. INTRODUCTION

The idea of space-time perturbation has been in use since the beginning of general relativity and it remains an essential tool for astrophysics, cosmology, and quantum field theory in curved space-times.[1-9] Nonetheless, the current definitions of space-time perturbations are not clear, at least in some respects. Intuitively one speaks of a fixed space-time background $(\overline{V}_4,\bar{g})$ and a perturbed space-time $(V_4(\epsilon),g(\epsilon))$, where the metric

$$g_{ij}=\bar{g}_{ij}+\epsilon h_{ij}+\cdots,$$

is a solution of Einstein's equations. Here ϵ is a parameter (or a collection of parameters), and h_{ij} is a field over the background. In one point of view the physical space is identified with the background and a perturbation of $(\overline{V}_4,\bar{g})$ would be the fictious manifold $(V_4(\epsilon),g(\epsilon))$. When g_{ij} is replaced in Einstein-s equations, we obtain an approximate equation for h_{ij} as seen from an observer who supposedly remains unperturbed in the background.

In another more realistic point of view, the physical space is identified with the perturbed manifold $(V_4(\epsilon).g(\epsilon))$ with respect to a fictious background $(\overline{V}_4,\bar{g})$. Here a tetrad frame initially defined in the backgound changes continuously with the parameter when the perturbation is carried on the tetrad field itself. Then the perturbed metric is obtained by calculating the physical components of $\bar{g}_{ij}$ in the perturbed tetrad field. Again, replacing this metric in Einstein's equations an approximate equation for h_{ij} is obtained, with respect to the background. Since now that background is no longer the physical space, the resulting equation would not be truly physical. However, this second point of view is more general as it would include the first one as the limit when $\epsilon\to 0$, provided such a limit is properly defined.

The limit of a space-time, when certain parameters tend to given values, was studied by Geroch[10] and later applied to a geometric definition of space-time perturbation by Stewart and Walker.[11] This definition, referred to here as the Geroch–Stewart–Walker (or GSW for short) definition, is currently used as the standard geometric definition of space-time perturbation.

In the GSW definition, the physical world in its dynamical evolution is set in correspondence with a five-dimensional space $(V_5,\mathscr{G})$, in which the various stages of that evolution are pictured as distinct members of a one-parameter family of embedded space-times $(V_4(\epsilon),g(\epsilon))$, including the background $(\overline{V}_4,\bar{g})$ as a boundary. This family is characterized by a vector field in V_5 transverse (not tangent) with respect to $\overline{V}_4$. Then a (one-parameter) perturbation of $(\overline{V}_4,\bar{g})$ is defined by a one-parameter diffeomorphism of $(V_5,\mathscr{G})$ which relates $(\overline{V}_4,\bar{g})$ to any other member of the family along the integral curves of the transverse vector field. All points thus obtained are identified with a single point of the physical space by the identification map $\mathscr{J}$: (family of embedded space-times) $\to$ (physical space). In a different language, this diffeomorphism can be described as a deformation of the background.[12]

A choice of distinct transverse vector fields corresponds to a choice of distinct families, distinct perturbations, and distinct identification maps. Since this choice is made independently of the coordinates of $\overline{V}_4$, it is referred to as a choice of identification gauge and a transformation between transverse vectors is called an identification gauge transformation. As is well known, the basic probelm associated with the GSW definition is that a choice of identification gauge usually imposes a coordinate condition on $\overline{V}_4$ (see Ref. 3). Thus the space-time perturbations, which are identification gauge invariant (igi), are said to be the only physically meaningful ones. However, when this definition is applied to the metric of $\overline{V}_4$, it turns out that it can never be an igi quantity.[11,13] Therefore the intuitive idea of space-time perturbation as given by small deviations of the metric does not seem to fit well within the GSW definition.

We notice that Geroch's space of space-times has been conveniently chosen to have zero metric signature along the fifth dimension, so as to avoid measuring distances between two space-times. In fact, no dynamical principle has been proposed in the definition of the geometry of that space, so that the fifth dimension is devoid of physical significance. This situation is distinct but it reminds one of the criticism made by Einstein and Bergmann to the lack of physical significance attached to the original five-space of Kaluza.[14,15] Quite conceivably, if a space or a small portion of that space is filled with submanifolds which are identified with physical space, then it is likely to have some physical meaning.

The purpose of this note is to present a modification of

[a] Permanent address.

 0022-2488/87/030647-04$02.50

the GSW definition of space-time perturbation where the space of space-times is replaced by a $(4+n)$-dimensional flat space M_{4+n}, $n>1$. Consequently we have an n-parameter perturbation theory where the families of space-times are characterized by n orthogonal vector fields. The other major distinction from the GSW definition is that the metric of M_{4+n} is Minkowski-like with nonzero signature in all dimensions.

In Sec. II we make a brief review of the GSW definition. Section III extends that definition to n parameters and constructs the corresponding field equations. The question of gauge invariance is left to a subsequent paper.

The index notation is as follows: Greek indices refer to the higher-dimensional space and run from 1 to 5 in Sec. II and to $4+n$ in Sec. III. Lowercase Latin indices always refer to the four-dimensional space-times and run from 1 to 4. All capital Latin indices run from 5 to $4+n$.

II. ONE-PARAMETER PERTURBATIONS

Given any space-time $(\overline{V}_4,\bar{g})$, it can always be locally and isometrically embedded in a curved five-dimensional space $(V_5,\mathscr{G})$. The embedding is specified by a set of coordinates $y^\alpha(x^i)$ functions of the space-time coordinates x^i such that

$$\bar{g}_{ij}=y^\alpha_{;i}y^\beta_{;j}\mathscr{G}_{\alpha\beta}, \tag{1}$$

where the semicolon denotes covariant derivatives with respect to $\bar{g}_{ij}$. If N^α is a vector field orthogonal to $\overline{V}_4$ we also have the equations

$$y^\alpha_{;i}N^\beta\mathscr{G}_{\alpha\beta}=0,\quad N^\alpha N^\beta\mathscr{G}_{\alpha\beta}=\pm K^2, \tag{2}$$

where K is a constant.

Equations (1) and (2) are the basic equations for determining the embedding of $(\overline{V}_4,\bar{g})$, assuming that the geometry of V_5 is known. As previously mentioned, the GSW definition does not prescribe any physical principle to determine the metric $\mathscr{G}_{\alpha\beta}$. Instead, Eqs. (2) are made trivial with the assumption that $K=0$:

$$\mathscr{G}_{\alpha\beta}=\begin{pmatrix}\mathscr{G}_{mn} & 0\\ 0 & 0\end{pmatrix}. \tag{3}$$

In this case the only relevant equation is (1), which reduces to

$$\bar{g}_{ij}=l^m_i l^n_j\mathscr{G}_{mn}.$$

Here $l^m_i=y^m_{;i}$ can be thought of as a tetrad field over $\overline{V}_4$. Since the $y^m_{;i}$ define an invertible matrix of rank 4 we may also write

$$\mathscr{G}_{mn}=l^{-1i}{}_m l^{-1j}{}_n\bar{g}_{ij}. \tag{4}$$

Consequently, an observer in $\overline{V}_4$ interprets $\mathscr{G}_{mn}$ as the tetrad components of $\bar{g}_{ij}$. This means that the choice of metric (3) reduces the extrinsic geometry of $\overline{V}_4$ to its Riemannian geometry, aided by tetrad formalism.

Nonetheless the existence of the embedding space means that the normal vector field cannot be ignored. Let ζ^α denote a vector field in V_5 such that it is not tangent to $\overline{V}_4$ (a transverse vector field),

$$\zeta^\alpha=a^i y^\alpha_{;i}+aN^\alpha,\quad a\neq 0, \tag{5}$$

where the a^i are arbitrary tangent components. To each ζ^α we associate a one-parameter diffeomorphism h_s of V_5 such that for a given point $p\in\overline{V}_4$, its orbit $h_s(p)$ is the integral curve of ζ^α. Now define a nonintersecting one-parameter family of four-manifolds $V_4(x^i,s)$, with the same differentiable structure as $\overline{V}_4$, and whose points lie in the orbits $h_s(p)$ for all p belonging to the embedding neighborhood of $\overline{V}_4$. In particular, $\overline{V}_4$ is the family member corresponding to $s=0$.

If $\overline{Q}$ denotes a geometrical object in $\overline{V}_4$, the corresponding object in $V_4(x^i,s)$ is given by the appropriate action of the derivative map h^*_s of h_s. The change of $\overline{Q}$ along ζ^α is given by the Lie derivative

$$\mathscr{L}_\zeta\overline{Q}=\lim_{s\to 0}\{[h^*_s(\overline{Q})-h^*_0(\overline{Q})]/s\}.$$

Therefore if s is sufficiently small so that its powers are neglected, we obtain a linear perturbation of $\overline{Q}$ defined by

$$\overset{(1)}{Q}(x^i,s)=h^*_s(\overline{Q})=\overline{Q}+s\mathscr{L}_\zeta\overline{Q}. \tag{6}$$

Higher-order perturbations may be obtained by repeated application of (6). Thus for a k th-order perturbation we have

$$\overset{(k)}{Q}=\sum_{j=0}^{k}\frac{s^j}{j!}Q_j,\quad Q_{j+1}=\mathscr{L}_\zeta Q_j,\quad Q_0=\overline{Q}.$$

In particular, the linear perturbation of the tetrad field is

$$\overset{(1)}{l}{}^m_i=\bar{l}^m_i+s\mathscr{L}_\zeta\bar{l}^m_i. \tag{7}$$

When contracted with a geometrical object $\overline{Q}$, this perturbed tetrad produces a perturbation on the tetrad (or physical) components of $\overline{Q}$, with perturbation order depending on the rank of $\overline{Q}$.

Now we are in position to restate the GSW definition of space-time perturbation. A member of the family $\overline{V}_4(x^i,s)$ constructed above is a perturbation of $\overline{V}_4$ (the background) when its metric is induced by the metric of V_5 via the perturbed tetrad

$$\overset{(2)}{g}_{ij}=\overset{(1)}{l}{}^m_i\overset{(1)}{l}{}^n_j\mathscr{G}_{mn}=t^m_i t^n_j\bar{g}_{mn}, \tag{8}$$

where we have denoted

$$t^m_i=\overset{(1)}{l}{}^m_q l^{-1q}_i.$$

Therefore after reaching expression (8) the problem can be handled without further mention of the five-space. However, considerations on V_5 are important to understand the difference between two perturbations of the same order. From (5) and (6) and the properties of Lie derivatives, the difference between two linear perturbations of $\overline{Q}$ produced by two transverse vector fields $\zeta^\alpha,\zeta'^\alpha$ is

$$\overset{(1)}{Q}{}'-\overset{(1)}{Q}=\mathscr{L}_\xi\overline{Q}+\mathscr{L}_\eta\overline{Q}, \tag{9}$$

where

$$\xi^\alpha=(s'a'^i-sa^i)y^\alpha_{;i}$$

is an arbitrary tangent vector and

$$\eta^\alpha=(s'a'-sa)N^\alpha$$

is a vector normal to $\overline{V}_4$. With the adoption of the metric (3) for V_5, we do not have a measure for a so that we can always

choose a, a' such that η^α is a zero while ξ^α remains arbitrary. Therefore the two perturbations of $\overline{Q}$ will be equal if $\mathcal{L}_\xi \overline{Q} = 0$ for any tangent vector ξ^α. This condition cannot be satisfied except for a very special class of objects. In particular if $\overline{Q} = \bar{g}_{ij}$ this would require the impossible condition that any tangent vector field should be a Killing vector field of $\overline{V}_4$ (see Ref. 11).

One possible way out of the above difficulty is to remove the arbitrariness of ξ^α by assigning a measure along the fifth dimension in such a way that η^α would still vanish but ξ^α is a Killing vector field of $\overline{V}_4$. Such a scheme would be better understood in a multiparameter perturbation program.

III. MULTIPARAMETER PERTURBATIONS

The embedding of a curved manifold into another curved manifold is a difficult problem of differential geometry. A simpler problem is to embed a Riemannian manifold into a flat space. In particular, for space-times there are numerous known examples with varying dimensions and signatures.[16] We may generically assume that $\overline{V}_4$ is locally and isometrically embedded in a flat space M_{4n} with $4+n$ dimensions and signature $p(+)+q(-)$. These embeddings are specified by a set of Cartesian coordinates $x^\mu(x^i)$ such that

$$\bar{g}_{ij} = X^\mu_{,i} X^\nu_{,j} \eta_{\mu\nu}, \tag{10}$$

where $\eta_{\mu\nu}$ denotes the Cartesian components of the metric of M_{4+n} and the comma denotes partial derivatives.

When $X^\mu(x^i)$ are real analytic functions of x^i we may use Friedman's adaptation of the Janet–Cartan theorem showing that ten dimensions are sufficient to embed analytically any four-dimensional space-time.[17] While most known embeddings fall in this category, there is no proof that all space-times can be analytically embedded and they probably cannot. In fact, if we consider the most general cases, including regions that are near singularities, then it is likely that the analytic condition fails and the best we can hope is that these functions remain differentiable. In this case it has been shown that the maximum number of required dimensions rises to 14 (see Ref. 18). This limit is irrelevant to our present considerations except for the fact that we expect to be dealing with 14 independent differential equations.

If N^μ_A denotes n vector fields orthogonal to $\overline{V}_4$, then besides (10) we also have the following equations:

$$X^\mu_{,i} N^\nu_A \eta_{\mu\nu} = 0, \quad N^\mu_A N^\nu_B \eta_{\mu\nu} = \bar{g}_{AB} = \kappa^2 \epsilon_A \delta_{AB}, \tag{11}$$

where now κ is a nonzero constant, which for simplicity we take to be 1 and $\epsilon_A = \pm 1$ are the signature numbers. For the purpose of perturbation theory we assume that if the background $(\overline{V}_4, \bar{g})$ has a certain embedding signature (p,q) then its perturbations also have the same signature.

A generic transverse vector field in M_{4+n} has the general expression

$$\zeta^\mu = \xi^\mu + X^A N^\mu_A,$$

where again ξ^μ denotes an arbitrary tangent vector. To compare with Sec. II we may introduce the notation $s = \sqrt{g_{AB} X^A X^B}$ and a single vector $N^\mu = x^A N^\mu_A / s$, bearing in mind that the independent parameters are x^A. With this notation the transverse vector reads

$$\zeta^\mu = \xi^\mu + s N^\mu.$$

As before, to each such vector we associate a diffeomorphism h_s of M_{4+n} such that for each $p \in \overline{V}_4$, its orbit is the integral curve of ζ^μ, with parameter s. We may now introduce an n-parameter family of embedded manifolds $V_4(x^i, s)$ whose points lie in the orbits of h_s. Therefore in the neighborhood of $\overline{V}_4$ the points of the family can be expressed by the coordinates

$$Z^\mu(x^i, s) = X^\mu(x^i) + s\left(\frac{\partial Z^\mu}{\partial s}\right) + \cdots .$$

Assuming that s is sufficiently small, this expression can be approximated by a straight line (actually, since our embedding space is flat, these lines are globally defined)

$$Z^\mu(x^i, s) = X^\mu(x^i) + s N^\mu. \tag{12}$$

The identification map can be constructed as in the previous section. Supposing that the manifolds described by (12) are space-times, the various points $p = h_s(p)$, $p' = h_{s'}(p)$,..., associated with different values of s in (12) correspond to a single point in the physical space. For each set of independent vectors N_A satisfying (11) we have one such identification map. In other words, different identification maps are generated by pseudorotations of the vectors N_A. For a fixed origin in $\overline{V}_4$ these transformations belong to the group $SO(r,s)$ where $r(+)+s(-)$ denotes the signature of $\bar{g}_{AB}$. This group is a noninvariant subgroup of the homogeneous group of isometries of M_{4+n}, $SO(p,q)$. This means that a transformation of $SO(r,s)$ induces a transformation in the subspace tangent to $\overline{V}_4$. Consequently, it is sufficient to calculate perturbations generated by the normal vectors N_A and use the transformations of $SO(r,s)$ to change the identification gauge. Obviously, such a situation cannot exist in the five-dimensional case of the last section. It is interesting to notice that such construction can be generalized to other structures where the high-dimensional space is not necessarily an embedding space.[19]

The linear perturbation of a geometric object $\overline{Q}$ defined in $\overline{V}_4$, corresponding to the normal direction N, or equivalently to a choice of n normal vectors N_A, is

$$\overset{(1)}{Q} = \overline{Q} + s\mathcal{L}_N \overline{Q} = \overline{Q} + x^A \mathcal{L}_{N_A} \overline{Q}, \tag{13}$$

and in particular for the metric $\bar{g}_{ij}$ we have the linear n-parameter perturbation

$$\overset{(1)}{g}_{ij} = \bar{g}_{ij} + x^A \mathcal{L}_{N_A} \bar{g}_{ij}.$$

We shall see that $\mathcal{L}_{N_A} \bar{g}_{ij}$ is given by the second quadratic form of $\overline{V}_4$.

Unlike (1), expression (10) cannot be reduced to a simple tetrad construction, but we can make use of a vielbein formulation. Defining the vielbein $\bar{l}^\mu_i = X^\mu_{,i}$, relating the Cartesian frame to a tangent frame in $\overline{V}_4$, its linear n-parameter perturbation is given by

$$\overset{(1)}{l}{}^\mu_i = \bar{l}^\mu_i + x^A \mathcal{L}_{N_A} \bar{l}^\mu_i = \bar{l}^\mu_i + x^A N^\mu_{A,i}, \tag{14}$$

where $N^\mu_{A,i}$ is given by Ref. 20,

$$N^{\mu}_{A,i} = -\bar{g}^{mn} b_{imA} X^{\mu}_{,n} + \bar{g}^{MN} A_{iMA} N^{\mu}_{N}, \tag{15}$$

and where

$$b_{imA} = -N^{\mu}_{A,i} X^{\mu}_{,m} \eta_{\mu\nu}$$

are the coefficients of the second quadratic form of $\bar{V}_4$ and

$$A_{iMA} = N^{\mu}_{A,i} N^{\nu}_{M} \eta_{\mu\nu}$$

are the components of the "torsion" vector. It has been observed that these components transform as the Lie algebra components of a gauge potential under SO(r,s) (see Ref. 21). Therefore, if the L^{AB} denote the Lie algebra generators of that group, these geometric gauge potentials are $A_i = A_{iAB} L^{AB}$. Using (15) the linear vielbein perturbation becomes

$$\overset{(1)}{l}{}^{\mu}_{i} = (\delta^{n}_{i} - X^{A} \bar{g}^{mn} b_{imA}) l^{\mu}_{n} + x^{A} \bar{g}^{MN} A_{iMA} N^{\mu}_{N}, \tag{16}$$

which is the same as $\partial Z^{\mu}/\partial x^{i}$, where Z^{μ} is given by (12).

Following the same idea as in Sec. II, we may define the n-parameter geometric perturbation of the background $\bar{V}_4$ as a member of the family of four-manifolds given by (12) whose metric is induced by $\eta_{\mu\nu}$ via the perturbed vielbein (16)

$$\overset{(2)}{g}{}_{ij} = \overset{(1)}{l}{}^{\mu}_{i} \overset{(1)}{l}{}^{\nu}_{j} \eta_{\mu\nu} = g_{ij} + X^{A} X^{B} \bar{g}^{MN} A_{iMA} A_{jNB}, \tag{17}$$

where we have denoted

$$g_{ij} = \bar{g}_{ij} - 2X^{A} b_{ijA} + x^{A} x^{B} \bar{g}^{mn} b_{imA} b_{jnB}. \tag{18}$$

Notice that by using (12), expression (17) is equivalent to

$$\overset{(2)}{g}{}_{ij} = Z^{\mu}_{,i} Z^{\mu}_{j} \eta_{\mu\nu}. \tag{19}$$

In the case of the GSW definition, the perturbed metric is simply replaced in Einstein's equations and these are solved in terms of t^{m}_{i}. Here we have a different situation because of the larger number of functions to be determined and therefore we also need an additional set of equations. These equations are derived from the integrability conditions for (10) and (11), the Gauss–Codazzi–Ricci equations.[20]

The next step would be to write Einstein's equations for g_{ij} and use the mentioned supplementary equations to determine the complete set of unknowns. A simpler but less straightforward method is to translate the metric of M_{4+n} to the Gaussian coordinate system formed by x^{i} and x^{A} and the equivalence of expressions (17) and (19) to obtain the following metric expression[21]:

$$\gamma_{AB} = Z^{\mu}_{,\alpha} Z^{\nu}_{,\beta} \eta_{\mu\nu} = \begin{pmatrix} g_{ij} + x^{A} x^{B} \bar{g}^{MN} A_{iMA} A_{jNB} & X^{A} A_{iMA} \\ X^{A} A_{iMA} & \bar{g}_{AB} \end{pmatrix}. \tag{20}$$

Then as follows from the analogy with the Kaluza–Klein metric ansatz, calculating $R(\gamma)\sqrt{\det \gamma}$,

$$R(g)\sqrt{-\det g} = -\tfrac{1}{4} \operatorname{tr} F^{2} \sqrt{-\det g},$$

where $R(\gamma)$ and $R(g)$ are the curvature scalars constructed with $\gamma_{\alpha\beta}$ and g_{ij}, respectively, and we have denoted

$$F_{ij} = \partial_i A_j - \partial_j A_i + \tfrac{1}{2}[A_i, A_j], \qquad F^{2} = g^{im} g^{jn} F_{ij} F_{mn}. \tag{21}$$

Therefore from (21) we may construct an equality of action functional whose variations with respect to g_{ij} and A_i give the equations

$$G_{ij}(g) = T_{ij}(F), \qquad D^{i} F_{ij} = 0, \tag{22}$$

where $D_i = \nabla_i + \frac{1}{2} A_i$ and ∇_i is the covariant derivative with respect to g_{ij}. Here $T_{ij}(F)$ denotes the Yang–Mills energy momentum tensor corresponding to the torsion vector A_i. All contractions are made with respect to g_{ij} (see Ref. 21).

Since $\bar{g}_{ij}$ is given as the background metric, the 14 Einstein–Yang–Mills equations (22) can be interpreted as equations on the second quadratic form and the torsion vector. While the latter has an interpretation as a Yang–Mills potential with gauge group SO (r,s), the former has not yet a clear physical interpretation.

ACKNOWLEDGMENTS

The author wishes to thank the warm hospitality received at the University of Washington where this work was completed.

This paper was partially supported by the Conselho Nacional de Pesquisas (CNPq), Brazil.

[1] S. Teulkolsky, Phys. Rev. Lett. **29**, 1114 (1972).
[2] S. Hawking, Astrophys. J. **145**, 544 (1966).
[3] J. M. Bardeen, Phys. Rev. D. **22**, 1882 (1980).
[4] S. Chandrasekar, *The Mathematical Theory of Black Holes* (Oxford U. P., Oxford, 1980).
[5] D. Brill and J. B. Hartle, Phys. Rev. B **135**, 271 (1964).
[6] R. A. Isaacson, Phys. Rev. **166**, 1263, 1272 (1968).
[7] R. Breuer, *Gravitational Perturbation Theory and Synchroton Radiation* (Springer, New York, 1975).
[8] Y. Choquet-Bruhat, *Coupling of High Frequency Gravitational and Electromagnetic Waves*, in *Proceedings of the 1st Marcel Grossman Meeting*, edited by R. Rufini (North-Holland, Amsterdam, 1977).
[9] M. D. Maia, J. Math. Phys. **22**, 538 (1981).
[10] R. Geroch, Commun. Math. Phys. **13**, 180 (1969).
[11] J. M. Stewart and M. Walker, Proc. R. Soc. London Ser. A **341**, 49 (1974).
[12] R. H. Gowdy, J. Math. Phys. **19**, 2294 (1978).
[13] R. K. Sachs, in *Relativity Groups and Topology*, edited by B. De Witt and C. De Witt (Gordon and Breach, New York, 1964).
[14] T. H. Kaluza, Sitzungsber, Preuss. Akad. Wiss. Phys. Math. Kl. **1921**, 966. [English translation in *Unified Field Theories in More than 4 Dimensions*, edited by V. de Sabata and E. Schmutzer (World Scientific, Singapore, 1982)].
[15] A. Einstein and P. Bergmann, Ann. Math. **39**, 683 (1938).
[16] J. Rosen, Rev. Mod. Phys. **37**, 201 (1965).
[17] A. Friedman, Rev. Mod. Phys. **37**, 201 (1965).
[18] R. E. Greene, Memoirs, Am. Math. Soc. **97**, (1970); see also, H. Goenner, in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980).
[19] F. Mansouri and L. Witten, Found Phys. **14**, 1095 (1984).
[20] L. P. Eisenhart, *Riemannian Geometry* (Princeton U. P., Princeton, NJ, 1966).
[21] M. D. Maia, Phys. Rev. D **31**, 262, 268 (1985).

Geometric space-time perturbation. II. Gauge invariance

M. D. Maia

Departmento de Matematica, Universidade de Brasilia, 70910 Brasilia, DF, Brazil[a] *and Department of Physics, FM-15, University of Washington, Seattle, Washington 98195*

(Received 26 August 1986; accepted for publication 8 October 1986)

Using a multiparameter definition of space-time perturbation in a $(4+n)$-dimensional flat space, the question of identification gauge invariance of the background metric is examined. It is shown that when the allowed identification gauge transformations are given by rotations in the parameter space, then the background metric is invariant. A possible association with Kaluza–Klein theory is also examined.

I. INTRODUCTION

In a previous paper we discussed the properties of space-time perturbations as defined by Geroch, Stewart, and Walker (GSW) and proposed a reformulation of that definition[1] (hereafter referred to as I). The new definition has n parameters with a nonsingular metric in the parameter space. Each perturbation is a solution of Einstein–Yang–Mills equations where the gauge potential is the torsion vector of the background.

The question of gauge invariance is discussed in the present paper. It is found that for the physically meaningful identification gauge changes, the background metric can be made gauge invariant. Because we have to work with a vielbein instead of a tetrad perturbation, the extra dimensions cannot be dispensed with as it is done in the case of the GSW definition. Consequently the space of space-times may have a physical meaning. In this respect the analogy with Kaluza–Klein theory mentioned in I is improved, with the assumption that the internal space of that theory is replaced by the space of perturbation parameters. Indeed, it is shown that this parameter space is naturally bounded and that the identification map provides the necessary identification of points located at the boundaries.

The notation and index convention is the same as in I and an equation (xx) of that paper will be referred to as (xx–I).

II. GAUGE INVARIANCE

As it was seen in I, the GSW definition of space-time perturbation produces a gauge dependence on the metric perturbations, essentially because the geometry of the five-dimensional space V_5 is chosen to have zero signature along the fifth dimension. The situation would be different if that metric had a nonzero signature. In fact let us suppose that $K=1$ in expression (2–I). Then a transverse vector field ζ^α in V_5 is given by ($N^\alpha=0$ for $\alpha\neq 5$)

$$\zeta^\alpha = a^i y^\alpha_{;i} + aN^\alpha,$$

where now the value of a can be measured. Therefore when calculating the difference between two perturbations as in (9–I)

$$\overset{(1)}{Q}{}' - \overset{(1)}{Q} = \mathcal{L}_\xi \overline{Q} + \mathcal{L}_\eta \overline{Q},$$

the normal vector $\eta^\alpha = (s'a' - sa)N^\alpha$ cannot be made zero by arbitrary choices of a. It will vanish when $a's' = as$, which corresponds to saying that the gauge transformation is an isometry in the parameter space, with the important consequence that in this case the resulting tangent vector ξ in (9–I) is no longer arbitrary, but depends on that isometry. That is, the condition for the two perturbations to be equal becomes an equation in ξ, $\mathcal{L}_\xi \overline{Q} = 0$. Therefore the resulting coordinate condition is improved (but not eliminated) with respect to the case where $K=0$. In order to completely eliminate the coordinate gauges, the equation $\mathcal{L}_\xi \overline{Q} = 0$ should reduce to an identity, but of course, this cannot be done for a generic $\overline{Q}$. It is, however, possible to select specific geometrical objects $\overline{Q}$ such that ξ generates a symmetry of $\overline{Q}$. In particular, because the current definitions of space-time perturbations rely on the perturbations of the metric, we would be interested in a gauge invariant metric perturbation and by the above argument this would be the case when ξ is a Killing vector field of $\overline{V}_4$.

The embedding of $\overline{V}_4$ in M_{4+n} is defined up to an isometry of M_{4+n}. In other words, besides the manifold mapping group of $\overline{V}_4$ we have an embedding symmetry group $SO(p,q)$. In terms of Gaussian coordinates an infinitesimal transformation of this group is $x'^\alpha = x^\alpha + \xi^\alpha$, where $\xi^{(\alpha;\beta)} = 0$, the covariant derivative being calculated with respect to $\gamma_{\alpha\beta}$ given by (20–I). These equations split as

$$x'^i = x^i + \xi^i, \quad \xi^{(i;j)} = 0, \quad \xi^{(i;A)} = 0,$$

$$x'^A = x^A + \xi^A, \quad \xi^{(A;B)} = 0.$$

Since we are interested in evaluating these transformations in the background, these Killing's equations must be projected in $\overline{V}_4$:

$$\xi^{(i;j)}|_{x^A=0} = 0, \quad \xi^{(i;A)}|_{x^A=0} = 0, \quad \xi^{(A;B)}|_{x^A=0} = 0.$$

After using the metric $\gamma_{\alpha\beta}$ and the corresponding Christoffel symbols we obtain the projected equations[2]

$$\bar{\xi}^{(i;j)} = \bar{g}^{k(i}\bar{g}^{j)m} b_{kmM}\bar{\xi}^M, \tag{1}$$

$$\xi_{(i,A)}|_{x^A=0} = b_{imA}\bar{\xi}^m - A_{iBA}\bar{\xi}^B, \tag{2}$$

$$\xi^{(A,B)}|_{x^A=0} = 0, \tag{3}$$

where we have denoted $\bar{\xi}^\alpha = \xi^\alpha|_{x^A=0}$ and the covariant derivative in (1) refers to $\bar{g}_{ij}$. Notice that in the last two equa-

[a] Permanent address.

 0022-2488/87/030651-03$02.50

tions we have partial derivatives. If $\bar{E}$ denotes the group of coordinate transformations in $\bar{V}_4$, $x'^i = x^i + \xi^i$, where $\bar{\xi}^i$ satisfy (1) and $\bar{G}$ denotes the group of transformations $x'^A = x^A + \xi^A$, where ξ^A satisfy (3), then Eq. (2) says that these groups are not invariant subgroups of the embedding symmetry SO (p,q), even when it is restricted to $\bar{V}_4$. Notice from (3) that for an observer in $\bar{V}_4$ the subgroup $\bar{G}$ describes local isometries of the parameter space with metric $\bar{g}_{AB}$. The general solution of (3) is

$$\xi^A = \theta^A_B(x^i)x^B + \theta^A_m(x^i)x^m, \tag{4}$$

where the θ^α_β depend only on x^i and $\theta^{(\alpha\beta)} = \gamma^{(\alpha\gamma}\theta^{\beta)}_\gamma = 0$. The projected $\bar{\xi}^A$ is

$$\bar{\xi}^A = \xi^A|_{x^A=0} = \theta^A_m(x^i)x^m.$$

Therefore points in the background are not necessarily mapped into the background after a transformation of $\bar{G}$. We may define a physical gauge transformation as the one that preserves the space-time definition ($x^A = 0 \Rightarrow x'^A = 0$), i.e., such that $\xi^A|_{x^A=0} = 0$, or

$$\xi^A = \theta^A_B(x^i)x^B. \tag{5}$$

In this case Eqs. (1)–(3) become

$$\bar{\xi}^{(i;j)} = 0, \quad \bar{\xi}_{i,A} = b_{imA}\bar{\xi}^m, \quad \xi^{(A,B)} = 0. \tag{6}$$

Therefore with the condition $\bar{\xi}^A = 0$, $\bar{E}$ and $\bar{G}$ become the groups of isometries of $\bar{g}_{ij}$ and $\bar{g}_{AB}$, respectively, but still not invariant subgroups of SO (p,q). In other words if we take an isometry of $\bar{g}_{AB}$ then we have an induced isometry in $\bar{V}_4$.

We may now return to the problem of identification gauge invariance for the multiparameter perturbation defined in I. As stated in that paper, our identification map is defined by a combination of the vectors N_A for a given set of parameters x^A

$$\zeta^\alpha = x^A N^\alpha_A,$$

leading to the perturbation (13–I). Applying to the background metric $\bar{g}_{ij}$ we obtain

$$\overset{(1)}{g}_{ij} = \bar{g}_{ij} + x^A \mathcal{L}_{N_A}\bar{g}_{ij}.$$

If we now change the vector ζ^α by means of an isometric transformation of the parameter space, a new perturbation is generated,

$$\overset{(1)}{g'_{ij}} = \bar{g}_{ij} + x'^A \mathcal{L}_{N'_A}\bar{g}_{ij},$$

where N'_A denotes the corresponding change in the set of vectors N_A. The difference between the two perturbations is

$$\overset{(1)}{g'_{ij}} - \overset{(1)}{g_{ij}} = \mathcal{L}_{(x'^A N'_A - x^A N_A)}\bar{g}_{ij}.$$

Now if the group of isometries of $\bar{g}_{AB}$ were an invariant subgroup of SO (p,q), then $x^A N_A$ would be an invariant and the two perturbations would be equal. However, as follows from (1)–(3) this is not the case and $x'^A N'_A - x^A N_A$ will have tangent and normal components. In fact, consider an infinitesimal transformation $x^A \to x^A + \xi^A$, where ξ^A is given by (4). From (1) and (2) it follows that there is an induced transformation $x^i \to x^i + \xi^i$ so that the vector field N_A transforms as

$$N_A \to N'_A = \frac{\partial x^B}{\partial x'^A}N_B + \frac{\partial x^m}{\partial x'^A}l_m,$$

where l_m is tangent to $\bar{V}_4$. Using (4) and writing $\xi^i(x^i,x^A) = \theta^i_m x^m + \theta^i_A x^A$, it follows that

$$N'_A = N_A - \theta^B_A N_B - \theta^m_A l_m,$$

so that

$$x'^A N'_A - x^A N_A = x^A\theta^m_A l_m + x^m l_m - \theta^B_m x^m N_B,$$

and

$$\overset{(1)}{g'_{ij}} - \overset{(1)}{g_{ij}} = \mathcal{L}_\chi\bar{g}_{ij} + \mathcal{L}_\xi\bar{g}_{ij} + \mathcal{L}_\eta g_{ij},$$

where χ and ξ are tangent vectors with components along l_m given by $\chi^m = \theta^m_A x^A$, $\xi^m = x^m$, and η is a normal vector with components along N_A given by $\eta^A = \theta^B_m x^m$. Therefore if the two perturbations differ by a coordinate transformation in M_{4+n} defined by (1)–(3), then they are distinct. However, the transformations (1)–(3) do not preserve the space-time definition (i.e., points in space-time do not necessarily remain in space-time) unless $\bar{\xi}^A = 0$, which means $\theta^A_m = 0$ and consequently if we restrict our gauge group to satisfy $\bar{\xi}^A = 0$, we have in view of (6) that ξ is a Killing vector field and $\overset{(1)}{g'_{ij}} - \overset{(1)}{g_{ij}} = \xi_{(i;j)} = 0$. In other words, gauge transformations such that $\bar{\xi}^A = 0$ produce gauge invariant metric perturbations. Notice that if $\bar{V}_4$ does not admit a Killing vector field, the only transformation induced in $\bar{V}_4$ from (6) is the identity. It is also important to note that the restriction of the gauge transformation to $\bar{\xi}^A = 0$, is the only meaningful type of gauge transformation allowed by an observer sitting in physical space.

III. GAUGE-FREE PERTURBATIONS

From the expression (17–I) we see that the perturbation of the background metric induced by the vielbein perturbation is given by the second quadratic form b_{ijA} and by the torsion vector A_{iAB}. On the other hand, comparing (20–I) with the Kaluza–Klein metric ansatz we notice that the background metric in that ansatz is replaced by

$$g_{ij} = \bar{g}_{ij} - 2x^A b_{ijA} + x^A x^B \bar{g}^{mn} b_{imA} b_{jnB}, \tag{7}$$

which does not depend on A_{iAB}. Therefore if we wish to pursue an analogy with Kaluza–Klein theory we should look at (7) as a "gauge-free" metric perturbation. Indeed the geometrical implication of A_{iAB} is to bend the family of embedded space-times with respect to N_A. This follows from (16–I), which shows that contrary to $\bar{l}^\mu_m$, the perturbed vielbein $\overset{(1)}{l}{}^\mu_M$ is not orthogonal to N_A. In other words the derivative map h_s^* does not preserve the orthogonality to N_A.

Supposing that the "gauge potentials" A_{iAB} are momentarily switched off, the perturbed vielbein becomes [from (14–I) and (16–I)]

$$\lim_{A\to 0}\overset{(1)}{l}{}^\mu_i = Y^\mu_{,i} = \bar{g}^{mn}(\bar{g}_{im} - x^A b_{imA})X^\mu_{,n}, \tag{8}$$

which is a set of vectors orthogonal to N_A for any x^A. The corresponding family of space-time is of course different

from that described by (12–I). The points of the new family can be obtained by integrating (8):

$$Y^{\mu}(x^{i},x^{A}) = X^{\mu} + x^{A}V^{\mu}_{A}, \tag{9}$$

where

$$V^{\mu}_{A} = -\oint b_{imA}\bar{g}^{mn}X^{\mu}_{,n}\,dx^{i}.$$

The metric tensor in each member of this family, induced by $\eta_{\mu\nu}$ via (8), is

$$g_{ij} = Y^{\mu}_{,i}Y^{\nu}_{,j}\eta_{\mu\nu} = \bar{g}^{mn}(\bar{g}_{im} - x^{A}b_{imA})(\bar{g}_{jn} - x^{B}b_{jnB}), \tag{10}$$

which is precisely (7). Therefore the family of manifolds described by (9) with metric (10) defines a new class of space-time perturbation induced solely by the second quadratic form without participation of the torsion vector. The identification map can be defined with the same parameter s as in I and with the vector field $V^{\mu} = x^{A}V^{\mu}_{A}/s$ so that

$$h^{\mu}_{s}(p) = Y^{\mu}(x^{i},s) = X^{\mu} + sV^{\mu}. \tag{11}$$

The identification map itself identifies all points along the curve (11) with a single point in physical space.

Since $Y^{\mu}_{,i}N^{\nu}_{A}\eta_{\mu\nu} = 0$, the metric of M_{4+n} in terms of the vielbein $Y^{\mu}_{,\alpha}$ is

$$\gamma'_{\alpha\beta} = \begin{pmatrix} g_{ij} & 0 \\ 0 & \bar{g}_{AB} \end{pmatrix}. \tag{12}$$

Therefore given a background $\bar{V}_4$ its "gauge-free" perturbation is any member of the family of space-times embedded in M_{4+n} with metric given by (10). These perturbations satisfy Einstein's vacuum field equations $G_{ij}(g) = 0$ derived from (22–I) as $A_{iAB} \to 0$. Interesting enough, when x^{A} is sufficiently small that equation produces a linear wave equation for b_{ijA}. Indeed for $g^{(1)}_{ij} = \bar{g}_{ij} + x^{A}b_{ijA}$ and following the general procedure of linearization[3] we obtain $\Box^{2}b_{ijA} = 0$, where $\Box^{2}$ is calculated with respect to the background $\bar{g}_{ij}$. Therefore for an observer in $\bar{V}_4$ it is the second quadratic form rather than the metric which is interpretable as the graviton.

Unlike Geroch's space of space-times, which is bounded at the background ($x^{A} = 0$) only, here we have another natural limit. Indeed, since $\det \gamma' \neq 0$ then $\det g \neq 0$ and from (10) it follows that the parameters x^{A} are subjected to

$$\det(\bar{g}_{im} - x^{A}b_{imA}) \neq 0. \tag{13}$$

In other words, x^{A} must not coincide with any of the curvature radii ρ^{A}_{m} of $\bar{V}_4$ corresponding to a principal direction dx^{m} and one of the normals N_{A} (Ref. 4). Notice that (13) is trivially satisfied when $x^{A} = 0$. Therefore the allowed domain of the perturbation parameters is $x^{A} \in [0, a(\rho)]$ where $a(\rho) < \rho^{A}_{m}$ for any values of A and m. In order to allow for group properties we should also include negative values of x^{A} so that we could expand the range of x^{A} to $x^{A} \in [-a(\rho), a(\rho)]$. The resulting picture is that of a local strip in M_{4+n} filled with space-time perturbations (such a picture also holds with slight modifications when the A_{iAB} are present) generated by b_{ijA}. When $a(\rho)$ is small these perturbations appear as oscillations of b_{ijA} around the background geometry. Evidently the values of $a(\rho)$ depend on the geometry of the background but from the point of view of high-dimensional physics, some other assumptions may be required to determine its size. [For example assuming a gravitational Casimir-like force between those boundaries then under certain conditions $a(\rho)$ becomes of the order of Planck's length.[5]]

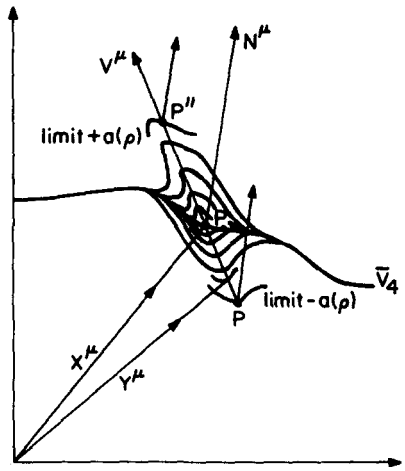


FIG. 1. Local space-time perturbations induced by b_{ija}.

A straightforward comparison with Kaluza–Klein theory fails when we ask about the nature of the ground state, which in that theory is of the form $M_4 \times B_n$, where M_4 is Minkowski's space and B_n is a compact space of small diameter. This compactness would respond for the periodic behavior of the extra dimensions. Here we have a local space which is bounded at $x^{A} = \pm a(\rho)$ but the desired periodicity is not apparent. It is our view that this periodicity is provided by the identification maps. Indeed, if all points of the family of space-times along the orbit of $h_{s}(p)$ are mapped into a single point of physical space, then for a four-dimensional observer sitting in that space the sequence of space-time perturbations will be seen with a periodicity in x^{A}. Therefore that observer may think of B_n as a compacted space in the geometrical sense. However all that is required is that B_n is compact in the topological sense. That is, bounded with identified points. This construction seems to be equivalent to the Geroch–Mansouri–Witten dimensional reduction procedure.[6,7]

ACKNOWLEDGMENTS

This paper was partially supported by the Conselho Nacional de Pesquisas (CNPq), Brazil.

The author is indebted to the warm hospitality of the University of Washington where this work was done.

[1] M. D. Maia, J. Math. Phys. **28**, 647 (1987).
[2] M. D. Maia, Phys. Rev. D **31**, 262 (1985).
[3] V. D. Zakharov, *Gravitational Waves in Einstein's Theory* (Halsted Willey, New York, 1973).
[4] L. P. Eisenhart, *Riemannian Geometry* (Princeton U. P., Princeton, NJ, 1966), pp. 143ff.
[5] T. Appelquist, A. Chodos, and E. Myers, Phys. Lett. B **127**, 51 (1983); A. Chodos and E. Myers, "The gravitational Casimir energy in Non-Abelian Kaluza–Klein theories," Brookhaven Natl. Lab preprint BNL35799, 1985.
[6] R. P. Geroch, J. Math Phys. **12**, 918 (1971).
[7] F. Mansouri and L. Witten, Found. Phys. **14**, 1095 (1984).

On the generalization of Szafron solutions of Einstein field equations

C. Bona and J. Stela
Departament de Física, Universitat Illes Balears, 07071 Palma de Mallorca, Spain
P. Palou
Institut Politècnic de Formació Profesional, 07071 Palma de Mallorca, Spain

(Received 6 June 1986; accepted for publication 12 November 1986)

The explicit form of the solutions of the Einstein field equations corresponding to a perfect fluid in geodesic, hypersurface-orthogonal motion is given with the following restrictions: (i) the comoving hypersurfaces are flat; and (ii) the second fundamental form of these surfaces is degenerate. These results are a natural extension of the metrics previously found by Szafron and co-workers [D. A. Szafron and J. Wainwright, J. Math. Phys. **18**, 1668 (1977); D. A. Szafron, *ibid.* **18**, 1673 (1977); D. A. Szafron and C. B. Collins, *ibid.* **20**, 2354 (1979)] as a perfect fluid generalization of the Szekeres dust solutions.

I. STATEMENT OF THE PROBLEM

Many years ago, Szafron[1] studied a large class of solutions of Einstein field equations corresponding to geodesic perfect fluid source and admitting the following form of the metric (comoving coordinates):

$$ds^2 = dt^2 - e^{2\alpha}\,dz^2 - e^{2\beta}(dx^2 + dy^2)\,, \qquad \alpha = \alpha(t,x,y,z),\quad \beta = \beta(t,x,y,z)\,. \tag{1}$$

The case of dust corresponds to the well known Szekeres metrics,[2] so that space-times with metric (1) corresponding to a perfect fluid source were called "Szekeres models." They were characterized in a geometric way by Szafron and Collins[3] as being those geodesic perfect fluid solutions of Einstein equations with conformally flat comoving slices so that both the second fundamental form and the Ricci tensor of every slice are degenerate (they possess at least two equal eigenvalues).

In this paper, we will study the subset of all Szekeres models such that the comoving surfaces are flat. This study fits into the program recently proposed by Stephani and Wolf[4] for finding such kind of solutions. To be concrete, we will obtain the explicit form of the metrics with the following properties.

(i) The matter content is a perfect fluid whose flow lines form a geodesic congruence orthogonal to a family S of spacelike hypersurfaces.

(ii) Each spacelike surface Σ in the family S is flat.

(iii) The second fundamental form of the hypersurfaces Σ of S is degenerate, that is, it possess at least two equal eigenvalues.

Our starting point will be the equations for α and β in (1) given by Szafron[1]; we shall adopt the same notation, in particular $(\,)' = \partial/\partial z$, $(\,)\dot{} = \partial/\partial t$, and we will use the pair of complex variables $\xi = x + iy$, $\bar{\xi} = x - iy$. The form of the equations is different when $\beta' \neq 0$ (class I) or when $\beta' = 0$ (class II), so that we will treat both cases separately.

II. METRICS OF CLASS I, $\beta' \neq 0$

The Szekeres models of class I are metrics of the form (1) with α and β defined as follows[1]:

$$\beta = \log\phi(t,z) + \nu(z,\xi,\bar{\xi})\,, \tag{2a}$$

$$e^{\alpha} = \phi' + \phi\nu'\,, \tag{2b}$$

$$e^{-\nu} = A(z)\xi\bar{\xi} + B(z)\xi + \bar{B}(z)\bar{\xi} + C(z)\,, \tag{2c}$$

where $A(z)$ and $C(z)$ are real functions, $B(z)$ is complex with

$$AC - B\bar{B} = 1/4[1 + k(z)]\,, \tag{3}$$

and $\phi(t,z)$ verifies the differential equation

$$2\ddot{\phi}/\phi + (\dot{\phi}/\phi)^2 + \kappa p(t) + k(z)/\phi^2 = 0\,, \tag{4}$$

where $p(t)$ is the pressure of the fluid and κ is the constant appearing in the Einstein field equations.

Proposition: The necessary and sufficient condition for the comoving three-dimensional slices in a class I Szekeres model to be flat is $k(z) = 0$.

Proof: It follows from a straightforward computation of the three-dimensional Ricci tensor. A partial result (sufficiency of the condition) was stated in Ref. 1 for a particular form of $\phi(t,z)$.

Allowing for this result, we will look for the metrics (1) with α and β defined by (2) with the following restrictions:

$$AC - B\bar{B} = \tfrac{1}{4}\,, \tag{5}$$

$$2\ddot{\phi}/\phi + (\dot{\phi}/\phi)^2 + \kappa p(t) = 0\,, \tag{6}$$

the last condition being just the propagation equation for the length scale in the Friedmann ($k = 0$) solution. Let us perform the standard substitution $\phi = G^{2/3}$. Then the equation becomes

$$\ddot{G} + \tfrac{3}{4}\kappa p(t) G = 0\,, \tag{7}$$

which is linear and of second order in G.

The general solution of (7) is

$$G(t,z) = [a(z) + b(z) f(t)](\dot{f})^{-1/2}\,, \tag{8}$$

where a and b are arbitrary functions of z and $f(t)$ is related with $p(t)$ by the following equation:

$$\dddot{f}/\dot{f} - \tfrac{3}{2}(\ddot{f}/\dot{f})^2 = \tfrac{3}{2}\kappa p(t)\,. \tag{9}$$

Theorem: The general form of the metrics of class I Szekeres models admitting flat comoving slices can be expressed as in Eqs. (1) and (2) with the following restrictions:

$$AC - B\bar{B} = \tfrac{1}{4}\,, \tag{10a}$$

$$\phi(t,z) = [a(z) + b(z) f(t)]^{2/3}(\dot{f})^{-1/3}\,, \tag{10b}$$

 0022-2488/87/030654-03$02.50

where $a(z)$, $b(z)$, and $f(t)$ are arbitrary functions of their arguments.

Proof: As far as the function $p(t)$ appearing in (6) was arbitrary, we can interpret (9) as a mere definition of $p(t)$ once $f(t)$ is given or vice versa. The function $\phi(t,z)$ given in (10b) is then the general solution of Eq. (6) corresponding to the function $p(t)$ defined by (9).

We have computed the energy density and the kinematical quantities associated to the fluid motion corresponding to this set of solutions. The results are given in Appendix A.

III. METRICS OF CLASS II, $\beta'=0$

The class II Szekeres models are again metrics of the form (1) with α and β defined now as follows[1]:

$$\beta = \log\phi(t) + \nu(\xi,\bar{\xi}) , \tag{11a}$$

$$e^{\alpha} = \lambda(t,z) + \phi(t)\sigma(z,\xi,\bar{\xi}) , \tag{11b}$$

$$e^{-\nu} = 1 + k/4\xi\bar{\xi} , \tag{11c}$$

$$\sigma = e^{\nu}[U(z)\xi\bar{\xi} + V(z)\xi + \bar{V}(z)\bar{\xi} + W(z)] , \tag{11d}$$

where $U(z)$ and $W(z)$ are real functions and $V(z)$ is complex, $k = 0, \pm 1$, and the functions $\phi(t)$ and $\lambda(t,z)$ are restricted by the differential equations

$$2\ddot{\phi}/\phi + (\dot{\phi}/\phi)^2 + \kappa p + k/\phi^2 = 0 , \tag{12}$$

$$\ddot{\lambda}\phi + \dot{\lambda}\dot{\phi} + \lambda\ddot{\phi} + \kappa p\lambda\phi = U(z) + k/4W(z) . \tag{13}$$

Proposition: The necessary and sufficient conditions for the comoving three-dimensional slices in a class II Szekeres model to be flat are $k = 0$, $U(z) = 0$.

Proof: It follows from a straightforward computation of the Ricci tensor. The sufficiency of the condition for some particular cases has been stated by Bonnor and Tomimura.[5]

In our case ($k = 0$), Eq. (12) reduces to Eq. (6). This suggests introducing again the auxiliary function $f(t)$, namely,

$$\phi(t) = (\dot{f})^{-1/3} , \tag{14}$$

so that we are led to the same expression (9) for $p(t)$ as in Sec. II. The general solution $\lambda(t,z)$ of the linear equation (13) in the case $U(z) = 0$, $k = 0$ is easily expressed in terms of $f(t)$:

$$\lambda(t,z) = [A(z)f(t) + B(z)](\dot{f})^{-1/3} , \tag{15}$$

where A and B are arbitrary functions of z. Note that, allowing for the definitions of α and β in (11), the arbitrary function $W(z)$ can be reabsorbed into $B(z)$ or vice versa.

We collect now our results into the following theorem.

Theorem: The general form of the metrics of class II Szekeres models admitting flat comoving slices can be expressed as in Eq. (1) with the following restrictions:

$$e^{\beta} = (\dot{f})^{-1/3} , \tag{16a}$$

$$e^{\alpha} = \lambda(t,z) + [xa(z) + yb(z)](\dot{f})^{-1/3} , \tag{16b}$$

$$\lambda(t,z) = [A(z)f(t) + B(z)](\dot{f})^{-1/3} , \tag{16c}$$

where A, B, a, b, and f are arbitrary functions of their arguments.

The expressions of the energy density and the kinematical quantities associated to the fluid motion corresponding to these metrics are given in Appendix B.

IV. CONCLUSIONS

We have obtained the explicit form of the metrics of all Szekeres models admitting flat comoving slices, that is, the metrics that fulfill conditions (i)–(iii) as given in Sec. I.

The class I solutions given in (10) are a generalization of the Szafron solutions[1] corresponding to

$$\kappa p = \tfrac{4}{3}q(1-q)t^{-2} , \tag{17}$$

$$\phi(t,z) = [g(z)t^{1-q} + h(z)t^{q}]^{2/3} , \tag{18}$$

as can be seen by substituting $f(t) = t^{1-2q}$ into Eq. (10). In the case $q = \frac{1}{2}$, however, we obtain a more general result, namely,

$$\phi(t,z) = [a(z) + b(z)\log(t)]^{2/3}t^{1/3} , \tag{19}$$

which is the complete solution of Eq. (6) with $p(t)$ given by (17) in the $q = \frac{1}{2}$ case.

Szafron and Wainwright[6] have obtained explicit expressions for class II Szekeres models with a time dependence of the pressure as given in (17). The subset of their solutions corresponding to our case $U(z) = 0$ [$C(z) = 0$ in their notation] is given by

$$\phi(t) = [C_1 t^{1-q} + C_2 t^{q}]^{2/3} , \tag{20}$$

where C_1 and C_2 are constants.

The geometrical properties of the whole class of Szekeres models, including an invariant classification of them, are given in Ref. 3. In the case of the solutions given in (10) and (16), their Killing structure can be obtained directly from the tables given in Refs. 1 and 6. The spherically symmetric case can then be seen to correspond to the class I metrics (10) with A, B, and C constant. This case has been considered in a recent work[7] as a generalization of the well-known Tolman ($k = 0$) dust solutions.[8]

ACKNOWLEDGMENTS

The authors are indebted to the unknown referee, who made useful suggestions concerning the final presentation of the results.

We also acknowledge financial support under CAICYT Project No. 1005/84.

APPENDIX A: SOME RESULTS FOR CLASS I METRICS

We list here some results concerning metrics of class I [described by Eq. (10)]. Notation is the same as Sec. II. The energy density μ of the fluid is given by

$$\kappa\mu = \tfrac{1}{3}\dot{f}^2[(1/\dot{f})^{\cdot} + 2b/F] \times [(1/\dot{f})^{\cdot} + 2(b' + 3/2bv')/(F' + 3/2Fv')] , \tag{A1}$$

where we have noted

$$F(t,z) = a(z) + b(z)f(t) , \tag{A2}$$

the pressure p being defined by Eq. (9) in the text.

The expansion θ of the fluid is

$$\theta = \dot{f}[(1/\dot{f})^{\cdot} + b/F + (b' + 3/2bv')/(F' + 3/2Fv')] \tag{A3}$$

and the components of the shear tensor are

$$\sigma_x^x = \sigma_y^y = -\tfrac{1}{2}\sigma_z^z$$
$$= \tfrac{1}{3}\dot{f}\,[b/F - (b' + 3/2bv')/(F' + 3/2Fv')]\,. \quad \text{(A4)}$$

APPENDIX B: SOME RESULTS FOR CLASS II METRICS

We list here some results concerning metrics of class II [described by Eq. (16)]. Notation is the same as in Sec. III. In this case,

$$\kappa\mu = -\tfrac{1}{3}\ddot{f}\,[(1/\dot{f})^{\cdot} + 2A(z)/H] \quad \text{(B1)}$$

is the energy density, where we have noted

$$H(t,x,y,z) = A(z)\,f(t) + B(z) + a(z)x + b(z)y\,, \quad \text{(B2)}$$

the pressure being defined by Eq. (9) in the text.

The expansion of the fluid is given by

$$\theta = \dot{f}\,[(1/\dot{f})^{\cdot} + A(z)/H] \quad \text{(B3)}$$

and the components of the shear tensor are

$$\tfrac{1}{2}\sigma_z^z = -\sigma_x^x = -\sigma_y^y = \tfrac{1}{3}\dot{f}A(z)/H\,. \quad \text{(B4)}$$

[1] D. A. Szafron, J. Math. Phys. **18**, 1673 (1977).
[2] P. Szekeres, Commun. Math. Phys. **41**, 55 (1975).
[3] D. A. Szafron and C. B. Collins, J. Math. Phys. **20**, 2354 (1979).
[4] H. Stephani and Th. Wolf, in *Galaxies, Axisymmetric Systems and Relativity*, edited by M. A. H. MacCallum (Cambridge U.P., Cambridge, 1985).
[5] W. B. Bonnor and N. Tomimura, Mon. Not. R. Astron. Soc. **175**, 85 (1976).
[6] D. A. Szafron and J. Wainwright, J. Math. Phys. **18**, 1668 (1977).
[7] C. Bona, J. Stela, and P. Palou, "Perfect fluid spheres admitting flat 3-dimensional slices," Gen. Relativ. Gravit. (to be published).
[8] R. C. Tolman, Proc. Natl. Acad. Sci. USA **20**, 169 (1934).

Bianchi VI_0 viscous fluid cosmology with magnetic field

Marcelo Byrro Ribeiro
Observatório Nacional, CEP 20921, Rio de Janeiro, Brazil

Abhik Kumar Sanyal
Department of Physics, Jadavpur University, Calcutta-700032, India

(Received 24 October 1985; accepted for publication 12 November 1986)

A spatially homogeneous Bianchi type VI_0 model containing a viscous fluid in the presence of an axial magnetic field has been studied. A barotropic equation of state together with a pair of linear relations among the square root of matter density, shear scalar, and expansion scalar have been assumed. Solutions are obtained in the presence of a magnetic field, only in two special cases, which are comparatively easy. The complete solutions for this model in the absence of a magnetic field are also obtained. The presence of a magnetic field in the former case, however, does not in effect cause any major modification in the fundamental nature of the initial singularity of the expanding model.

I. INTRODUCTION

The investigation of cosmological models in Einstein's theory usually chooses the energy momentum tensor of matter as that due to a perfect fluid. These models lead to an initial singular state. Of course, it is important to investigate more realistic models that take into account dissipative processes due to viscosity.

The first suggestion was investigated by Misner[1] and he proposed that the neutrino viscosity acting in the early era might have considerably reduced the present anisotropy of the black-body radiation during the process of evolution. Murphy[2] in 1973 showed that the bulk viscosity can push the initial singularity in Friedman universe to the infinite past but at the cost of violating the Hawking–Penrose energy conditions. Belinskii and Khalatnikov[3] studied the behavior of anisotropic spatially homogeneous models with viscous fluid in the asymptotic limits. They assumed that the fluid viscosity coefficients could be expressed as power functions of the matter density. It was found by them that the dissipative mechanism due to the presence of viscosity not only modifies the nature of the initial big bang singularity, but also can account for the anomalously large entropy per baryon in the present day universe. Similar properties were shown by Banerjee, Duttachoudhury, and Sanyal[4] by constructing particular Bianchi I models consisting of a viscous fluid. Other models with viscosity terms included in the stress energy tensor were constructed by Banerjee and Santos,[5,6] Banerjee, Duttachoudhury, and Sanyal,[7] Coley and Tupper,[8,9] and Santos, Dias, and Banerjee.[10] Also problems with axial magnetic fields in Bianchi I and III and Kantowski–Sachs viscous fluid models were previously investigated by Banerjee and Sanyal.[11] Though the recently developed theory of inflationary cosmology,[12] using GUT, claims to have given a plausible explanation for the outstanding cosmological problems, such as the high degree of isotropy and large entropy per baryon in the present universe, the theory itself appears to be incomplete yet in many aspects. It is therefore worthwhile to investigate if the classical relativity theory is successful in dealing with the above problems by introducing dissipative phenomena in the matter content of the universe.

In this paper we proceed to investigate the Bianchi VI_0 model filled with a viscous fluid characterized by both bulk and shear viscosities including a magnetic field in the axial direction. Evidently the task of obtaining exact solutions in a viscous fluid model becomes more difficult than the corresponding perfect fluid case, due to a larger number of unknown quantities to be determined. In this way, in the present paper we attempt to find exact solutions under the assumption that the ratio of the shear to expansion rate (σ/θ) and the density to the square of the expansion (ρ/θ^2) were both constants. The perfect fluid solutions for the Bianchi II model with these assumptions were first obtained by Collins and Stewart.[13]

In Sec. II we consider Einstein's field equations for a Bianchi VI_0 cosmological model and show the dynamical importance of matter density and shear scalar. The entropy variation is also explicitly stated.

In Sec. III we obtain two particular solutions in the presence of the magnetic field and complete solutions in the absence of it.

II. EINSTEIN'S FIELD EQUATIONS AND SOME GENERAL RESULTS

The metric for the spatially homogeneous Bianchi VI_0 space time is taken in the following form:

$$ds^2 = -dt^2 + e^{2\alpha}dx^2 + e^{2(\beta+mx)}dy^2 + e^{2(\gamma-mx)}dz^2, \tag{2.1}$$

where α, β, and γ are functions of time alone and m is a constant. The energy momentum tensor for a viscous fluid is

$$T_\mu{}^\nu = (\rho+\bar{p})v_\mu v^\nu + \bar{p}\delta_\mu{}^\nu - \eta U_\mu{}^\nu, \tag{2.2}$$

with

$$\bar{p} = p - (\zeta - \tfrac{2}{3}\eta)v^\alpha{}_{;\alpha}, \tag{2.3}$$

and

$$U_{\mu\nu} = v_{\mu;\nu} + v_{\nu;\mu} + v_\mu v^\beta v_{\nu;\beta} + v_\nu v^\beta v_{\mu;\beta}.$$

In the above p is the thermodynamic pressure and η and ζ are the shear viscosity and bulk viscosity coefficients, respectively. Here v^μ is the four-velocity vector so that $v_\mu v^\mu = -1$. Since there is a magnetic field along the x direction, we have F_{23} as the only nonvanishing component of the electromagnetic field tensor. From Maxwell's equation it can easily be

 0022-2488/87/030657-04$02.50

seen that $F_{23}=A$, where A is a constant of integration. If we have for the stress energy tensor of electromagnetic field the expression

$$E_\mu{}^\nu=(1/4\pi)\left[F_{\mu\alpha}F^{\nu\alpha}-\tfrac{1}{4}\delta_\mu{}^\nu F_{\alpha\beta}F^{\alpha\beta}\right],$$

the nonvanishing components are

$$-E_0^0=-E_1^1=E_2^2=E_3^3=(A^2/8\pi)e^{2(\beta+\gamma)}. \tag{2.4}$$

Einstein's field equations are (choosing $8\pi G=C=1$)

$$R_\mu{}^\nu-\tfrac{1}{2}\delta_\mu{}^\nu R=-(T_\mu{}^\nu+E_\mu{}^\nu), \tag{2.5}$$

and in the comoving coordinates $v^\mu=\delta_0{}^\mu$. Thus in view of Eq. (2.5) and using Eqs. (2.1)–(2.4) we find the following equations:

$$\tfrac{2}{2}\dot R^2/R^2-\tfrac{1}{2}(\dot\alpha^2+\dot\beta^2+\dot\gamma^2)-m^2e^{-2\alpha}$$
$$=+(A^2/8\pi)e^{-2(\beta+\gamma)}, \tag{2.6a}$$

$$\ddot\beta+\ddot\gamma+\tfrac{3}{2}(\dot R/R)(\dot\beta+\dot\gamma-\dot\alpha)+\tfrac{1}{2}(\dot\alpha^2+\dot\beta^2+\dot\gamma^2)$$
$$+m^2e^{-2\alpha}=-(\bar p-2\eta\dot\alpha)+(A^2/8\pi)e^{-2(\beta+\gamma)}, \tag{2.6b}$$

$$\ddot\gamma+\ddot\alpha+\tfrac{3}{2}(\dot R/R)(\dot\gamma+\dot\alpha-\dot\beta)+\tfrac{1}{2}(\dot\alpha^2+\dot\beta^2+\dot\gamma^2)$$
$$-m^2e^{-2\alpha}=-(\bar p-2\eta\dot\beta)-(A^2/8\pi)e^{-2(\beta+\gamma)}, \tag{2.6c}$$

$$\ddot\alpha+\ddot\beta+\tfrac{3}{2}(\dot R/R)(\dot\alpha+\dot\beta-\dot\gamma)+\tfrac{1}{2}(\dot\alpha^2+\dot\beta^2+\dot\gamma^2)$$
$$-m^2e^{-2\alpha}=-(\bar p-2\eta\dot\gamma)-(A^2/8\pi)e^{-2(\beta+\gamma)}, \tag{2.6d}$$

and $\dot\beta-\dot\gamma=0$.

A dot represents time differentiation and R stands for

$$R^3=\exp(\alpha+\beta+\gamma). \tag{2.7}$$

In view of Eq. (2.6a) and with a suitable coordinate transformation we can have

$$\beta=\gamma. \tag{2.8}$$

Combining the field equations (2.6a)–(2.6d) and using Eq. (2.8), we get the following set of equations:

$$\tfrac{1}{3}\theta^2-\rho-\sigma^2=m^2e^{-2\alpha}+n^2e^{-4\beta}, \tag{2.9a}$$

$$\ddot\beta+(\theta+2\eta)\dot\beta-(\rho-p)/2-\tfrac{1}{2}\zeta\theta-\tfrac{2}{3}\eta\theta=n^2e^{-4\beta}, \tag{2.9b}$$

$$\ddot\beta+(\theta+2\eta)\dot\beta-(\rho+p)+\zeta\theta-\tfrac{2}{3}\eta\theta-\tfrac{1}{3}\theta^2-2\sigma^2-\dot\theta$$
$$=2n^2e^{-4\beta}. \tag{2.9c}$$

Here we have used the relation (2.9a) to derive the other two. In the set of equations (2.9a)–(2.9c), $A^2/8\pi$ has been replaced by n^2. The expansion and the shear scalars θ and σ^2 are defined in the usual way:

$$\theta=v^\mu{}_{;\mu}=\dot\alpha+2\dot\beta, \tag{2.10}$$

and

$$2\sigma^2=\sigma_{\mu\nu}\sigma^{\mu\nu}=\dot\alpha^2+2\dot\beta^2-\tfrac{1}{3}\theta^2, \tag{2.11}$$

where the shear tensor $\sigma_{\mu\nu}$ has the usual expression

$$\sigma_{\mu\nu}=\tfrac{1}{2}(v_{\mu;\nu}+v_{\nu;\mu})+\tfrac{1}{2}(v_\mu v^\beta v_{\nu;\beta}+v_\nu v^\beta v_{\mu;\beta})$$
$$+\tfrac{1}{3}(g_{\mu\nu}+v_\mu v_\nu)\theta.$$

From Eqs. (2.9b) and (2.9c) we get

$$\dot\theta=-2\sigma^2-\tfrac{1}{3}\theta^2-\tfrac{1}{2}[\rho+3(p-\zeta\theta)]-n^2e^{-4\beta}, \tag{2.12}$$

and the divergence relation $(T^{\mu\nu}+E^{\mu\nu})_{;\nu}=0$ yields

$$\dot\rho=-(\rho+p)\theta+\zeta\theta^2+4\eta\sigma^2. \tag{2.13}$$

It is interesting to observe from Eq. (2.12) that for a contracting model (that is, for $\theta<0$) the time derivative for the expansion scalar θ is less than zero. It means that θ remains negative always and thus collapse cannot be halted for a physically reasonable fluid ($\rho>0$, $p>0$). On the other hand, if the bulk viscosity ζ is very small and can be ignored, one has $\dot\theta<0$ independent of whether the model is expanding or contracting. Thus there may be a maximum but no minimum of the volume. One can easily verify that in both the cases $R_{\mu\nu}v^\mu v^\nu<0$, and thus Hawking's energy condition is satisfied.

Another relation showing the dynamical importance of matter density and shear scalars can be derived in view of the field equations (2.9a)–(2.9c) and also using Eqs. (2.12) and (2.13). This expression is explicitly given in the form

$$(\rho/\theta^2)^\cdot=-[(1/\theta^2)(\sigma^2+m^2e^{-2\alpha}+n^2e^{-4\beta})]^\cdot$$
$$=(\sigma^2/\theta^2)[3(\rho-p)\theta^{-1}+3\zeta+4\eta]$$
$$+(m^2e^{-2\alpha}/\theta^2)[3\zeta-(\rho+3p)\theta^{-1}]$$
$$+(n^2e^{-4\beta}/\theta^2)[3\zeta+(\rho-3p)\theta^{-1}]. \tag{2.14}$$

From the definitions of θ and σ given by Eqs. (2.10) and (2.11) it is possible to write

$$\sigma^2=\tfrac{1}{3}\theta^2-\dot\beta(2\theta-3\dot\beta), \tag{2.15}$$

which in turn yields

$$\dot\beta=(\theta/3\pm\sigma/\sqrt{3}). \tag{2.16}$$

Now differentiating Eq. (2.15) with respect to time and substituting $\ddot\beta$ from the field equation (2.9b) and $\dot\beta$ and $\dot\theta$ from Eqs. (2.16) and (2.12), respectively, one can finally obtain, after a little manipulation and utilizing (2.9a), the relation for shear dissipation in the form

$$(\sigma^2)^\cdot=-2(2\eta+\theta)^2\pm(4\sigma/\sqrt{3})(n^2e^{-4\beta}-m^2e^{-2\alpha}). \tag{2.17}$$

Using Eq. (2.16) in the above relation it is possible to obtain further a very similar kind of relation

$$\frac{(\sigma^2R^6)^\cdot}{R^6}=-4\eta\sigma^2-n^2\frac{(e^{-4\beta}R^4)^\cdot}{R^4}$$
$$-m^2\frac{(e^{-2\alpha}R^2)^\cdot}{R^2}. \tag{2.18}$$

The relations (2.17) and (2.18) are generalizations of the corresponding equations derived for Bianchi I space-time in a previous communication.[4] Further, as considered by Belinskii and Khalatnikov,[3] let the time derivative of the entropy density be

$$\dot\Sigma/\Sigma=\dot\rho/(\rho+p),$$

where Σ is the entropy density. The total entropy can be defined as $s=R^3\Sigma$, the time derivative of which can be found in view of Eqs. (2.13) and the above one, as

$$\dot s/s=(\zeta\theta^2+4\eta\sigma^2)/(\rho+p). \tag{2.19}$$

Now, since $\rho+p>0$ and $\zeta>0$, $\eta>0$, so $\dot s>0$, which implies

that the total entropy will always increase with the change of proper time irrespective of any model (expanding or contracting).

III. EXACT SOLUTIONS OF EINSTEIN'S FIELD EQUATIONS

We have obtained a set of three field equations [viz., (2.9a)–(2.9c)] with six unknown quantities (viz., $\alpha, \beta, \rho, p, \eta$, and ζ) to be determined. So in order to obtain exact solutions of the field equations we consider three more physically reasonable equations: one is a barotropic equation of state between matter density and thermodynamic pressure and the other two are a pair of linear relations connecting matter density, expansion, and shear scalars. They are

$$p = \epsilon\rho, \quad \rho = C^2\theta^2, \quad \sigma^2 = D^2\theta^2, \tag{3.1}$$

where C and D are two constant quantities. Hence the field equations (2.9a)–(2.9c) can now be written in view of Eqs. (2.10)–(2.13) and (3.1) as

$$(\tfrac{1}{3} - C^2 - D^2)\theta^2 = m^2 e^{-2\alpha} + n^2 e^{-4\beta}, \tag{3.2a}$$

$$\ddot{\beta} + (\theta + 2\eta)\dot{\beta} - [(1-\epsilon)/2]C^2\theta^2 - \tfrac{1}{2}\zeta\theta - \tfrac{2}{3}\eta\theta = n^2 e^{-4\beta}, \tag{3.2b}$$

$$\ddot{\beta} + (\theta + 2\eta)\dot{\beta} - [\tfrac{1}{3} + (1+\epsilon)C^2 + 2D^2]\theta^2 - \zeta\theta - \tfrac{2}{3}\eta\theta - \dot{\theta} = 2n^2 e^{-4\beta}. \tag{3.2c}$$

In view of Eqs. (2.16) and (3.1) we have

$$\dot{\beta} = (\tfrac{1}{3} \pm D/\sqrt{3})\,\theta, \tag{3.3}$$

which, when substituted in Eq. (2.10), gives us

$$\dot{\alpha} = (\tfrac{1}{3} \mp 2D/\sqrt{3})\theta. \tag{3.4}$$

The above two equations [(3.3) and (3.4)] lead to the relation $\dot{\alpha} = a\dot{\beta}$, where $a = (\tfrac{1}{3} \mp 2D/\sqrt{3})/(\tfrac{1}{3} \pm D/\sqrt{3})$, so that

$$e^{-2\alpha} = be^{-2a\beta}. \tag{3.5}$$

In Eq. (3.5) b is an integration constant of positive magnitude. Now in view of Eqs. (3.3) and (3.2a) we get

$$(\tfrac{1}{3} - C^2 - D^2)/(\tfrac{1}{3} \pm D/\sqrt{3})^2 = m^2 be^{-2a\beta} + n^2 e^{-4\beta},$$

which can be written as

$$\dot{\beta} = [C_1 e^{-2a\beta} + C_2 e^{-4\beta}]^{1/2}, \tag{3.6}$$

C_1 and C_2 being two constants. It is not difficult to show that both C_1 and C_2 are greater than zero. Writing x for $e^{2\beta}$, relation (3.6) can be integrated to yield

$$\frac{1}{2}\int \frac{dx}{[C_1 x^{(2-a)} + C_2]^{1/2}} = t - t_0. \tag{3.7}$$

The explicit value for x (that is, $e^{2\beta}$) is obtainable upon choosing specific values for a. We consider here two special cases: $a = 2$ and $a = 0$. It is evident that the parameter a cannot be unity because then from its definition $D = 0$ or, in other words, the shear vanishes. When $a = 2$, we have

$$x = e^{2\beta} = 2(C_1 + C_2)^{1/2}(t - t_0) = C_3(t - t_0), \tag{3.8}$$

where C_3 is a constant and is equal to $2(C_1 + C_2)^{1/2}$. Thus from Eq. (3.5)

$$e^{2\alpha} = (1/b)e^{4\beta}. \tag{3.9}$$

The expansion scalar $\theta = 2/(t - t_0)$ and the proper volume $R^3 = e^{\alpha + 2\beta} = (C_3^2/\sqrt{b})(t - t_0)^2$. From the above solutions, one can conclude that for such a model as $t \to t_0$, $R^3 = 0$, i.e., the proper volume vanishes, the expansion scalar $\theta \to \infty$, and in consequence $\rho \to \infty$, $\sigma^2 \to \infty$. It is a point singularity. The magnetic field B being proportional to $e^{-2\beta}$ also increases to an indefinitely large value at the singularity. On the other hand as $t \to \infty$, we have $\theta \to 0$, $\sigma^2 \to 0$, $\rho \to 0$, and $R^3 \to \infty$. The second case is for $a = 0$, when we have the integral

$$\frac{1}{2}\int \frac{dx}{(C_1 x^2 + C_2)^{1/2}} = t - t_0. \tag{3.10}$$

On integration we obtain, since $C_2 > 0$,

$$(1/2\sqrt{C_1})\,\ln[x\sqrt{C_1} + \sqrt{C_1 x^2 + C_2}] = t - t_0,$$

so that the solution for $e^{2\beta}$ is given by

$$x = e^{2\beta} = (1/2\sqrt{C_1})\, e^{-2\sqrt{C_1}(t - t_0)}[e^{4\sqrt{C_1}(t - t_0)} - C_2]. \tag{3.11}$$

In this case $\alpha = \text{const}$. Now as $t \to t_1$, such that

$$e^{4\sqrt{C_1}(t - t_0)} = C_2,$$

we have $e^{2\beta} \to 0$, the proper volume $R^3 \to 0$, $\theta \to \infty$, so that $\rho \to \infty$, $\sigma^2 \to \infty$, and the magnetic field $B \to \infty$. On the other hand, as t increases $e^{2\beta}$ increases and approaches an infinitely large magnitude as $t \to \infty$. In this limit $\theta \to 2\sqrt{C_1}$, that is, a finite magnitude so that the scalars like ρ, σ^2, etc. also remain finite.

Now, eliminating both ζ and p from Eqs. (2.12) and (2.13) and using Eq. (3.1) the explicit expression for the shear viscosity coefficient can be obtained. This is given by

$$\eta = (1/2D^2)[(C^2 - \tfrac{1}{3})\dot{\theta}/\theta - \tfrac{1}{3}\theta(2D^2 + \tfrac{1}{3} - C^2) - \tfrac{1}{3}n^2 e^{-4\beta}/\theta]. \tag{3.12}$$

When the metric is known the exact magnitude of η can be calculated independently of any equation of state relating density and pressure of the fluid. But the calculations for the bulk viscosity coefficient ζ involve pressure and therefore one has to know the pressure in order to write the final form of ζ. Let us assume the barotropic equation of state $p = \epsilon\rho$, as considered earlier in Eq. (3.1), to be valid for the fluid under consideration. In this case the relation (2.13) yields

$$\zeta = \tfrac{2}{3}[\dot{\theta}/\theta + (\tfrac{1}{3} + 2D^2 + [(1+3)/2]C^2)\theta + n^2 e^{-4\beta}/\theta]. \tag{3.13}$$

We now consider a more simple case where there is no magnetic field. Here one can obtain the exact solution for Bianchi type VI_0 spatially homogeneous space-time filled with viscous fluid. Now since a magnetic field is absent, we have $n^2 = 0$ and from Eq. (3.2a) using (3.4) we obtain the equation

$$[(\tfrac{1}{3} - C^2 - D^2)/(\tfrac{1}{3} \mp (2/\sqrt{3})D)^2]\dot{\alpha}^2 = m^2 e^{-2\alpha}. \tag{3.14}$$

Integrating Eq. (3.14) and with a suitable time transformation we get the solutions for α and β as

$$e^{\alpha} = t, \quad e^{\beta} = l_0 t^{1/a}, \tag{3.15}$$

where a is the same constant as mentioned in Eq. (3.5) and l_0 is another integration constant. The proper volume and expansion scalars are given by

$$R^3 = e^{\alpha + 2\beta} = l_0 t^{(1+2/a)}, \tag{3.16}$$

and

$$\theta = (1 + 2/a)/t, \tag{3.17}$$

where $1 + 2/a = (\frac{1}{3} \mp 2D/\sqrt{3})$.

From (3.17) it is evident that $\dot{\theta}$ is proportional to θ^2 and further in this case $n^2 = 0$, so that the relations (3.12) and (3.13) lead us to the conclusion that both the shear and bulk viscosity coefficients are proportional to the expansion scalar θ. This in turn suggests that these viscosity coefficients are linearly proportional to the square root of the matter density, i.e., $\eta = \eta_0 \rho^{1/2}$, $\zeta = \zeta_0 \rho^{1/2}$, with η_0 and ζ_0 being constants. Now since $\frac{1}{3} - C^2 - D^2 > 0$, we have $(\frac{1}{3} \pm D/\sqrt{3}) > 0$, but $(1 \mp (2/\sqrt{3})D)$ may be greater than or less than zero depending on the magnitude of D. For an expanding case $\theta > 0$ and one finds that $(1 + 2/a) > 0$, that is, $(1 \mp (2/\sqrt{3})D) > 0$. In this case as $t \to 0$ we have the proper volume $R^3 \to 0$, and the expansion scalar $\theta \to \infty$, so that the density ρ, shear σ^2, and the viscosity coefficients η and ζ all approach infinitely large magnitudes. It represents a point-like singularity, the model exploding from a singularity state and asymptotically approaching an infinite expansion stage at $t \to \infty$. In this limit ρ, σ^2, η, and ζ all vanish.

There is one particular case $[(1 - 2D/\sqrt{3})] < 0$ when the solution is different. Here $(1 + 2/a) < 0$ and so $\theta < 0$. It represents a contracting model. When $t \to 0$, the model starts from an infinitely large volume ($R^3 \to \infty$), but since $e^{\alpha} \to 0$, $e^{\beta} \to \infty$, at this epoch we may say that the model is initially in the form of an infinite disk at the start of contraction. At $t \to \infty$ we get $R^3 \to 0$, but now $e^{\alpha} \to \infty$, $e^{\beta} \to 0$, so that the singularity is in the form of a line. The peculiarity of this situation is that at this limit of zero volume the expansion scalar θ becomes vanishingly small, so that the density shear and the viscosity coefficients all vanish in this limit.

IV. CONCLUSION

In the present paper we analyzed a Bianchi VI_0 model with viscous fluid and in the presence of magnetic field in the axial direction. The viscous fluid is characterized by bulk and shear viscosities. We assumed that $\sigma^2/\theta^2 = D^2$ is a constant and $\rho/\theta^2 = C^2$ is also one. We obtained solutions for only two special cases in the presence of a magnetic field, which, however, do not change the nature of the singularity. In the absence of the magnetic field complete solutions were obtained. Here the viscosity coefficients η and ζ are found to be power functions of the fluid density being proportional to $\rho^{1/2}$. In a particular case of the latter the model is a contracting one with a peculiar feature of the density, viscosity coefficients, shear, etc. all approaching negligible values in the limit of zero volume.

ACKNOWLEDGMENTS

Thanks are due to Dr. A. Banerjee, Dr. N. O. Santos, and to the referee for their valuable suggestions.

We also wish to thank U.G.C. (India) and CNPq (Brazil) for financial support.

[1] C. W. Misner, Astrophys. J. **151**, 431 (1968).
[2] G. L. Murphy, Phys. Rev. D **8**, 4231 (1973).
[3] V. A. Belinskii and I. M. Khalatnikov, Sov. Phys. JETP **42**, 205 (1976).
[4] A. Banerjee, S. B. Duttachoudhury, and A. K. Sanyal, J. Math. Phys. **26**, 3010 (1985).
[5] A. Banerjee and N. O. Santos, J. Math. Phys. **24**, 2689 (1983).
[6] A. Banerjee and N. O. Santos, Gen. Relativ. Gravit. **16**, 217 (1984).
[7] A. Banerjee, S. B. Duttachoudhury, and A. K. Sanyal, Gen. Relativ. Gravit. **18**, 461 (1986).
[8] A. A. Coley and B. O. J. Tupper, Phys. Rev. D **29**, 2701 (1984).
[9] A. A. Coley and B. O. J. Tupper, Astrophys. J. **271**, 1 (1983).
[10] N. O. Santos, R. S. Dias, and A. Banerjee, J. Math. Phys. **26**, 878 (1985).
[11] A. Banerjee and A. K. Sanyal, Gen. Relativ. Gravit. (in press).
[12] A. H. Guth, Phys. Rev. D **23**, 347 (1981).
[13] C. B. Collins and J. M. Stewart, Mon. Not. R. Astron. Soc. **153**, 419 (1971).

On the completion of the post-Newtonian gravitational two-body problem with spin

B. M. Barker
Department of Physics and Astronomy, The University of Alabama, Tuscaloosa, Alabama 35487
R. F. O'Connell
Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803

(Received 20 June 1986; accepted for publication 5 November 1986)

Previous work by the authors [B. M. Barker and R. F. O'Connell, Phys. Rev. D **12**, 329 (1975); **14**, 861 (1976); B. M. Barker, G. G. Byrd, and R. F. O'Connell, Astrophys. J. **305**, 623 (1986); B. M. Barker and R. F. O'Connell, Gen. Relativ. Gravit. **18**, 1055 (1986)] on the post-Newtonian (order c^{-2}) gravitational two-body problem with spin and parametrized post-Newtonian parameters γ and β was concerned with the relative position $\mathbf{r}=\mathbf{r}_1-\mathbf{r}_2$. Here this work is completed by finding the individual positions $\mathbf{r}_1$ and $\mathbf{r}_2$, which is necessary for the interpretation of certain binary-system observations. First the center of inertia $\mathbf{r}_{\rm CI}$ is found. This makes it possible to obtain the positions $\mathbf{r}_1$ and $\mathbf{r}_2$ and the center of mass $\mathbf{r}_{\rm CM}$ as a function of the relative position $\mathbf{r}$, relative velocity $\mathbf{v}$, and spin angular momenta $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ of the two bodies. Thus, if a solution $\mathbf{r}=\mathbf{r}(t)$ can be obtained, then solutions $\mathbf{r}_1=\mathbf{r}_1(t)$ and $\mathbf{r}_2=\mathbf{r}_2(t)$ can also be obtained. The final results are given in a very general coordinate system specified by four arbitrary dimensionless parameters. In particular, the spin-orbit potential energy terms V_{S1} and V_{S2} are given *without* going to a frame of reference where the total momentum is zero.

I. INTRODUCTION

In our previous work[1–4] involving equations of motion arising from the post-Newtonian gravitational two-body problem with spin, we were interested only in the relative position $\mathbf{r}$. However, for some binary systems—such as the binary pulsar[5,6] PSR 1913 + 16—it is necessary to have equations of motion for the positions $\mathbf{r}_1$ and $\mathbf{r}_2$ in order to connect theory and observation.[7–9] In Sec. II we define three coordinate systems. Our most general coordinates $\mathbf{r}_1$ and $\mathbf{r}_2$ are related to the Einstein–Infeld–Hoffman (EIH) coordinates $\mathbf{r}_{E1}$ and $\mathbf{r}_{E2}$ by four arbitrary dimensionless parameters. In Sec. III we give the spin-orbit potential energy terms V_{S1} and V_{S2} in a frame of reference where the total momentum is *not* zero and include parametrized post-Newtonian (PPN) parameters γ and β. In the Appendix, we give a more elaborate treatment of V_{S1} and V_{S2} for general relativity. In Sec. IV, we find the center of inertia $\mathbf{r}_{\rm CI}$ for our most general coordinate system and display the positions $\mathbf{r}_1$ and $\mathbf{r}_2$ and the center of mass $\mathbf{r}_{\rm CM}$ as a function of the relative position $\mathbf{r}$, relative velocity $\mathbf{v}$, and spin angular momenta $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ of the two bodies. In Sec. V we present our conclusions.

II. COORDINATE SYSTEMS

In this paper, we use coordinates $\mathbf{r}_{EN}$, $\mathbf{r}_{*N}$, and $\mathbf{r}_N$, where N for body N always equals 1 or 2. The relative coordinates $\mathbf{r}_E$, $\mathbf{r}_*$, and $\mathbf{r}$ are given by

$$\mathbf{r}_E=\mathbf{r}_{E1}-\mathbf{r}_{E2},\quad \mathbf{r}_*=\mathbf{r}_{*1}-\mathbf{r}_{*2},\quad \mathbf{r}=\mathbf{r}_1-\mathbf{r}_2. \tag{2.1}$$

The $\mathbf{r}_{*N}$ coordinates are related to the EIH coordinates[1,2] $\mathbf{r}_{EN}$ by the coordinate transformation[10]

$$\mathbf{r}_{EN}=\mathbf{r}_{*N}+(-1)^N\alpha G\left[(1-a_0)m_N+a_0\frac{m_1m_2}{m_N}\right]\frac{\mathbf{r}_*}{c^2r_*}, \tag{2.2}$$

where α and a_0 are arbitrary dimensionless parameters, m_N is the mass of body N, G is Newton's constant of gravitation, and c is the speed of light. From Eqs. (2.1) and (2.2), we obtain[1,2,10]

$$\mathbf{r}_E=\mathbf{r}_*(1-\alpha GM/c^2r_*), \tag{2.3}$$

where $M\equiv m_1+m_2$.

The $\mathbf{r}_N$ coordinates are related to the $\mathbf{r}_{*N}$ coordinates by the coordinate transformation[2]

$$\mathbf{r}_{*N}=\mathbf{r}_N-\lambda_N\,\mathbf{v}_N\times\mathbf{S}^{(N)}/m_Nc^2, \tag{2.4}$$

where λ_1 and λ_2 are arbitrary dimensionless parameters, $\mathbf{v}_N$ is the velocity of body N, and $\mathbf{S}^{(N)}$ is the spin angular momentum of body N. From Eqs. (2.1) and (2.4), we obtain[2]

$$\mathbf{r}_*=\mathbf{r}+\sum_{N=1}^{2}(-1)^N\lambda_N\frac{\mathbf{v}_N\times\mathbf{S}^{(N)}}{m_Nc^2}. \tag{2.5}$$

If we are in a frame of reference where the total momentum is equal to zero (i.e., center-of-mass system) then to first order $m_1\mathbf{v}_1+m_2\mathbf{v}_2=0$. We then obtain $\mathbf{v}_N=-(-1)^N\mu\mathbf{v}/m_N$, where $\mathbf{v}=\mathbf{v}_1-\mathbf{v}_2$ and $\mu\equiv m_1m_2/M$. Using the above in Eqs. (2.4) and (2.5) we obtain,[2] respectively (correct to the post-Newtonian approximation),

$$\mathbf{r}_{*N}=\mathbf{r}_N+(-1)^N\lambda_N\frac{\mu\mathbf{v}\times\mathbf{S}^{(N)}}{m_N^2c^2}, \tag{2.6}$$

$$\mathbf{r}_*=\mathbf{r}-\sum_{N=1}^{2}\lambda_N\frac{\mu\mathbf{v}\times\mathbf{S}^{(N)}}{m_N^2c^2}. \tag{2.7}$$

To the same approximation, we also obtain

$$\mathbf{r}_{EN}=\mathbf{r}_N+(-1)^N\alpha G\left[(1-a_0)m_N+a_0\frac{m_1m_2}{m_N}\right]\frac{\mathbf{r}}{c^2r}+(-1)^N\lambda_N\frac{\mu\mathbf{v}\times\mathbf{S}^{(N)}}{m_N^2c^2}, \tag{2.8}$$

 0022-2488/87/030661-07$02.50

$$\mathbf{r}_E = \mathbf{r}\left(1 - \alpha\frac{GM}{c^2 r}\right) - \sum_{N=1}^{2} \lambda_N \frac{\mu \mathbf{v}\times\mathbf{S}^{(N)}}{m_N^2 c^2}, \tag{2.9}$$

where Eqs. (2.8) and (2.9) hold only for center-of-mass system. Our most general coordinates $\mathbf{r}_N$ are, thus, related to the EIH coordinates $\mathbf{r}_{EN}$ by the four arbitrary dimensionless parameters α, a_0, λ_1, and λ_2 as given by Eq. (2.8).

The notation of Refs. 3 and 4 is consistent with this paper. The coordinates $\mathbf{r}_{EN}$, $\mathbf{r}_{*N}$, and $\mathbf{r}_N$ of this paper correspond to $\mathbf{r}_{\mathrm{EIH},\alpha\beta,N}$, $\mathbf{r}_N$, and $\mathbf{r}_{N(\lambda N)}$ of Ref. 2, respectively. The coordinates $\mathbf{r}_{EN}$ and $\mathbf{r}_{*N}$ of this paper correspond to $\mathbf{r}_{NB}$ and $\mathbf{r}_N$ of Ref. 10 (Sec. I and II), respectively, if the electromagnetic part of Ref. 10 is omitted. Reference 1 is all in EIH coordinates.

We also will be using coordinates for the center of mass $\mathbf{r}_{\mathrm{ECM}}$, $\mathbf{r}_{*\mathrm{CM}}$, and $\mathbf{r}_{\mathrm{CM}}$, where

$$\mathbf{r}_{\mathrm{ECM}} = \sum_{N=1}^{2} \nu_N \mathbf{r}_{EN}, \quad \mathbf{r}_{*\mathrm{CM}} = \sum_{N=1}^{2} \nu_N \mathbf{r}_{*N}, \quad \mathbf{r}_{\mathrm{CM}} = \sum_{N=1}^{2} \nu_N \mathbf{r}_N, \tag{2.10}$$

and where in general ν_1 and ν_2 can take any values such that $\nu_1 + \nu_2 = 1$. However, in this paper we will always set $\nu_N = m_N/M$. Using Eqs. (2.2) and (2.10), we obtain

$$\mathbf{r}_{\mathrm{ECM}} = \mathbf{r}_{*\mathrm{CM}} - \alpha G(1 - a_0)\delta m \mathbf{r}_*/c^2 r_*, \tag{2.11}$$

where $\delta m \equiv m_1 - m_2$. Using Eqs. (2.6), (2.8), and (2.10), we obtain

$$\mathbf{r}_{*\mathrm{CM}} = \mathbf{r}_{\mathrm{CM}} + \sum_{N=1}^{2} (-1)^N \lambda_N \frac{\mu \mathbf{v}\times\mathbf{S}^{(N)}}{m_N M c^2}, \tag{2.12}$$

$$\mathbf{r}_{\mathrm{ECM}} = \mathbf{r}_{\mathrm{CM}} - \alpha G(1 - a_0)\delta m \frac{\mathbf{r}}{c^2 r} + \sum_{N=1}^{2} (-1)^N \lambda_N \frac{\mu \mathbf{v}\times\mathbf{S}^{(N)}}{m_N M c^2}, \tag{2.13}$$

where Eqs. (2.12) and (2.13) hold only for a center-of-mass system.

III. SPIN-ORBIT TERMS

Cho and Dass[11] have given the potential energy terms V_{S1} and V_{S2} for general relativity in EIH coordinates and in a frame of reference where the total momentum is *not* zero. Their results—derived from Schwinger's source theory[12]—are

$$V_{S1} = (Gm_2/c^2 r_E^3) \times [\tfrac{3}{2}\mathbf{S}^{(1)}\cdot(\mathbf{r}_E\times\mathbf{v}_{E1}) - 2\mathbf{S}^{(1)}\cdot(\mathbf{r}_E\times\mathbf{v}_{E2})], \tag{3.1}$$

$$V_{S2} = (Gm_1/c^2 r_E^3) \times [-\tfrac{3}{2}\mathbf{S}^{(2)}\cdot(\mathbf{r}_E\times\mathbf{v}_{E2}) + 2\mathbf{S}^{(2)}\cdot(\mathbf{r}_E\times\mathbf{v}_{E1})]. \tag{3.2}$$

Using the time derivative Eq. (2.10), we can put Eqs. (3.1) and (3.2) in the form

$$V_{SN} = \frac{G\mu}{c^2 r_E^3}\left(\frac{3}{2}\frac{m_1 m_2}{m_N^2} + 2\right)\mathbf{S}^{(N)}\cdot(\mathbf{r}_E\times\mathbf{v}_E) + (-1)^N \frac{G\mu}{c^2 r_E^3}\left(\frac{M}{2m_N}\right)\mathbf{S}^{(N)}\cdot(\mathbf{r}_E\times\mathbf{v}_{ECM}), \tag{3.3}$$

which is in agreement with the earlier general relativity results of Tulczyjew[13] [see his Eq. (3.10) and his Errata] who derived them using an "improved" EIH formalism.

The generalization of Eq. (3.3) to include PPN parameters γ and β is

$$V_{SN} = \frac{G\mu}{c^2 r_E^3}\left[\left(\gamma + \frac{1}{2}\right)\frac{m_1 m_2}{m_N^2} + \gamma + 1\right]\mathbf{S}^{(N)}\cdot(\mathbf{r}_E\times\mathbf{v}_E) + (-1)^N \frac{G\mu}{c^2 r_E^3}\left(\frac{M}{2m_N}\right)\mathbf{S}^{(N)}\cdot(\mathbf{r}_E\times\mathbf{v}_{ECM}). \tag{3.4}$$

The γ and β dependence of the first term in Eq. (3.4)—it turns out to be independent of β—has been given by us[2] previously and is consistent with the results of Börner, Ehlers, and Rudolph.[14] The V_{SN} term will contribute a term $-\partial V_{SN}/\partial \mathbf{v}_{\mathrm{ECM}}$ to the total momentum $\mathbf{P}_{\mathrm{ECM}}$, where

$$-\frac{\partial V_{SN}}{\partial \mathbf{v}_{\mathrm{ECM}}} = (-1)^N \frac{G\mu}{c^2 r_E^3}\left(\frac{M}{2m_N}\right)\mathbf{r}_E\times\mathbf{S}^{(N)}. \tag{3.5}$$

The second term in Eq. (3.4) must be γ (and β) independent so that its contribution to the total momentum will be γ (and β) independent. We shall now explain why the total momentum $\mathbf{P}_{\mathrm{ECM}}$ must be independent of γ (and β). Consider the n-body post-Newtonian Lagrangian with PPN parameters γ and β for (uncharged) point bodies (see Sec. IV C of Ref. 10). For this case[10] $\mathbf{P}_{\mathrm{ECM}}$ is independent of γ (and β). Because our two spinning bodies can be considered to be made up of n-point bodies, the total momentum $\mathbf{P}_{\mathrm{ECM}}$ for the two spinning bodies must also be independent of γ (and β).

IV. CENTER OF INERTIA

Let us start in the $\mathbf{r}_{*N}$ coordinate system where $\mathbf{r}_{*\mathrm{CI}}$, $\mathbf{v}_{*\mathrm{CI}}$, and $\mathbf{a}_{*\mathrm{CI}}$ are the position, velocity, and acceleration, respectively, of the center of inertia, and where $\mathscr{E}_*$, $\mathbf{P}_{*\mathrm{CM}}$, and $\mathbf{P}_{*N}$ are the total conserved energy, total conserved (canonical) momentum, and (canonical) momentum of body N, respectively. The center of inertia must satisfy the equation[10,15,16]

$$\frac{d}{dt}((\mathscr{E}_*/c^2)\mathbf{r}_{*\mathrm{CI}}) = \mathbf{P}_{*\mathrm{CM}} = \mathbf{P}_{*1} + \mathbf{P}_{*2}, \tag{4.1}$$

from which it follows that $(\mathscr{E}_*/c^2)\mathbf{v}_{*\mathrm{CI}} = \mathbf{P}_{*\mathrm{CM}}$ and $\mathbf{a}_{*\mathrm{CI}} = 0$.

In order to satisfy Eq. (4.1), we set

$$\mathscr{E}_*\mathbf{r}_{*\mathrm{CI}} = \sum_{N=1}^{2}\left[\mathscr{E}_{*N}\mathbf{r}_{*N} + (-1)^N \frac{\mu}{2m_N}\mathbf{S}^{(N)}\times\mathbf{v}_*\right], \tag{4.2}$$

where

$$\mathscr{E}_{*N} = m_N c^2 + \tfrac{1}{2}m_N v_{*N}^2 - \left[\frac{1}{2} - (-1)^N \alpha(1 - a_0)\frac{\delta m}{\mu}\right]\frac{Gm_1 m_2}{r_*}. \tag{4.3}$$

We must also have

$$\mathscr{E}_* = \mathscr{E}_{*1} + \mathscr{E}_{*2} = Mc^2 + \sum_{N=1}^{N} \frac{1}{2} m_N v_{*N}^2 - \frac{Gm_1 m_2}{r_*}, \tag{4.4}$$

which is in agreement with Eq. (4.3). The terms $m_N c^2$ in Eq. (4.3) and Mc^2 in Eq. (4.4) must include rotational kinetic energy in order that these equations be accurate to the Newtonian approximation (i.e., order c^0). Thus we have

$$m_N c^2 = \left(m_{0N} + \tfrac{1}{2}\frac{I^{(N)}\omega^{(N)2}}{c^2}\right)c^2, \tag{4.5}$$

$$Mc^2 = (m_1 + m_2)c^2, \tag{4.6}$$

where m_{0N}, $I^{(N)}$, and $\omega^{(N)}$ are the nonrotating rest mass, moment of inertia, and angular velocity, respectively, of body N. We deduced Eqs. (4.2) and (4.3) by using the results for nonrotating bodies given in Secs. IV B and IV C of Ref. 10 and then adding spin terms to Eq. (4.2) that are consistent with Eq. (3.5). The spin terms in Eq. (4.2) are consistent with Eq. (3.5) because

$$\frac{d}{dt}\left[(-1)^N\frac{\mu}{2m_N c^2}\mathbf{S}^{(N)}\times\mathbf{v}_*\right] = (-1)^N\frac{G\mu M}{2r_*^3 m_N c^2}\mathbf{r}_*\times\mathbf{S}^{(N)}. \tag{4.7}$$

In evaluating the left-hand side of Eq. (4.7) which is a post-Newtonian term (i.e., of order c^{-2}), we have used $d\mathbf{S}^{(N)}/dt = 0$ and $\mathbf{a}_* = -GM\mathbf{r}_*/r_*^3$, which are correct to first order. The spin-orbit terms and Eq. (3.5) in Sec. III are of order c^{-2} and, thus, to this order we can replace the $\mathbf{r}_{EN}$ coordinates with $\mathbf{r}_{*N}$ coordinates. It is an interesting fact[10] that $\mathbf{P}_{*\mathrm{CM}}$ expressed in $\mathbf{r}_{*N}$ coordinates is explicitly independent of the parameters γ and β (and α if $a_0 = 1$). It should be noted that the post-Newtonian potential energy terms for the spin–spin interaction[1–4] $V_{S1,S2}$ and the Nordtvedt effect[2–4,9] as well as the quadrupole moment interactions[1,3,4] (small Newtonian terms and thus treated as if they were post-Newtonian terms from an order of magnitude point of view) V_{Q1} and V_{Q2} are velocity independent and hence will not contribute to $\mathbf{P}_{*\mathrm{CM}}$. We conclude that Eqs. (4.1)–(4.3) are still valid when these terms are included in the Lagrangian.

In the center-of-mass coordinate system $\mathbf{P}_{*\mathrm{CM}} = 0$ and, thus, $\mathbf{v}_{*\mathrm{CI}} = 0$ and $\mathbf{r}_{*\mathrm{CI}}$ is a constant. We shall now set $\mathbf{r}_{*\mathrm{CI}} = 0$ and obtain from Eq. (4.2)

$$\sum_{N=1}^{2}\left[\mathscr{E}_{*N}\mathbf{r}_{*N} + (-1)^N\frac{\mu}{2m_N}\mathbf{S}^{(N)}\times\mathbf{v}_*\right] = 0. \tag{4.8}$$

From Eq. (4.8), it follows (to post-Newtonian order) that

$$\mathbf{r}_{*1} = -\frac{\mathscr{E}_{*2}}{\mathscr{E}_{*1}}\mathbf{r}_{*2} - \frac{1}{m_1c^2}\sum_{N=1}^{2}(-1)^N\frac{\mu}{2m_N}S^{(N)}\times\mathbf{v}_*, \tag{4.9}$$

$$\mathbf{r}_{*2} = -\frac{\mathscr{E}_{*1}}{\mathscr{E}_{*2}}\mathbf{r}_{*1} - \frac{1}{m_2c^2}\sum_{N=1}^{2}(-1)^N\frac{\mu}{2m_N}\mathbf{S}^{(N)}\times\mathbf{v}_*, \tag{4.10}$$

and thus

$$\mathbf{r}_* = -(-1)^N\left[\frac{\mathscr{E}_*\mathscr{E}_{*N}}{\mathscr{E}_{*1}\mathscr{E}_{*2}}\mathbf{r}_{*N} + \frac{m_N}{m_1m_2c^2}\times\sum_{N=1}^{2}(-1)^N\frac{\mu}{2m_N}\mathbf{S}^{(N)}\times\mathbf{v}_*\right], \tag{4.11}$$

which can be inverted to give us

$$\mathbf{r}_{*N} = -(-1)^N\frac{\mathscr{E}_{*1}\mathscr{E}_{*2}}{\mathscr{E}_*\mathscr{E}_{*N}}\mathbf{r} - \frac{1}{Mc^2}\times\sum_{N=1}^{2}(-1)^N\frac{\mu}{2m_N}\mathbf{S}^{(N)}\times\mathbf{v}_*. \tag{4.12}$$

Because we are in center-of-mass system, we can use $\mathbf{v}_{*N} = -(-1)^N\mu\mathbf{v}_*/m_N$ in Eqs. (4.4) and (4.3) to obtain, respectively,

$$\mathscr{E}_* = Mc^2 + \tfrac{1}{2}\mu v_*^2 - GM\mu/r_*, \tag{4.13}$$

$$\mathscr{E}_{*N} = m_Nc^2 + \tfrac{1}{2}\mu^2v_*^2/m_N - [\tfrac{1}{2} - (-1)^N\alpha(1-a_0)\delta m/\mu]Gm_1m_2/r_*. \tag{4.14}$$

Inverting Eq. (2.10) we obtain

$$\mathbf{r}_{*N} = -(-1)^N(m_1m_2/Mm_N)\mathbf{r}_* + \mathbf{r}_{*\mathrm{CM}}. \tag{4.15}$$

Using Eqs. (4.13) and (4.14) in (4.12) and comparing the result with Eq. (4.15), we obtain (to the post-Newtonian approximation)

$$\mathbf{r}_{*\mathrm{CM}} = \frac{\mu\delta m}{2M^2c^2}\left[v_*^2 - \frac{GM}{r_*} + \frac{2\alpha(1-a_0)GM^2}{\mu r_*}\right]\mathbf{r}_* - \frac{1}{Mc^2}\sum_{N=1}^{2}(-1)^N\frac{\mu}{2m_N}\mathbf{S}^{(N)}\times\mathbf{v}_*. \tag{4.16}$$

The masses M and m_N in the first terms (terms of order c^2) in Eqs. (4.13) and (4.14), respectively, are given by Eqs. (4.5) and (4.6) and the same must be true for the masses in the first term of Eq. (4.15) (i.e., for the masses in $\nu_N = m_N/M$) if Eq. (4.16) is to be correct.

The EIH coordinate versions of Eqs. (4.15) and (4.16) are given by setting $\alpha = 0$. We obtain

$$\mathbf{r}_{EN} = -(-1)^N(m_1m_2/Mm_N)\mathbf{r}_E + \mathbf{r}_{\mathrm{ECM}}, \tag{4.17}$$

$$\mathbf{r}_{ECM} = \frac{\mu\delta m}{2M^2c^2}\left[v_E^2 - \frac{GM}{r_E}\right]\mathbf{r}_E - \frac{1}{Mc^2}\sum_{N=1}^{2}(-1)^N\frac{\mu}{2m_N}\mathbf{S}^{(N)}\times\mathbf{v}_E. \tag{4.18}$$

The above with $\mathbf{S}^{(N)} = 0$ (the result for nonrotating bodies) has been given by Wagoner and Will.[7] Using Eqs. (2.11) and (4.18), we can regain Eq. (4.16) correct to the post-Newtonian approximation.

Let us next consider our most general coordinate system, the $\mathbf{r}_N$ coordinate system. Inverting Eq. (2.10) and combining Eqs. (2.12) and (4.16), we obtain (to the post-Newtonian approximation)

$$\mathbf{r}_N = -(-1)^N\frac{m_1m_2}{Mm_N}\mathbf{r} + \mathbf{r}_{\mathrm{CM}}, \tag{4.19}$$

$$\mathbf{r}_{\mathrm{CM}} = \frac{\mu\delta m}{2M^2c^2}\left[v^2 - \frac{GM}{r} + \frac{2\alpha(1-a_0)GM^2}{\mu r}\right]\mathbf{r} - \frac{1}{Mc^2}\sum_{N=1}^{N}(-1)^N\left(\frac{\mu}{2m_N}\right)(1-2\lambda_N)\mathbf{S}^{(N)}\times\mathbf{v}, \tag{4.20}$$

where the above are valid in the center-of-mass coordinate system where the total momentum is zero. Tulczyjew[13] [see his Eq. (3.14) and his Errata] has given an EIH coordinate

system version (i.e., α and λ_N are zero) of our Eq. (4.20). If we start with a Lagrangian in the $\mathbf{r}_{*N}$ coordinate system and make the coordinate transformation of Eq. (2.4) to the $\mathbf{r}_N$ coordinate system, we will obtain an acceleration-dependent Lagrangian.[4] To obtain this acceleration-dependent Lagrangian start with the Lagrangian in the $\mathbf{r}_{*N}$ coordinate system, replace $\mathbf{r}_{*N}$ by $\mathbf{r}_N$ and $\mathbf{v}_{*N}$ by $\mathbf{v}_N$, and then add the terms $-V_{\lambda 1}$ and $-V_{\lambda 2}$, where

$$V_{\lambda N} = -\frac{\lambda_N}{m_N c^2}\mathbf{S}^{(N)}\cdot\left\{\left[\mathbf{a}_N - (-1)^N\frac{Gm_1m_2}{r^3m_N}\mathbf{r}\right]\times m_N\mathbf{v}_N\right\}. \tag{4.21}$$

The center of inertia must then satisfy the equation[16]

$$\frac{d}{dt}\left(\frac{\mathscr{E}}{c^2}\mathbf{r}_{\mathrm{CI}}\right) = \mathbf{\Pi}_1 + \mathbf{\Pi}_2, \tag{4.22}$$

where

$$\mathbf{\Pi}_N = \frac{\partial\mathscr{L}}{\partial\mathbf{v}_N} - \frac{d}{dt}\left(\frac{\partial\mathscr{L}}{\partial\mathbf{a}_N}\right), \tag{4.23}$$

and it follows from Eq. (4.22) that $(\mathscr{E}/c^2)\mathbf{v}_{\mathrm{CI}} = \mathbf{\Pi}_1 + \mathbf{\Pi}_2$ and $\mathbf{v}_{\mathrm{CI}}$ is a constant. In order to satisfy Eq. (4.22), we set

$$\mathscr{E}\mathbf{r}_{\mathrm{CI}} = \sum_{N=1}^{2}\left[\mathscr{E}_N\mathbf{r}_N + (-1)^N\times\left(\frac{\mu}{2m_N}\right)(1-2\lambda_N)\mathbf{S}^{(N)}\times\mathbf{v}\right], \tag{4.24}$$

where $\mathscr{E} = \mathscr{E}_1 + \mathscr{E}_2$ and

$$\mathscr{E}_N = m_Nc^2 + \tfrac{1}{2}m_Nv_N^2 - [\tfrac{1}{2} - (-1)^N\alpha(1-a_0)\delta m/\mu]Gm_1m_2/r. \tag{4.25}$$

It should be noted that we could also add the constant terms $K_1\mathbf{S}^{(1)}\times\mathbf{v}_{\mathrm{CM}}$ and $K_2\mathbf{S}^{(2)}\times\mathbf{v}_{\mathrm{CM}}$, where K_1 and K_2 are arbitrary dimensionless constants, to the right-hand side of Eq. (4.24) and also to the right-hand side of Eq. (4.2). This would not alter our other results because we use Eqs. (4.2) and (4.24) in a frame of reference where $\mathbf{v}_{\mathrm{CM}}$ is zero. We can also obtain Eqs. (4.24) and (4.25) by noting that these results are exactly what is needed to obtain Eq. (4.20). This can easily be seen by comparing Eqs. (4.2), (4.3), and (4.16) with Eqs. (4.24), (4.25), and (4.20), respectively.

The spin supplementary condition[2,17] of Price[18] and of Newton and Wigner[19] corresponds to setting λ_N equal to zero and has the advantage that the Langrangian will have no acceleration-dependent spin terms.[4] The spin supplementary condition[2,17] of Corinaldesi–Papapetrou[20] corresponds to setting $\lambda_N = \frac{1}{2}$ and has the advantage that Eqs. (4.20) and (4.24) will have no spin-dependent terms. The spin supplementary condition[2,17] of Pirani[21] corresponds to setting $\lambda_N = -\frac{1}{2}$ and has the advantage that the equations of motion take on a particularly simple form. The equations of motion can be written as

$$\begin{aligned}\mathbf{a}_N - (-1)^N Gm_P\mathbf{r}/r^3 &= \text{SPIN-INDEPENDENT TERMS}\\ &\quad + \text{SPIN TERMS} + \text{SPIN–SPIN TERMS}\\ &\quad + \text{QUADRUPOLE MOMENT TERMS},\end{aligned} \tag{4.26}$$

where the SPIN TERMS, which depend on λ_1 and λ_2, are given by

$$\begin{aligned}\text{SPIN TERMS} = \frac{Gm_P}{m_Nc^2r^3}&\left[(2\gamma+1)\mathbf{S}^{(N)}\times\mathbf{v}_N - \left(2\gamma+\frac{3}{2}+\lambda_N\right)\mathbf{S}^{(N)}\times\mathbf{v}_P + \left(3\gamma+\frac{3}{2}-3\lambda_N\right)\frac{\mathbf{S}^{(N)}\cdot(\mathbf{r}\times\mathbf{v}_N)}{r^2}\mathbf{r}\right.\\ &\left. - (3\gamma+3)\frac{\mathbf{S}^{(N)}\cdot(\mathbf{r}\times\mathbf{v}_P)}{r^2}\mathbf{r} + (-1)^N\left(3\gamma+\frac{3}{2}+3\lambda_N\right)\frac{\mathbf{v}\cdot\mathbf{r}}{r^2}\mathbf{S}^{(N)}\times\mathbf{r}\right]\\ +\frac{G}{c^2r^3}&\left[(2\gamma+2)\mathbf{S}^{(P)}\times\mathbf{v}_N + \left(\lambda_P - 2\gamma - \frac{3}{2}\right)\mathbf{S}^{(P)}\times\mathbf{v}_P + \left(3\lambda_P - 3\gamma - \frac{3}{2}\right)\frac{\mathbf{S}^{(P)}\cdot(\mathbf{r}\times\mathbf{v}_P)}{r^2}\mathbf{r}\right.\\ &\left. + (3\gamma+3)\frac{\mathbf{S}^{(P)}\cdot(\mathbf{r}\times\mathbf{v}_N)}{r^2}\mathbf{r} + (-1)^N(3\gamma+3)\frac{\mathbf{v}\cdot\mathbf{r}}{r^2}\mathbf{S}^{(P)}\times\mathbf{r}\right],\end{aligned} \tag{4.27}$$

and where $P \equiv 3 - N$. If one now sets $\lambda_1 = \lambda_2 = -\frac{1}{2}$, the above can be put in the form

$$\begin{aligned}\text{SPIN TERMS} = &\frac{Gm_P}{m_Nc^2r^3}(-1)^P\left[(2\gamma+1)\mathbf{S}^{(N)}\times\mathbf{v} + (3\gamma+3)\times\frac{\mathbf{S}^{(N)}\cdot(\mathbf{r}\times\mathbf{v})}{r^2}\mathbf{r} - 3\gamma\frac{\mathbf{v}\cdot\mathbf{r}}{r^2}\mathbf{S}^{(N)}\times\mathbf{r}\right]\\ &+\frac{G}{c^2r^3}(-1)^P\left[(2\gamma+2)\mathbf{S}^{(P)}\times\mathbf{v} + (3\gamma+3)\times\frac{\mathbf{S}^{(P)}\cdot(\mathbf{r}\times\mathbf{v})}{r^2}\mathbf{r} - (3\gamma+3)\frac{\mathbf{v}\cdot\mathbf{r}}{r^2}\mathbf{S}^{(P)}\times\mathbf{r}\right].\end{aligned} \tag{4.28}$$

It should be noted that Eqs. (4.26)–(4.28) are correct for a non-center-of-mass coordinate system and that Eq. (4.28) takes on the identical form in a center-of-mass coordinate system where $\mathbf{v}_{\mathrm{CM}}$ is zero (to first order). If one sets $\gamma = 1$ and uses the vector identity

$$\begin{aligned}(\mathbf{r}\times\mathbf{v})\mathbf{S}^{(N)}\cdot\mathbf{r} &\equiv [\mathbf{S}^{(N)}\cdot(\mathbf{r}\times\mathbf{v})]\mathbf{r} + (\mathbf{S}^{(N)}\times\mathbf{v})r^2 - (\mathbf{v}\cdot\mathbf{r})(\mathbf{S}^{(N)}\times\mathbf{r}),\end{aligned} \tag{4.29}$$

and the center-of-mass system relation $\mathbf{v}_N = -(-1)^N\times\mu\mathbf{v}/m_N$ in Eq. (4.28) one can put this equation into the form of the center-of-mass general relativity result of D'Eath[22] [see his Eq. (6.7)].

V. CONCLUSIONS

We completed the solution to the post-Newtonian gravitational two-body problem with spin and PPN parameters γ and β by giving [see Eqs. (4.19) and (4.20)] the positions $\mathbf{r}_1$ and $\mathbf{r}_2$ and the center of mass $\mathbf{r}_{CM}$ of the two bodies as a function of the relative position $\mathbf{r}$, relative velocity $\mathbf{v}$, and spin angular momenta $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ of the two bodies. Thus, if we have a solution $\mathbf{r}=\mathbf{r}(t)$, correct to the post-Newtonian approximation, we also have solutions $\mathbf{r}_1=\mathbf{r}_1(t)$ and $\mathbf{r}_2=\mathbf{r}_2(t)$, correct to the post-Newtonian approximation. For coordinate systems [see Eq. (2.8)] where $\alpha\neq 0$ and $a_0\neq 1$, the position $\mathbf{r}_{CM}$ has a nonzero term $2\alpha(1-a_0)\times GM^2/\mu r$ inside the square bracket of Eq. (4.20) that is due to the potential energy term $-Gm_1m_2/r$ not being split[10] equally between $\mathscr{E}_1$ and $\mathscr{E}_2$ of Eq. (4.25). Clearly Eqs. (4.19), (4.20), (4.24), and (4.25) remain valid if the potential energy terms for the spin–spin interaction[1–4] $V_{S1,S2}$ and the Nordtvedt effect[2–4,9] as well as the quadrupole moment interactions[1,3,4] V_{Q1} and V_{Q2} (which are all velocity and acceleration independent and hence will not contribute to the total momentum $\mathbf{\Pi}_1+\mathbf{\Pi}_2$) are included in the Lagrangian.

In the Appendix, we gave a quantum field theory derivation of the spin-orbit potential energy terms V_{S1} and V_{S2} for general relativity in a frame of reference where the total momentum was *not* zero and our results were in agreement with those derived by Cho and Dass[11] from Schwinger's source theory[12] and those derived by Tulczyjew[13] using an "improved" EIH formalism. In Sec. III, we included PPN parameters γ and β in the spin-orbit potential energy terms and concluded that the second term in Eq. (3.4) had to be independent of γ and β.

APPENDIX: FIELD THEORY DERIVATION OF V_{S1} AND V_{S2}

In this Appendix, we shall give a quantum field theory derivation of the spin-orbit potential energy terms V_{S1} and V_{S2} of Eqs. (3.1) and (3.2), respectively, for general relativity in a frame of reference where the total momentum is not zero. Let us consider the one-graviton exchange interaction between a spin-$\frac{1}{2}$ particle of mass m_1 and a spin-0 particle of mass m_2. Let the initial and final propagation four-vectors for the spin-$\frac{1}{2}$ particle be p_μ and p'_μ, respectively, and those for the spin-0 particle be q_μ and q'_μ, respectively. We also have

$$\mathbf{P}_1=\hbar\mathbf{p},\quad E_1=c\hbar p_0,\quad \lambda_1=m_1c/\hbar,\qquad p^2=\mathbf{p}^2-p_0^2=-\lambda_1^2, \tag{A1}$$

$$\mathbf{P}_2=\hbar\mathbf{q},\quad E_2=c\hbar q_0,\quad \lambda_2=m_2c/\hbar,\qquad q^2=\mathbf{q}^2-q_0^2=-\lambda_2^2, \tag{A2}$$

where $\mathbf{P}_N$ and E_N are the momentum and energy, respectively, of particle N and $\hbar$ is Planck's constant divided by 2π. Note that the λ_1 and λ_2 of this Appendix are *not* the same quantities as the λ_1 and λ_2 used in the rest of this paper.

The graviton coupling constant κ is related to Newton's constant of gravitation G and the speed of light c by the relation

$$\kappa^2=16\pi G/c^4. \tag{A3}$$

Let ψ be a spin-$\frac{1}{2}$ field with mass m_1 and U_0 be a spin-0 field with mass m_2. The interaction terms (using ordered products[23]) with the graviton field $h_{\mu\nu}$ are[24,25] (to order κ)

$$:L_{\rm int}:=-\frac{1}{8}\kappa c\hbar:\left[\bar\psi\gamma_\mu\frac{\partial\psi}{\partial x_\nu}+\bar\psi\gamma_\nu\frac{\partial\psi}{\partial x_\mu}-\frac{\partial\bar\psi}{\partial x_\nu}\gamma_\mu\psi-\frac{\partial\bar\psi}{\partial x_\mu}\gamma_\nu\psi\right]h_{\mu\nu}:$$
$$-\frac{1}{2}\kappa:\left[\frac{\partial U_0}{\partial x_\mu}\frac{\partial U_0}{\partial x_\nu}-\frac{\delta_{\mu\nu}}{2}\frac{\partial U_0}{\partial x_\rho}\frac{\partial U_0}{\partial x_\rho}-\frac{\delta_{\mu\nu}}{2}\lambda_2^2U_0^2\right]h_{\mu\nu}:, \tag{A4}$$

and the contractions are[24,25]

$$\dot{h}_{\mu\nu}(x)\dot{h}_{\lambda\rho}(x')=-ic\hbar(\delta_{\mu\lambda}\delta_{\nu\rho}+\delta_{\mu\rho}\delta_{\nu\lambda}-\delta_{\mu\nu}\delta_{\lambda\rho})D_F(x-x'), \tag{A5}$$

where

$$D_F(x-x')=\lim_{\epsilon\to+0}\frac{1}{(2\pi)^4}\int dk\,e^{ik(x-x')}\frac{1}{k^2-i\epsilon}. \tag{A6}$$

We also have[23]

$$S_2=\frac{-1}{2c^2\hbar^2}\int dx\int dx'\,T[:H_{\rm int}(x):\,:H_{\rm int}(x'):], \tag{A7}$$

where $:H_{\rm int}:$ may be replaced[23,26] by $-:L_{\rm int}:$.

Using Eqs. (A4) in (A7), we obtain

$$S_2=\frac{-\kappa^2}{16c\hbar}\int dx\int dx':\left[\bar\psi(x)\gamma_\mu\frac{\partial\psi(x)}{\partial x_\nu}+\bar\psi(x)\gamma_\nu\frac{\partial\psi(x)}{\partial x_\mu}-\frac{\partial\bar\psi(x)}{\partial x_\nu}\gamma_\mu\psi(x)-\frac{\partial\bar\psi(x)}{\partial x_\mu}\gamma_\nu\psi(x)\right]\dot{h}_{\mu\nu}(x)$$
$$\times\left[\frac{\partial U_0(x')}{\partial x'_\alpha}\frac{\partial U_0(x')}{\partial x'_\beta}-\frac{\delta_{\alpha\beta}}{2}\frac{\partial U_0(x')}{\partial x'_\lambda}\frac{\partial U_0(x')}{\partial x'_\lambda}-\frac{\delta_{\alpha\beta}}{2}\lambda_2^2U_0(x')U_0(x')\right]\dot{h}_{\alpha\beta}(x'):. \tag{A8}$$

Using Eqs. (A5) and (A6) along with

$$\psi(x)=(1/V)^{1/2}\psi^+(\mathbf{p})e^{ipx}, \tag{A9}$$

$$\bar\psi(x)=(1/V)^{1/2}\bar\psi^-(\mathbf{p}')e^{-ip'x}, \tag{A10}$$

$$U_0(x')=(c\hbar/2q_0V)^{1/2}a(\mathbf{q})e^{iqx'}+(c\hbar/2q'_0V)^{1/2}a^*(\mathbf{q}')e^{-iq'x'}, \tag{A11}$$

in Eq. (A8), we find

$$S_2 = \frac{ic\hbar\kappa^2(2\pi)^4}{4V^2(q_0'q_0)^{1/2}}\delta(p-p'+q-q') \times \frac{1}{(p'-p)^2}:[-\lambda_1\lambda_2^2\bar{\psi}^-(\mathbf{p}')\psi^+(\mathbf{p}) + \tfrac{1}{4}(p+p')(q+q')\bar{\psi}^-(\mathbf{p}') \times i(q+q')\gamma\psi^+(\mathbf{p})]a^*(\mathbf{q}')a(\mathbf{q}):, \tag{A12}$$

where V is a volume factor, and only the appropriate terms in the Fourier expansion have been included on the right-hand side of Eqs. (A9), (A10), and (A11).

The quantity $V(\mathbf{k})$, correct to first order in G, is defined in terms of the second-order S matrix as[23]

$$S_2 = (-i/c\hbar V^2)(2\pi)^4\delta(p+q-p'-q') \times \psi_L^{*-}(\mathbf{p}')a^*(\mathbf{q}')V(\mathbf{k})a(\mathbf{q})\psi_L^+(\mathbf{p}), \tag{A13}$$

where $\bar{\psi}^-(\mathbf{p}') = \psi^{*-}(\mathbf{p}')\gamma_4$ and [23]

$$\psi^{*-}(\mathbf{p}') = \left(\frac{\lambda_1+p_0'}{2p_0'}\right)^{1/2}(\psi_L^{*-}(\mathbf{p}') \quad \psi_S^{*-}(\mathbf{p}')), \tag{A14}$$

$$\psi^+(\mathbf{p}) = \left(\frac{\lambda_1+p_0}{2p_0}\right)^{1/2}\begin{pmatrix}\psi_L^+(\mathbf{p})\\ \psi_S^+(\mathbf{p})\end{pmatrix}, \tag{A15}$$

$$\psi_S^{*-}(\mathbf{p}') = \psi_L^{*-}(\mathbf{p}')\frac{(\mathbf{p}'\cdot\boldsymbol{\sigma}^{(1)})}{(\lambda_1+p_0')}, \tag{A16}$$

$$\psi_S^+(\mathbf{p}) = \frac{(\mathbf{p}\cdot\boldsymbol{\sigma}^{(1)})}{(\lambda_1+p_0)}\psi_L^+(\mathbf{p}), \tag{A17}$$

and where $k \equiv p'-p = q-q'$ so that

$$\mathbf{k} = \mathbf{p}'-\mathbf{p} = \mathbf{q}-\mathbf{q}', \quad k_0 = p_0'-p_0 = q_0-q_0'. \tag{A18}$$

Let us now set

$$V(\mathbf{k}) = V_1(\mathbf{k}) + V_{S1}(\mathbf{k}), \tag{A19}$$

where $V_1(\mathbf{k})$ is spin independent and $V_{S1}(\mathbf{k})$ is spin dependent (i.e., depends on $\boldsymbol{\sigma}^{(1)}$). Using Eqs. (A12)–(A19), we obtain

$$V_1(\mathbf{k}) = \frac{-c^2\hbar^2\kappa^2}{4(q_0q_0')^{1/2}}\left(\frac{\lambda_1+p_0'}{2p_0'}\right)^{1/2}\left(\frac{\lambda_1+p_0}{2p_0}\right)^{1/2}\frac{1}{\mathbf{k}^2-k_0^2}\left[-\lambda_1\lambda_2^2\left(1-\frac{\mathbf{p}'\cdot\mathbf{p}}{(\lambda_1+p_0')(\lambda_1+p_0)}\right) - \frac{1}{4}(p+p')(q+q')(q_0+q_0')\left(1+\frac{\mathbf{p}'\cdot\mathbf{p}}{(\lambda_1+p_0')(\lambda_1+p_0)}\right) + \frac{1}{4}(p+p')(q+q')\left(\frac{\mathbf{p}\cdot(\mathbf{q}+\mathbf{q}')}{\lambda_1+p_0}+\frac{\mathbf{p}'\cdot(\mathbf{q}+\mathbf{q}')}{\lambda_1+p_0'}\right)\right], \tag{A20}$$

$$V_{S1}(\mathbf{k}) = \frac{-c^2\hbar^2\kappa^2}{4(q_0q_0')^{1/2}}\left(\frac{\lambda_1+p_0'}{2p_0'}\right)^{1/2}\left(\frac{\lambda_1+p_0}{2p_0}\right)^{1/2}\frac{1}{\mathbf{k}^2-k_0^2}\left[\left(\lambda_1\lambda_2^2-\frac{1}{4}(p+p')(q+q')(q_0+q_0')\right) \times\frac{i\boldsymbol{\sigma}^{(1)}\cdot[\mathbf{k}\times(\mathbf{p}+\mathbf{p}')]}{2(\lambda_1+p_0')(\lambda_1+p_0)}+\frac{1}{4}(p+p')(q+q') \times\left(\frac{i\boldsymbol{\sigma}^{(1)}\cdot[\mathbf{k}\times(\mathbf{q}+\mathbf{q}')]}{2(\lambda_1+p_0)}+\frac{i\boldsymbol{\sigma}^{(1)}\cdot[\mathbf{k}\times(\mathbf{q}+\mathbf{q}')]}{2(\lambda_1+p_0')}-\frac{ik_0\boldsymbol{\sigma}^{(1)}\cdot[(\mathbf{p}+\mathbf{p}')\times(\mathbf{q}+\mathbf{q}')]}{2(\lambda_1+p_0)(\lambda_1+p_0')}\right)\right]. \tag{A21}$$

Using Eq. (A18), we can express k_0 as[10] $k_0 = \mathbf{k}\cdot(\mathbf{p}'+\mathbf{p})/(p_0'+p_0)$ or as $k_0 = \mathbf{k}\cdot(\mathbf{q}+\mathbf{q}')/(q_0+q_0')$. Similar to what was done in Ref. 10, we shall put the k_0^2 that is in Eq. (A20) and (A21) in the form

$$k_0^2 = [1+2\alpha(a_{12}+a_{21})]\left(\frac{\mathbf{k}\cdot(\mathbf{p}'+\mathbf{p})}{p_0'+p_0}\right)\left(\frac{\mathbf{k}\cdot(\mathbf{q}+\mathbf{q}')}{q_0+q_0'}\right) - 2\alpha\left[a_{12}\left(\frac{\mathbf{k}\cdot(\mathbf{p}'+\mathbf{p})}{p_0'+p_0}\right)^2+a_{21}\left(\frac{\mathbf{k}\cdot(\mathbf{q}+\mathbf{q}')}{q_0+q_0'}\right)^2\right], \tag{A22}$$

where

$$a_{12} = [(1-a_0)m_1+a_0m_2]/m_2, \tag{A23}$$

$$a_{21} = [(1-a_0)m_2+a_0m_1]/m_1. \tag{A24}$$

We also have a k_0 in Eq. (A21), a situation that did not arise in Ref. 10. Because there is no simple way to obtain a square root of the k_0^2 of Eq. (A22), we suggest that this k_0 be put in the form

$$k_0 = \bar{a}_0\left(\frac{\mathbf{k}\cdot(\mathbf{p}'+\mathbf{p})}{p_0'+p_0}\right)+(1-\bar{a}_0)\left(\frac{\mathbf{k}\cdot(\mathbf{q}+\mathbf{q}')}{q_0+q_0'}\right), \tag{A25}$$

where $\bar{a}_0$ is another arbitrary dimensionless constant. It should be noted that Eqs. (A20)–(A22) and (A25) have been put into a form such that they remain the same if $\mathbf{p}\rightleftarrows\mathbf{p}'$ and $p_0\rightleftarrows p_0'$ or if $\mathbf{q}\rightleftarrows\mathbf{q}'$ and $q_0\rightleftarrows q_0'$ while $\mathbf{k}$ is *not* altered (i.e., we do *not* change $\mathbf{k}$ into $-\mathbf{k}$). The quantity $V(\mathbf{k})$ can also be defined in terms of the potential energy $V(\mathbf{r})$, correct to first order in G, as

$$V(\mathbf{k}) = \int d\mathbf{r}\, e^{-i(\mathbf{p}'\cdot\mathbf{r}_1+\mathbf{q}'\cdot\mathbf{r}_2)}V(\mathbf{r})e^{i(\mathbf{p}\cdot\mathbf{r}_1+\mathbf{q}\cdot\mathbf{r}_2)}. \tag{A26}$$

The potential energy $V(\mathbf{r})$ is a Hermitian operator and is also momentum dependent [i.e., $V(\mathbf{r}) \equiv V(\mathbf{r},\mathbf{p}_{op},\mathbf{q}_{op})$ where $\mathbf{p}_{op}$ and $\mathbf{q}_{op}$ are operators].

If we are only interested in the classical results, as we are in this paper, the ordering of the factors in $V(\mathbf{r},\mathbf{p}_{op},\mathbf{q}_{op})$

makes no difference (i.e., we can neglect delta function terms). To obtain the classical results corresponding to Eqs. (A20)–(A22) and (A25) set $\mathbf{p}' = \mathbf{p}, p_0' = p_0, \mathbf{q}' = \mathbf{q}$, and $q_0' = q_0$ while now considering $\mathbf{k}$ to be an independent variable (i.e., do *not* set $\mathbf{k}$ equal to zero). The classical results are

$$V_1(\mathbf{k}) = -\frac{c^2\hbar^2\kappa^2}{4p_0q_0}\frac{1}{\mathbf{k}^2 - k_0^2}\left[2(pq)^2 - \lambda_1^2\lambda_2^2\right], \quad \text{(A27)}$$

$$\begin{aligned} V_{S1}(\mathbf{k}) = &-\frac{c^2\hbar^2 q_0\kappa^2}{8(\lambda_1 + p_0)}\frac{1}{\mathbf{k}^2 - k_0^2} \\ &\times\left\{\left[2 - \frac{2\mathbf{p}\cdot\mathbf{q}}{p_0q_0} + \lambda_1\left(\frac{1}{p_0} - \frac{\mathbf{q}^2}{p_0q_0^2}\right)\right]i\boldsymbol{\sigma}^{(1)}\cdot(\mathbf{k}\times\mathbf{p})\right. \\ &+\left[-\frac{2p_0}{q_0} + \frac{2\mathbf{p}\cdot\mathbf{q}}{q_0^2} + \lambda_1\left(-\frac{2}{q_0} + \frac{2\mathbf{p}\cdot\mathbf{q}}{p_0q_0^2}\right)\right]i\boldsymbol{\sigma}^{(1)} \\ &\left.\cdot(\mathbf{k}\times\mathbf{q}) + k_0\left(\frac{2}{q_0} - \frac{2\mathbf{p}\cdot\mathbf{q}}{p_0q_0^2}\right)i\boldsymbol{\sigma}^{(1)}\cdot(\mathbf{p}\times\mathbf{q})\right\}, \end{aligned} \quad \text{(A28)}$$

where

$$\begin{aligned} k_0^2 = &[1 + 2\alpha(a_{12} + a_{21})]\frac{(\mathbf{k}\cdot\mathbf{p})(\mathbf{k}\cdot\mathbf{q})}{p_0q_0} \\ &- 2\alpha\left[a_{12}\left(\frac{\mathbf{k}\cdot\mathbf{p}}{p_0}\right)^2 + a_{21}\left(\frac{\mathbf{k}\cdot\mathbf{q}}{q_0}\right)^2\right], \end{aligned} \quad \text{(A29)}$$

$$k_0 = \bar{a}_0\frac{\mathbf{k}\cdot\mathbf{p}}{p_0} + (1 - \bar{a}_0)\frac{\mathbf{k}\cdot\mathbf{q}}{q_0}. \quad \text{(A30)}$$

The inverse of Eq. (A26) for the classical result is

$$V(\mathbf{r}) = \frac{1}{(2\pi)^3}\int d\mathbf{k}\, e^{i\mathbf{k}\cdot\mathbf{r}}V(\mathbf{k}), \quad \text{(A31)}$$

and we shall put [as in Eq. (A19)]

$$V(\mathbf{r}) = V_1(\mathbf{r}) + V_{S1}(\mathbf{r}). \quad \text{(A32)}$$

In Ref. 10, we considered the one-gravitation exchange interaction between a spin-0 particle of mass m_1 and a spin-0 particle of mass m_2. The result [Eq. (5) of Ref. 10] corresponding to Eq. (A20) was not identical, but the classical result [Eq. (8) of Ref. 10] corresponding to Eq. (A27) was identical. The post-Newtonian result (i.e., result to order c^{-2}) for $V_1(\mathbf{r})$ is given by Eq. (33) of Ref. 10. Because of the choice of k_0^2 in the form of Eq. (A29), the post-Newtonian result for $V_1(\mathbf{r})$ is actually in the $\mathbf{r}_{*N}$ coordinate system as defined in Sec. II of this paper.

The post-Newtonian approximation to Eq. (A28) is

$$V_{S1}(\mathbf{k}) = -\frac{c^2\hbar^2\kappa^2}{4\mathbf{k}^2}\left[\frac{3\lambda_2}{4\lambda_1}i\boldsymbol{\sigma}^{(1)}\cdot(\mathbf{k}\times\mathbf{p}) - i\boldsymbol{\sigma}^{(1)}\cdot(\mathbf{k}\times\mathbf{q})\right]. \quad \text{(A33)}$$

Using Eq. (A1)–(A3) and letting $\frac{1}{2}\hbar\boldsymbol{\sigma}^{(1)} \to \mathbf{S}^{(1)}$ in Eq. (A33), we find, after using Eq. (A31) and

$$\frac{1}{(2\pi)^3}\int d\mathbf{k}\, e^{i\mathbf{k}\cdot\mathbf{r}}\frac{\mathbf{k}}{\mathbf{k}^2} = \frac{i\mathbf{r}}{4\pi r^3}, \quad \text{(A34)}$$

that the post-Newtonian approximation to $V_{S1}(\mathbf{r})$ is

$$\begin{aligned} V_{S1}(\mathbf{r}) = &(G/c^2r^3)[(3m_2/2m_1)\mathbf{S}^{(1)}\cdot(\mathbf{r}\times\mathbf{P}_1) \\ &- 2\mathbf{S}^{(1)}\cdot(\mathbf{r}\times\mathbf{P}_2)]. \end{aligned} \quad \text{(A35)}$$

To obtain both $V_{S1}(\mathbf{r})$ and $V_{S2}(\mathbf{r})$, one could consider the one-graviton exchange interaction between a spin-$\frac{1}{2}$ particle of mass m_1 and a spin-$\frac{1}{2}$ particle of mass m_2. However, it is much easier to obtain $V_{S2}(\mathbf{r})$ from $V_{S1}(\mathbf{r})$ by letting $1 \to 2$ and $2 \to 1$ and $\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2 \to -\mathbf{r}$, which gives us

$$\begin{aligned} V_{S2}(\mathbf{r}) = &(G/c^2r^3)[-(3m_1/2m_2)\mathbf{S}^{(2)}\cdot(\mathbf{r}\times\mathbf{P}_2) \\ &+ 2\mathbf{S}^{(2)}\cdot(\mathbf{r}\times\mathbf{P}_1)]. \end{aligned} \quad \text{(A36)}$$

Finally, noting that $\mathbf{P}_N = m_N\mathbf{v}_N$ to first order, we see that Eqs. (A35) and (A36) are in agreement with Eqs. (3.1) and (3.2).

[1] B. M. Barker and R. F. O'Connell, Phys. Rev. D **12**, 329 (1975).
[2] B. M. Barker and R. F. O'Connell, Phys. Rev. D **14**, 861 (1976).
[3] B. M. Barker, G. G. Byrd, and R. F. O'Connell, Astrophys. J. **305**, 623 (1986).
[4] B. M. Barker and R. F. O'Connell, Gen. Relativ. Gravit. **18**, 1055 (1986).
[5] R. A. Hulse and J. H. Taylor, Astrophys. J. Lett. **195**, L51 (1975).
[6] J. H. Taylor and J. M. Weisberg, Astrophys. J. **253**, 908 (1982).
[7] R. V. Wagoner and C. M. Will, Astrophys. J. **210**, 764 (1976).
[8] R. Epstein, Astrophys. J. **216**, 92 (1977).
[9] C. M. Will, *Theory and Experiment in Gravitational Physics* (Cambridge, New York, 1981).
[10] B. M. Barker and R. F. O'Connell, J. Math. Phys. **20**, 1427 (1979).
[11] C. F. Cho, and N. D. Hari Dass, Ann. Phys. (NY) **96**, 406 (1976).
[12] J. Schwinger, *Particles, Sources, and Fields* (Addison-Wesley, Reading, MA, 1970), Vol. I; Am. J. Phys. **42**, 507 (1974).
[13] W. Tulczyjew, Acta. Phys. Pol. **18**, 37, 534 E (1959).
[14] G. Börner, J. Ehlers, and E. Rudolph, Astron. Astrophys. **44**, 417 (1975).
[15] B. M. Barker and R. F. O'Connell, Phys. Lett. A **68**, 289 (1978). There are two misprints in Eq. (1) of this paper—a misplaced square bracket and missing exponent of 2. This equation is correctly given by Eq. (68) of Ref. 10.
[16] B. M. Barker and R. F. O'Connell, Ann. Phys. (NY) **129**, 358 (1980).
[17] B. M. Barker and R. F. O'Connell, Gen. Relativ. Gravit. **5**, 539 (1974).
[18] M. H. L. Pryce, Proc. R. Soc. London Ser. A **195**, 62 (1948).
[19] T. D. Newton and E. P. Wigner, Rev. Mod. Phys. **21**, 400 (1949).
[20] E. Corinaldesi and A. Papapetrou, Proc. R. Soc. London Ser. A **209**, 259 (1951).
[21] F. A. E. Pirani, Acta Phys. Pol. **15**, 389 (1956).
[22] P. D. D'Eath, Phys. Rev. D **12**, 2183 (1975).
[23] S. N. Gupta, *Quantum Electrodynamics* (Gordon and Breach, New York, 1977).
[24] B. M. Barker, S. N. Gupta, and R. D. Haracz, Phys. Rev. **149**, 1027 (1966).
[25] S. N. Gupta, Proc. Phys. Soc. London Ser. A **65**, 161, 608 (1952); Phys. Rev. **96**, 1683 (1954); Rev. Mod. Phys. **29**, 334 (1957); in *Recent Development in General Relativity* (Pergamon, New York, 1962), p. 251; Phys. Rev. **172**, 1302 (1968); Phys. Rev. D **14**, 2596 (1976).
[26] P. T. Matthews, Phys. Rev. **76**, 684 (1949).

An initial value gravitational quadrupole radiation theorem

Jeffrey Winicour
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260 and Max-Planck-Institut für Physik und Astrophysik, Institut für Astrophysik, 8046 Garching bei München, West Germany

(Received 9 September 1986; accepted for publication 5 November 1986)

A rigorous version of the quadrupole radiation formula is derived using the characteristic initial value formulation of a general relativistic fluid space-time. Starting from initial data for a Newtonian fluid, an algorithm is presented that determines characteristic initial data for a one-parameter family of general relativistic fluid space-times. At the initial time, a one-parameter family of space-times with this initial data osculates the evolution of the Newtonian fluid and has leading order news function equal to the third time derivative of the transverse Newtonian quadrupole moment.

I. INTRODUCTION

The goal of this paper is to derive a version of the quadrupole radiation formula which is based upon the characteristic initial value problem for a general relativistic fluid. Initial data for a λ-parameter family of general relativistic space-times are constructed from initial data for a spatially compact Newtonian fluid which provides a $\lambda = 0$ Newton–Cartan background. At the initial time, the quadrupole formula is obtained as an equality between the leading λ-order term of the Bondi news function and the third time derivative of the transverse Newtonian quadrupole moment. All assumptions concerning λ differentiability of the system are manifestly consistent. All asymptotic properties follow from the asymptotic properties of the Newton–Cartan background.

Many important results in general relativity have centered about initial value problems. In the case of the spacelike Cauchy problem, one notable example is the global existence and uniqueness theorems for solutions to the vacuum Einstein equations having asymptotically Euclidean spatial sections.[1] In the case of the characteristic initial value problem, one difficult feature of the ordinary Cauchy problem disappears, namely the differential constraints on the initial data. Yet there has been only slight progress toward proving global existence and uniqueness theorems.[2] Null hypersurfaces, with nonvanishing domain of dependence, contain caustics so that new difficulties arise from the lack of smoothness of the initial hypersurface.[3] In the presence of fluid sources, very little has been established in regard to global existence and uniqueness theorems even in the spacelike case; and for compact fluid sources no such theorems are known even in Newtonian hydrodynamics.

In this state of affairs, it is nevertheless possible to establish properties of formal solutions obtained by series expansion. In this manner, several derivations of the quadrupole formula have been based upon a harmonic coordinate approach.[4] Also, using the method of Newton–Cartan limits on null cones,[5,6] a derivation has been given in null coordinates, in which the calculations are simple enough so that the underlying ideas are not hidden.[7] The weakness of such results is that assumptions implicit in a perturbation expansion might, at higher orders, lead to mathematical inconsistencies in the case of radiative space-times. An alternative possibility is to avoid such inconsistencies by establishing properties of initial data sets rather than of solutions. Since data on an initial null hypersurface immediately exhibits gravitational radiation at null infinity, such a study of the quadrupole radiation problem is very natural. (In contrast, initial data on a spacelike hypersurface, though they contain this information, do not exhibit it directly. However, perhaps the *radiation reaction* problem could be formulated in terms of such data.)

Thus, rather than basing a strong version of the quadrupole radiation formula on overly optimistic assumptions as in Ref. 7, this paper establishes a weak initial value version using well founded assumptions. This at least provides some mathematically rigorous domain for its validity. Such a derivation was first given for a highly simplified dust model[8] but the calculations were far too complicated to generalize. The strategy here is to retain the same basic formulation as for the dust model but to employ the more powerful calculational approach of Ref. 7.

The main success of the characteristic initial value problem has been the description of gravitational waves in the asymptotic region far from the sources.[9] This has been extended to a global formalism, with fluid interior, by assuming the existence of outgoing null cones emanating from some central geodesic.[10] The null cone assumption appears to be justified for a wide class of astrophysical systems with intense gravitational fields. The original purpose was to apply this global formalism to a numerical study of the generation of gravitational waves. Similar numerical studies using more general null hypersurfaces have also been initiated.[3] A serious problem arises here concerning the presence of incoming radiation in the initial data. In the vacuum case, the requirement that the initial cone be shear-free gives the correct data for Minkowski space but, in the presence of matter, these data generally contain more incoming radiation than the amount of outgoing radiation generated by the matter.[10] Without some control over this incoming radiation, any numerical evolution or any statement of the quadrupole formula would be of little physical value.

 0022-2488/87/030668-09$02.50

How should one choose gravitational null data which, for a given matter distribution, exclude "too much" incoming radiation unrelated to the matter source? A tentative answer to this, proposed in Refs. 5 and 6 and used below, is the requirement of a Newton–Cartan limit based upon a family of outgoing null cones. In this limit, the gravitational null datum is determined by the background Newtonian potential. (Boundary conditions at infinity eliminate homogeneous contributions to the potential which might arise as a Newtonian limit of incoming radiation. In this context, see the Newtonian limit of plane waves.[11]) A λ-parameter system of space-times is described on a common manifold so that they share a family of outgoing null cones emanating from a timelike geodesic worldline. In a Bondi-type null coordinate system $x^\alpha = (x^0, x^1, x^A) = (u, r, \cos\theta, \phi)$ based upon these cones, the λ metric takes the λ-dependent Bondi form

$$ds^2 = [e^{2\lambda^2\beta}(1 + \lambda^2 W/r) - \lambda^4 r^2 h^{AB} U_A U_B]\,du^2 + 2\lambda e^{2\lambda^2\beta}\,du\,dr + 2\lambda^3 r^2 U_A\,du\,dx^A - \lambda^2 r^2 h_{AB}\,dx^A\,dx^B, \quad (1.1)$$

where $h^{AB}h_{BC} = \delta^A{}_C$, $\det(h_{AB}) = 1$, and $h_{AB} = q_{AB} + \lambda^2\gamma_{AB}$, with q_{AB} the unit-sphere metric. The fields β, W/r, rU_A, and γ_{AB} are required to be $O(r^2)$, near $r = 0$. Then, an admissible coordinate system in the neighborhood of the origin is given by $t = u + \lambda r$, $x = r\sin\theta\cos\phi$, $y = r\sin\theta \times \sin\phi$, and $z = r\cos\theta$, which agree with Fermi coordinates up to terms which do not involve the curvature. The explicit λ factors in (1.1) ensure that the $\lambda = 0$ limit yields the metric of Newton–Cartan theory with absolute time slices $u = \text{const}$, provided β, U^A, W, and γ_{AB} are smooth in this limit.

For matter source we adopt the λ-dependent ideal fluid stress-tensor

$$T_{\mu\nu} = (\rho + \lambda^2 p) w_\mu w_\nu - \lambda^2 p g_{\mu\nu} \quad (1.2)$$

with four-velocity w_μ of the form $w_\mu = t_{,\mu} + \lambda^2 v_\mu$. The explicit λ factors are sufficient to ensure that Einstein's equations yield, for $\lambda = 0$, the Euler–Poisson equations, which emerge in a polar coordinate system with freely falling origin. In this local inertial frame, the Newtonian potential Φ^* satisfies the boundary conditions

$$\Phi^*(u,0,x^A) = 0, \quad \Phi^*{}_{,1}(u,0,x^A) = 0 \quad (1.3)$$

and has the asymptotic behavior

$$\Phi^* = r\sum_m a_{1m}(u) Y_{1m} + a(u) + O\left(\frac{1}{r}\right) \quad (1.4)$$

at infinity. The remaining condition for the existence of a Newton–Cartan limit is that the $\lambda = 0$ limit of the Christoffel connection of (1.1) be the Newton–Cartan connection (whose geodesics are free-fall trajectories), which requires that

$$\lim_{\lambda\to 0}\left(\frac{W}{2r} + \beta\right) = \Phi^*.$$

This can be reformulated, using the hypersurface equations (see below), as the condition[5]

$$\lim_{\lambda = 0} q^A q^B (r^2\gamma_{AB,1})_{,1} = -4 q^A q^B \Phi^*_{:AB}, \quad (1.5)$$

where q_A is a complex null dyad for the unit sphere, $q_{AB} = 2q({}_A\bar{q}_B)$, and a colon denotes covariant differentiation with respect to q_{AB}. The left-hand side of (1.5) is the Weyl tensor version of the gravitational null datum. Here we only require (1.5) to hold in an osculatory sense in time [see Eq. (2.6)].

Einstein's equations $E_{\mu\nu} := G_{\mu\nu} + 8\pi T_{\mu\nu} = 0$, for this system, are equivalent to the four hypersurface equations

$$E_{1\mu} = 0, \quad (1.6)$$

the two gravitational evolution equations

$$E_{AB} - \tfrac{1}{2} g_{AB} g^{CD} E_{CD} = 0, \quad (1.7)$$

and the four fluid evolution equations

$$T^{\mu\nu}{}_{;\nu} = 0. \quad (1.8)$$

The remaining Einstein equations then follow from the Bianchi identities provided the smoothness conditions at the origin are satisfied.[10] All integration constants are uniquely determined by the vanishing of β, U^A, W, and γ_{AB} at the origin, which also follows from the smoothness conditions at the origin. Further details of this λ system are given in Ref. 6 and will be presented as they are required.

Section II reduces the quadrupole issue to a calculational problem by establishing the following lemma.

Lemma 1: Given smooth, compact initial fluid data for a Newtonian system, there exists initial null cone data for a λ-dependent family of general relativistic (GR) systems whose formal evolution, in the $\lambda = 0$ limit, osculates the Newtonian system in the sense of (2.6). The free gravitational null datum is a cubic polynomial in λ with uniformly smooth asymptotic behavior as described by (2.16) and (2.17).

Now the remaining problem is strictly calculational. Find the leading order news function for these data and compare it with the third time derivative of the quadrupole moment of the Newtonian background. To expedite this calculation a slow motion conformal Bondi frame is introduced in Sec. III. The quadrupole formula is then obtained in Sec. IV. The assumptions underlying this result and the conclusion can be summarized by the theorem below.

Initial Quadrupole Radiation Theorem: Given smooth, compact initial data for a Newtonian fluid, at u_0, there exists logarithmically asymptotic flat null data for a λ-dependent general relativistic fluid space-time with the following property. Any λ-dependent fluid space-time, which can be smoothly represented in null cone coordinates in a neighborhood of u_0, with this null data at u_0, has (i) a $\lambda = 0$ limit whose evolution osculates the Newtonian system in the sense of (2.6) and (ii) a leading order news function which satisfies the quadrupole radiation formula (4.8) at u_0.

Conventions and Notation: Our conventions are adopted to agree, as closely as possible, with those of Refs. 5–8. We use signature $+ - - -$; units for which $G = c = 1$; Greek letters ranging over 0–3 for space-time indices; lowercase Latin letters ranging over 1–3 for spatial indices: capital Latin letters ranging over 2–3 for indices on topologically spherical two-spaces; a semicolon to denote space-time covariant differentiation; a colon to represent covariant differentiation with respect to the unit sphere metric q_{AB}; a com-

ma for partial differentiation; a unit sphere dyad for which $q_{AB} = 2q(_A\bar{q}_{AB})$; and curvature conventions, for which $v_{\mu;\alpha\beta} - v_{\mu;\beta\alpha} = v_\nu R^\nu{}_{\mu\alpha\beta}$, $R_{\mu\nu} = R^\alpha{}_{\mu\nu\alpha}$, $R = R^\alpha{}_\alpha$, and the intrinsic scalar curvature of the unit sphere equals -2. The numerical conventions for the unit sphere spin-weight ladder operator $\eth$ are fixed by the examples $v_{A:B}q^Aq^B = \bar{\eth}(v_Aq^A)/\sqrt{2}$, $f^{:A}{}_A = \eth\bar{\eth}f$, and $(\bar{\eth}\eth - \eth\bar{\eth})\eta = 2s\eta$, for a spin-weight s quantity η. Acting on spherical harmonics, $\eth\bar{\eth}Y_{lm} = -l(l+1)Y_{lm}$. We write $f = \Sigma f^{(n)}\lambda^n$ for the expansions of λ-dependent fields. We use the shorthand notation

$$\int f = \int_0^r f(s)ds .$$

We denote by $\mathscr{P}_0$ the operator which projects out $l = 0$ harmonics and by $\mathscr{P}_1$ the operator which projects out $l = 0$ and $l = 1$ harmonics. In Sec. III, we introduce a Penrose compatification with conformally rescaled metric denoted by $\hat{g}_{\mu\nu}$.

II. QUASI-NEWTONIAN INITIAL DATA

We now discuss how smooth compact initial data ρ, v_i and an equation of state for a Newtonian fluid at initial time u_0 determine initial null data for the λ system through the requirement of the Newton–Cartan limit in the form (1.5).[5,6] We do not require that this limit exist for any finite time interval but only to an osculatory degree sufficient to formulate an initial value version of the quadrupole formula. For this purpose, we require that the limit condition (1.5) and its first three u derivatives are satisfied at u_0. In order to discuss such time derivatives determined from initial data by evolution equations, we introduce a λ family of Lorentz space-times in a neighborhood of u_0 which is described in null cone coordinates. Our primary interest is in those relationships at u_0 in which all time derivatives may be reexpressed in terms of derivatives tangential to u_0 by using evolution equations and coordinate conditions.

At u_0, the fluid data ρ and v_i for the λ system have their λ-independent Newtonian values. As gravitational null data at u_0, it is sufficient to adopt the cubic λ dependence

$$q^Aq^B\gamma_{AB} = q^Aq^B[\gamma^{(0)}_{AB} + \lambda\gamma^{(1)}_{AB} + \lambda^2\gamma^{(2)}_{AB} + \lambda^3\gamma^{(3)}_{AB}] . \tag{2.1}$$

For analyzing Einstein's equations it is convenient to introduce complex scalar potentials α and Z by

$$\eth^2\alpha = q^Aq^B\gamma_{AB,1} , \tag{2.2}$$

$$\eth Z = \sqrt{2}\, U_Aq^A , \tag{2.3}$$

and the Weyl tensor version of the null data

$$\psi = (r^2\alpha)_{,1} . \tag{2.4}$$

We fix the freedom in these potentials by requiring $\mathscr{P}_1\alpha = \alpha$ and $\mathscr{P}_0Z = Z$, i.e., α has no $l = 0$ and $l = 1$ parts and Z has no $l = 0$ part. This differs from the original gauge choice in Ref. 5 but it leads to some simplifications while introducing no changes in geometrically significant quantities. In terms of ψ the initial data takes the form

$$\psi = \psi^{(0)} + \lambda\psi^{(1)} + \lambda^2\psi^{(2)} + \lambda^3\psi^{(3)} . \tag{2.5}$$

(The boundary conditions at the origin imply that ψ, α, and $q^Aq^B\gamma_{AB}$ are equivalent ways to specify the gravitational data.) The osculatory version of the Newton–Cartan limit at u_0 now becomes

$$\frac{\partial^n}{\partial u^n}\psi^{(0)} = -2\mathscr{P}_1\frac{\partial^n}{\partial u^n}\Phi^* , \quad 0 \leqslant n \leqslant 3 . \tag{2.6}$$

The hypersurface equations (1.6) take the form

$$-4r\beta_{,1} = J_\beta , \tag{2.7}$$

$$(r^4Z_{,1})_{,1} = 2r^4(\beta/r^2)_{,1} - (2 + \eth\bar{\eth})r^2\alpha + J_Z , \tag{2.8}$$

$$W_{,1} = \frac{1}{4}\eth^2\bar{\eth}^2\int(\alpha + \bar{\alpha}) + (2 - \eth\bar{\eth})\beta + \frac{1}{4r^2}[r^4\eth\bar{\eth}(Z + \bar{Z})]_{,1} + J_W , \tag{2.9}$$

where the J's are quantities intrinsic to a single null hypersurface. They have the hierarchical form that J_β depends only upon the null data, J_z depends only upon the null data and β, and J_W depends only on the null data β, and Z. Thus they can be integrated in turn to find β, Z, and W given the null data. Smoothness at the origin determines the radial integration constants by integrating these equations from the origin, e.g., $\beta = -\frac{1}{4}\int J_\beta/r$. Furthermore, the J's are constructed using a finite number of differential and algebraic operations within a null hypersurface. As a result, given the polynomial λ dependence of the null data at u_0, they will be analytic functions of λ (including $\lambda = 0$) at u_0. Expressions for the J's, accurate up to order λ^4, are given in Appendix A. At each order, each $J^{(n)}$ is determined by gravitational null data up to $\psi^{(n-2)}$ and matter data up to $\rho^{(n)}$ and $v_i^{(n-1)}$.

The gravitational evolution equation (1.7) takes the form

$$2\lambda r\int\frac{\psi_{,0}}{r} = 2\beta - (r^2Z)_{,1} - J_\psi , \tag{2.10}$$

where J_ψ is again a hypersurface quantity which is determined by null data up to $\psi^{(n-2)}$ and matter data up to $\rho^{(n)}$ and $v_i^{(n-1)}$. Again, at $u = u_0$, J_ψ is an analytic function of λ. An expression for J_ψ accurate up to order λ^4 is given to Appendix A. The evolution equation can be combined with the hypersurface equation to yield[5,6]

$$r^2\nabla^2\psi = \frac{2\lambda}{r}[r^3\psi_{,0}]_{,1} + \frac{1}{r}\left[r^4\left(\frac{J_\psi}{r}\right)_{,1}\right]_{,1} + \mathscr{P}_1(J_\beta + J_{Z,1}) , \tag{2.11}$$

where ∇^2 is the Laplacian in spherical coordinates. Equation (2.11) gives a set of Poisson equations for the initial gravitational data. The source for $\psi^{(0)}$ is the matter density $\rho^{(0)}$ of the Newtonian background.

$$\nabla^2\psi^{(0)} = -8\pi\mathscr{P}_1\rho^{(0)}$$

in consistency with the Newton–Cartan limit condition (2.6).

The Poisson equation for $\psi^{(1)}$ has source depending upon $\psi^{(0)}{}_{,0}$ and initial data, so that the source can be constructed out of previously known quantities at u_0. Similarly, by u differentiation, the Poisson equation for $\psi^{(1)}{}_{,0}$ has source which can be constructed out of previously known quantities at u_0, after using the evolution equations to replace time derivatives in terms of derivatives tangential to u_0.

This continues for $\psi^{(1)}{}_{,00}$. That gives enough information to construct, in turn, sources for $\psi^{(2)}$, $\psi^{(2)}{}_{,0}$, and $\psi^{(3)}$, all at u_0, in terms of either previously known quantities or solutions of prior Poisson equations. Thus we have a scheme for determining the initial gravitational data $\psi^{(n)}$ $(0\leqslant n\leqslant 3)$ by solving a sequence of Poisson equations.

It remains to check whether the asymptotic properties of the source terms are consistent with the boundary condition that ψ vanish at infinity so that they determine unique solutions. The ability to eliminate homogeneous solutions of the Poisson equations, which do not vanish at infinity, is in fact our mechanism for restricting incoming radiation. It is self-evident from (2.6) that in the region exterior to the sources $\psi^{(0)}$ has analytic $1/r$ dependence. Specifically, $\psi^{(0)}=O(1/r^3)$ with

$$\eth^2\psi^{(0)}=3(Q/r^3)+O(1/r^4)\,. \tag{2.12}$$

Here Q is a pure spin-weight 2, $l=2$ quantity describing the transverse Newtonian quadrupole moment

$$Q=q^Aq^B(x^i/r)_{,A}\,(x^j/r)_{,B}Q_{ij}\,, \tag{2.13}$$

in terms of the Cartesian quadrupole components

$$Q_{ij}=\int\rho^{(0)}(x_ix_j-\tfrac{1}{3}r^2\delta_{ij})dV\,.$$

The behavior of $\psi^{(1)}$ and $\psi^{(2)}$ in the exterior has been worked out in Ref. 6. For $\psi^{(1)}$ the exterior self-gravitational source terms are linear and simple to analyze. At u_0, we again have analytic $1/r$ dependence in the exterior region with

$$\psi^{(1)}=O(1/r^3)\,. \tag{2.14}$$

The same behavior applies to the first two u derivatives of $\psi^{(1)}$, at u_0, which can be calculated from our knowledge of the first three u derivatives of $\psi^{(0)}$, at u_0. For $\psi^{(2)}$, the solution is complicated by nonlinear, noncompact source terms. Not all details of the exterior behavior have been worked out but it has been shown[6] that in the exterior, $\psi^{(2)}$ has the form

$$\psi^{(2)}=\psi_A{}^{(2)}+\psi_L{}^{(2)}\ln r/r\,, \tag{2.15}$$

where $\psi^{(2)}{}_A$ and $\psi_L{}^{(2)}$ are analytic in $1/r$ and both have $O(1/r^3)$ asymptotic behavior. At u_0, these results hold in the present case for both $\psi^{(2)}$ and its first u derivative. Whether $\psi_L{}^{(2)}$ can actually be nonzero has not been established but the important feature here is that at worst it has $O(1/r^3)$ dependence.

The exterior behavior of $\psi^{(3)}$ was not explored in Ref. 6 but the results for the dust model[8] explicitly show the existence of a $\ln r/r^3$ term. This is the first order at which asymptotic behavior inconsistent with the peeling property arises. However, most of the standard properties of null infinity remain intact for this weaker logarithmic version of asymptotic flatness, e.g., the existence of a conformal boundary at null infinity, the BMS group and the properties of the Bondi mass and news function.[12] In Appendix A we show that the asymptotic behavior of $\psi^{(3)}$ for the dust solution characterizes the worst possible behavior in the general case,

$$\psi^{(3)}=\psi_A{}^{(3)}+\psi_L{}^{(3)}\ln r,$$

where $\psi_A{}^{(3)}$ is analytic in $1/r$ and both $\psi_A{}^{(3)}$ and $\psi_L{}^{(3)}$ are $O(1/r^3)$. Whether $\psi_L{}^{(3)}$ is analytic in $1/r$ has not been investigated but it is shown in Appendix A that it satisfies the asymptotic uniform smoothness conditions that if $f(u_0,r,x^A)=O(g(r))$ then

$$\frac{\partial f}{\partial x^A}=O(g)\quad\text{and}\quad\frac{\partial f}{\partial r}=O\left(\frac{g}{r}\right). \tag{2.16}$$

These conditions are also possessed by all the metric variables and their u derivatives at u_0, as determined by evolution equations.

In summary, at u_0 we can set

$$\psi=\psi_A+\lambda^2\psi_L{}^{(2)}\ln r/r+\lambda^3\psi_L{}^{(3)}\ln r\,, \tag{2.17}$$

where ψ_A is analytic, the ψ_L's are uniformly smooth, and all the ψ's are $O(1/r^3)$. We have thus established Lemma 1, stated in the Introduction.

III. THE CONFORMAL BONDI FRAME

An asymptotic approach can be used to expedite the calculation of the news function for the data determined in the last section.[7] Since the news function is a purely geometrical field it may be calculated in any coordinate system. The first step is to demonstrate the existence of an asymptotic Bondi frame in which (1.5) is still valid. For this purpose it is convenient to construct a Penrose compactification[13] of a neighborhood of the exterior region of the initial null cone by attaching a portion of $\mathscr{I}^+$ to the physical space-time. An appropriate conformal factor and new radial coordinate is the inverse luminosity distance[14] $\Omega=l:=1/r$. Then the conformal metric $\hat g_{\mu\nu}=\Omega^2g_{\mu\nu}$ is given by

$$\begin{aligned}d\hat s^2=&\,[e^{2\lambda^2\beta}l^2(1+\lambda^2Wl)\\&-\lambda^4h^{AB}U_AU_B]du^2-2\lambda e^{2\lambda^2\beta}\,du\,dl\\&+2\lambda^3U_A\,du\,dx^A-\lambda^2h_{AB}\,dx^A\,dx^B\,.\end{aligned} \tag{3.1}$$

In this coordinate system (u,l,x^A), ψ as well as the auxiliary variables β, U_A, W, and γ_{AB} are still analytic in λ and satisfy the uniform smoothness conditions (2.16) with respect to l at u_0.

However, the metric variables do not vanish at $\mathscr{I}^+$. Their asymptotic forms may be inferred from the equations

$$(\hat\nabla_\mu\Omega)\hat\nabla^\mu\Omega=O(\Omega)$$

and

$$\hat\nabla_\mu\hat\nabla_\nu\Omega-\tfrac{1}{4}\hat g_{\mu\nu}\hat\nabla^\rho\hat\nabla_\rho\Omega=O(\Omega)\,,$$

which follow from the existence of a conformal boundary and the vacuum Einstein equations.[14] This gives

$$h_{AB}=q_{AB}+\lambda^2K_{AB}+O(l)\,, \tag{3.2}$$

$$\beta=H+O(l^2)\,, \tag{3.3}$$

$$U_A=h_{AB}(L^B+2re^{2\lambda^2H}D^BH)+O(l^2)\,, \tag{3.4}$$

$$Wl^2=D_AL^A+O(l)\,, \tag{3.5}$$

where H,L^A, and K_{AB} are independent of l, with

$$\lambda K_{AB,u}=(-h_{AC}D_B-h_{BC}D_A+\tfrac{1}{2}h_{AB}D_C)L^C+O(l)\,, \tag{3.6}$$

and where D_A is the covariant derivative with respect to h_{AB}. (Alternatively, these properties could be established directly

in the physical space from an asymptotic integration of the hypersurface and evolution equations.)

The asymptotic metric terms (H, L^B, and K_{AB}) arise because our coordinates were chosen to be locally inertial at the origin. They would vanish in a conformal Bondi frame, which corresponds as closely as possible to an asymptotic inertial frame. However, carrying out such a transformation directly would disrupt the λ analyticity of the metric variables, i.e., $1/\lambda$ terms would arise. This would then complicate the bookkeeping of λ orders in the calculation of the news. Such problems are avoided by introducing a slow motion time $(u-u_0)=\lambda(v-v_0)$ and a further conformal rescaling $\tilde{g}_{\mu\nu}=\hat{g}_{\mu\nu}/\lambda^2$ so that

$$d\tilde{s}^2=[e^{2\lambda^2\beta}l^2(1+\lambda^2 Wl) -\lambda^4 h^{AB}U_A U_B]dv^2-2e^{2\lambda^2\beta}\,dv\,dl +2\lambda^2 U_A\,dv\,dx^A-h_{AB}dx^A\,dx^B. \tag{3.7}$$

In the $\lambda=0$ limit, $d\tilde{s}^2$ leads to a background Minkowski geometry. Equation (3.7) provides a slow motion post-Minkowski representation of the post-Newton–Cartan initial value system.[6]

The goal now is to remove the asymptotic metric terms by a coordinate transformation $(v',l',x^{A'})=x^{\alpha'}(x^\beta)$ leading to a conformal Bondi frame. In doing so, the conformal metric must also undergo a conformal transformation $\tilde{g}'^{\mu\nu}=\omega^{-2}\tilde{g}^{\mu\nu}$ in order to retain $\det(h_{AB})=1$. The total change is

$$\tilde{g}'^{\mu'\nu'}=\left[\det\left(\frac{\partial x^{A'}}{\partial x^B}\right)\right]^{-1}x^{\mu'}{}_{,\alpha}x^{\nu'}{}_{,\beta}g^{\alpha\beta} \tag{3.8}$$

and the new conformal factor $\Omega'=l'$ is given by

$$(l')^2=l^2\det\left(\frac{\partial x^{A'}}{\partial x^B}\right). \tag{3.9}$$

This determines the new l' in terms of the new $x^{A'}$. Also, once the new v' and $x^{A'}$ are known at $\mathscr{I}^+$ their values interior to $\mathscr{I}^+$ are determined by requiring that the null coordinate conditions

$$\tilde{g}'^{0'0'}=0 \quad\text{and}\quad \tilde{g}'^{0'A'}=0 \tag{3.10}$$

are maintained.

Consider first transformations with $v'=v$, with an eye toward setting K_{Ab} and L^A to zero in the new frame. This requires that the new $\tilde{g}^{1A}$ must vanish at $\mathscr{I}^+$ or, according to (3.8),

$$x^{A'}{}_{,v}=-\lambda^2 x^{A'}{}_{,B}L^B.$$

We thus have $x^{A'}=x^A+\lambda^2 y^A$, where y^A is zeroth order in λ and, at $\mathscr{I}^+$, satisfies

$$y^A_{,v}=-L^A-\lambda^2 y^A{}_{,B}L^B. \tag{3.11}$$

We need only set the first two v derivatives of L^A equal to zero, at u_0. (We continue to refer to the initial null hypersurface as u_0.) At u_0, we choose $x^A+\lambda^2 y^A$ to generate the conformal mapping that initially sets $K_{AB}=0$. The first two v derivatives of y^A, at $\mathscr{I}^+$, are then determined by (3.11) and then (3.6) ensures K_{AB} remains zero up to its first three v derivatives.

Transforming K_{AB} and L^A to zero in this way also changes $\gamma^{(0)}_{AB}$. It is important to check that there is no change in $l^2\gamma^{(0)}_{AB,ll}=(r^2\gamma^{(0)}_{AB,r})_{,r}$ so that the Newton–Cartan limit condition (1.5) is still valid in the new coordinate system. Interior to $\mathscr{I}^+$, (3.10) gives $y^A{}_{,l}=0$ so that y^A has no l dependence and the l derivatives of $\gamma^{(0)}_{AB}$ are not changed, as desired.

Assuming now that L^A and K_{AB} have been transformed to zero. We next transform H to zero while keeping $x^{A'}=x^A$ at $\mathscr{I}^+$, so that L^A and K_{AB} remain zero. After setting $v'=v+\lambda^2 y$, the condition that the new H' vanishes reduces to

$$y_{,v}=(1-e^{2\lambda^2 H})/\lambda^2 \tag{3.12}$$

at $\mathscr{I}^+$. At u_0, we take $y=0$ so that the initial null hypersurface is unchanged. Then (3.12), or its higher v derivative counterparts, determines all necessary v derivatives at $\mathscr{I}^+$. It remains to check the change in $\gamma^{(0)}_{AB,ll}$ under this transformation. At interior points, (3.10) gives

$$y_{,l}=O(\lambda^2)\quad\text{and}\quad y^A{}_{,1}=q^{AB}y_{,B}+O(\lambda^2),$$

where we again set $x^{A'}=x^A+\lambda^2 y^A$ (but now with $y^A=0$ at $\mathscr{I}^+$). Thus $y^A=q^{AB}y_{,B}l+O(\lambda^2)$. As a result, $\gamma^{(0)}_{AB}$ and the leading order shear tensor $\gamma^{(0)}_{AB,l}$ change, but $\gamma^{(0)}_{AB,ll}$ remains unchanged.

Thus for the λ-dependent system there exists a conformal Bondi frame, at u_0, in which

$$q^Aq^Bl^2\gamma_{AB,ll}=-4q^Aq^B\Phi^*{}_{;AB}+O(\lambda),$$

or, equivalently,

$$\psi=-2\mathscr{P}_1\Phi^*+O(\lambda).$$

The transformation to this frame preserves analyticity with respect to λ and uniform smoothness with respect to l. Recall that the original coordinates for the physical spacetime covered a neighborhood of u_0 which was introduced to represent time derivatives at u_0 without explicitly reexpressing them, via evolution equations, in terms of derivatives intrinsic to u_0. The same applies to the conformal Bondi coordinates.

The original physical time u and the slow motion Bondi time v are related, at u_0, by

$$\lambda\frac{\partial}{\partial u}=\frac{\partial}{\partial v}+O(\lambda^2), \tag{3.13}$$

where the remainder satisfies $(\partial/\partial u)O(\lambda^2)=O(\lambda^2)$.

As a result, the osculating Newton–Cartan limit conditions (2.6), at u_0, take the form in the Bondi frame

$$\left(\frac{\partial}{\partial v}\right)^n\psi=-2\mathscr{P}_1\left(\lambda\frac{\partial}{\partial u}\right)^n\Phi^*+O(\lambda^{n+1}),\qquad 0\leqslant n\leqslant 3. \tag{3.14}$$

This implies certain slow motion properties of the Bondi frame, e.g., $\partial\psi/\partial v=O(\lambda)$ which follows from (3.13) and the fact that the $\lambda=0$ limit of $\partial\psi/\partial u$ exists. In particular, according to (2.12) and (3.13), we have, at u_0,

$$\lim_{l\to 0}l^{-3}\eth^2\psi_{,vvv}=3\lambda^3 Q_{,uuu}+O(\lambda^4). \tag{3.15}$$

This is the key equation relating the quadrupole moment of

the background Newtonian system to the exterior field that will be used in the next section.

IV. THE CALCULATION

In the interior space-time, four of Einstein's equations were replaced by the matter equations $T^{\mu\nu}{}_{;\nu} = 0$. In the exterior, although $T^{\mu\nu}$ vanishes, the remaining hypersurface and gravitational evolution equations do not imply the vacuum equations unless the supplementary conditions $R_{00} = R_{0A} = 0$ are imposed on some world tube.[14] With an interior, all radial integration constants can be fixed by smoothness conditions at the origin but, otherwise, those describing the mass and angular momentum must be initially specified on a spherical cross section of the world tube. The supplementary conditions then determine their time dependence. In the conformal Bondi frame constructed in the last section, we can choose this world tube to be $\mathscr{I}^+$, with the initial cross section at u_0.

An analysis of Einstein's equation in a conformal Bondi frame is given in Ref. 14 and the modifications to include $\ln r$ terms are discussed in Ref. 12. A brief λ-dependent version based upon those references is presented in Appendix B.

Here we need only consider the leading asymptotic behavior. In the conformal Bondi frame, the null data for the λ-dependent system corresponding to (2.17) has the asymptotic form

$$\psi = kl^3 + \lambda^2 jl^3 \ln l + O(l^4 \ln l) , \tag{4.1}$$

where $k_{,v} = O(\lambda)$ and $j_{,v} = O(\lambda)$, in terms of the slow time v. Integration of $q^A q^B l^2 \gamma_{AB,ll} = \eth^2 \psi$ then leads to

$$q^A q^B \gamma_{AB} = \eth^2 \left[cl + \tfrac{1}{6}(k - \tfrac{5}{6}\lambda^3 j) l^3 + \tfrac{1}{6}\lambda^3 j l^3 \ln l \right] + O(l^4 \ln l) , \tag{4.2}$$

where $c(v, x^A, \lambda)$ is a radial integration constant whose time derivative determines the Bondi news function.

The evolution equation (B5) gives $j_{,v} = 0$ and

$$\eth^2 k_{,v} = -\tfrac{1}{2}\eth^2 \eta , \tag{4.3}$$

where η is a spin-weight 0 potential for the angular momentum aspect. The supplementary conditions (B6) and (B7) give

$$M_{,v} = \tfrac{1}{8}\eth^2 \bar{\eth}^2 (c + \bar{c})_{,v} - (\lambda^2/4)(\eth^2 c_{,v})(\bar{\eth}^2 \bar{c}_{,v}) , \tag{4.4}$$

$$\eth \eta_{,v} = \eth \left[-2M - \tfrac{1}{2}\eth^2 \bar{\eth}^2 (\bar{c} - c) \right] + \lambda^2 \{ \tfrac{3}{8}\eth [(\eth^2 c)\bar{\eth}^2 \bar{c}]_{,v} - 2\eth [(\bar{\eth}^2 \bar{c})\eth^2 c_{,v}] - \tfrac{3}{2}(\eth^3 c)\bar{\eth}^2 \bar{c}_{,v} + \tfrac{1}{2}(\eth \bar{\eth}^2 \bar{c})\eth^2 c_{,v} \} , \tag{4.5}$$

where M is the mass aspect. By taking the second v derivative of (4.3), M and η can be eliminated via (4.4) and (4.5). There results

$$\begin{aligned} \eth^2 k_{,vvv} = {} & \eth^2 \{ \tfrac{1}{8}\eth^2 \bar{\eth}^2 (\bar{c} + c) + \tfrac{1}{4}\eth^2 \bar{\eth}^2 (\bar{c} - c) \}_{,v} \\ & + \tfrac{1}{2}\lambda^2 \eth \{ -\tfrac{1}{2}\eth [(\eth^2 c_{,v})\bar{\eth}^2 \bar{c}_{,v}] \\ & - \tfrac{3}{8}\eth [(\eth^2 c)(\bar{\eth}^2 \bar{c})]_{,vv} + 2\eth [(\bar{\eth}^2 \bar{c})\eth^2 c_{,v}]_{,v} \\ & + \tfrac{3}{2}[(\eth^3 c)\bar{\eth}^2 \bar{c}_{,v}]_{,v} - \tfrac{1}{2}[(\eth \bar{\eth}^2 \bar{c})\eth^2 c_{,v}]_{,v} \} . \end{aligned} \tag{4.6}$$

We now apply these results to our initial data at u_0. Since $\psi_{,vvv} = k_{,vvv} l^3 + O(l^4 \ln l)$, (3.15) implies that the right-hand side of (4.6) must be $O(\lambda^3)$. That requires $c_{,v} = O(\lambda^3)$. Similary, since Q is a pure spin-2 electric quadrupole, $(\bar{c} - c) = O(\lambda^4)$ and $\eth^2 \bar{\eth}^2 c_{,v} = 24 c_{,v} + O(\lambda^4)$. Then (3.15) and (4.6) combine to give

$$2\eth^2 c_{,v} = \lambda^3 Q_{,uuu} + O(\lambda^4) .$$

Here the left-hand side is the news function (with the numerical factor chosen to agree with Bondi's original definition). We have thus established, at u_0, the quadrupole radiation formula

$$N^{(0)} = Q_{,uuu} , \tag{4.7}$$

where $N^{(0)}$ is the leading order news function for the λ-dependent system. By construction of the initial null data, any space-time determined by this data must satisfy (4.7) at u_0. This establishes the initial Quadrupole Radiation Theorem stated in the Introduction.

V. DISCUSSION

The theorem just established provides a rigorous model for the widely accepted idea that the quadrupole formula should approximately describe the radiation from a quasi-Newtonian system. The theorem gives no error estimates but it does point to some possible causes of error. These can be separated into truncation error and evolution error.

Given data at u_0 for third-order Newtonian osculation, i.e., satisfying (2.6), the truncation error $N(\lambda) - N^{(0)}$, in the news function at u_0, arises from considering only the leading order term $N^{(0)}$. Although a general error bound might be difficult to obtain, the error for any specific system could be calculated numerically, using already developed codes in the axisymmetric case.[10] From physical considerations, one would expect this error to be significant, say for $\lambda = 1$, when the background Newtonian system had fluid velocities or escape velocities comparable to the velocity of light. And one would expect the approximation to be meaningless were the background incompatible with a λ family of null cones because of caustics arising from strong light bending properties. Such nonperturbative effects would not be accessible by this approach. Also, there is the degenerate case in which the Newtonian $Q_{,uuu}$ vanishes, at u_0, so that $N^{(0)} = 0$. On the basis of linearized theory, the next order term in the news function $N^{(1)}$ should then represent octupole radiation but, in our formalism, that would require an additional order of osculation with the Newtonian background.

This leads us to the issue of errors arising in the evolution. In the future of u_0, the $\lambda = 0$ limit of the space-time evolved from the λ-dependent general relativistic data no longer exactly equals the background Newton–Cartan space-time. Thus the quadrupole formula would not continue to hold exactly. The consequent error in applying the formula could be attributed to a basic inadequacy in approximating a general relativistic system by a Newtonian one over both large time and distance scales. It would then be important to know over what time intervals the quadrupole formula remains a good approximation.

Another possible interpretation of the evolution error is in terms of the inadequacy of the initial data. The osculation at u_0 with the Newtonian background might be imposed to a higher order than (2.6). Osculation to all orders might even be demanded, as assumed in Ref. 7. A potential problem here

is that asymptotic flatness is already weakened to a logarithmic version for third-order osculation so that for higher orders the asymptotic behavior might not remain physically reasonable. Models of slow motion expansions for the nonlinear wave equation $\Box\Psi + \Psi^p = S$, for $p \geqslant 3$, indicate that such drastic asymptotic behavior is inevitable at some order of osculation.[15]

Such deviations from asymptotic flatness stem from nonlinear, noncompact source terms. These terms are also responsible for backscattering, which obscures the degree to which the requirement of a Newton–Cartan limit eliminates incoming radiation. Can both of these difficulties be avoided by removing the troublesome nonlinear terms in setting up the initial data? (A method for removing logarithmic terms at third order was found for the dust model[8] but it might not generalize naturally.) For example, the linearized version of the null cone formulation of Einstein's equations might be used for determining initial data, i.e., demand osculation to all orders between the Newton–Cartan and linearized spacetimes. No logarithm terms would arise at each order since the source terms would be compact. Given this data, one could compare its full nonlinear evolution with the background. All these considerations deserve further exploration.

ACKNOWLEDGMENTS

The ideas of this paper have been sharpened by insights and suggestions from the Munich Relativity Group, in particular J. Ehlers and B. Schmidt. I am especially indebted to J. Ehlers for a thorough discussion of the manuscript.

This research was supported by the Alexander-von-Humboldt Stiftung and by NSF Grant No. PHY-8403708.

APPENDIX A: QUASI-NEWTONIAN FORMULA

To calculate the J's introduced in Sec. II, Einstein's equation is expanded in terms of β, Z, W, α, ρ, v_i, and p. In this process, it is convenient to express the contravariant two-metric h^{AB}, in terms of a dyad $h^{AB} = 2m^{(A}\bar{m}^{B)}$ with the expansion

$$m^A = (1 + \lambda^4 Q)q^A + \lambda^2 P \bar{q}^A , \tag{A1}$$

in terms of the auxiliary variables P and Q. The phase freedom in m^A is fixed here by the requirements

$$[m^A]_{r=0} = q^A \quad \text{and} \quad m^A{}_{,1}\bar{m}_A = 0 .$$

To the order required in this paper, only $P^{(0)}$ appears in the J's and may be reexpressed in terms of $\alpha^{(0)}$ by

$$2P^{(0)} = -\eth^2 \int \alpha^{(0)} . \tag{A2}$$

Straightforward calculation then leads to

$$J_\beta = -8\pi r^2(\rho + \lambda^2 p)(1 + \lambda v_1)^2 - \tfrac{1}{2}\lambda^2 r^2(\eth^2\alpha)\bar{\eth}^2\bar{\alpha} + O(\lambda^4) , \tag{A3}$$

$$\begin{aligned}\eth J_Z = {} & 16\pi\lambda\sqrt{2}r^2(\rho + \lambda^2 p)(1 + \lambda v_1)q^A v_A + 2\lambda^2(r^4\beta\eth Z_{,1})_{,1} \\ & + \lambda^2[r^4(\eth^2\alpha)\bar{\eth}\bar{Z}]_{,1} + r^2\lambda^2[P\eth\bar{\eth}^2\bar{\alpha} - \bar{P}\eth^3\alpha - 2(\eth\bar{P})\eth^2\alpha] + O(\lambda^4) ,\end{aligned} \tag{A4}$$

$$\begin{aligned}J_w = {} & -4\pi[\rho r^2 + \lambda^2\rho(2r^2\beta + q^{AB}v_A v_B) - \lambda^2 p r^2] - (3\lambda^2/2)\eth\bar{\eth}(P\bar{P}) + (\lambda^2/2)\eth(P\bar{\eth}\bar{P}) + (\lambda^2/2)\bar{\eth}(\bar{P}\eth P) \\ & - \lambda^2\beta(\bar{\eth}^2 P + \eth^2\bar{P}) + 2\lambda^2\beta(1 - \eth\bar{\eth})\beta - \lambda^2(\eth\beta)\bar{\eth}\beta - \lambda^2\bar{\eth}(P\bar{\eth}\beta) - \lambda^2\eth(\bar{P}\eth\beta) \\ & + (\lambda^2/2r^2)[r^4\bar{\eth}(P\bar{\eth}\bar{Z}) + r^4\eth(\bar{P}\eth\bar{Z})]_{,1} - (\lambda^2 r^4/4)(\bar{\eth}\bar{Z}_{,1})\eth Z_{,1} + O(\lambda^4),\end{aligned} \tag{A5}$$

$$\begin{aligned}\eth^2 J_\psi = {} & -16\pi\lambda^2\rho(v_A q^A)^2 - \lambda^2(rW\eth^2\alpha)_{,1} - 4\lambda^2\beta\eth^2\beta - 2\lambda^2(\eth\beta)^2 - 4\lambda^2 P\eth\bar{\eth}\beta + 2\lambda^2[(\eth\beta)\bar{\eth}P - (\bar{\eth}\beta)\eth P] \\ & + \lambda^2 P\eth\bar{\eth}[r^2(Z + \bar{Z})]_{,1} - (\lambda^2 r^4/2)(\eth Z_{,1})^2 + \lambda^2 r^2(\bar{\eth}\eth^2\alpha)\eth Z + (\lambda^2 r^2/2)(\eth^2\alpha)\eth\bar{\eth}(Z - \bar{Z}) \\ & + \lambda^2[(\eth P)(r^2\bar{\eth}\bar{Z})_{,1} - (\bar{\eth}P)(r^2\eth Z)_{,1}] + O(\lambda^4) .\end{aligned} \tag{A6}$$

It is convenient to have the following combination which appears in S:

$$\begin{aligned}\eth(J_\beta + J_{Z,1}) = {} & -8\pi r^2\eth[(\rho + \lambda^2 p)(1 + \lambda v_1)^2] + 16\pi\lambda\sqrt{2}[r^2(\rho + \lambda^2 p)(1 + \lambda v_1)q^A v_A]_{,1} \\ & + \lambda^2\{r^4[2\beta\eth Z_{,1} + (\eth^2\alpha)\bar{\eth}\bar{Z}]\}_{,11} + \lambda^2[P\eth\bar{\eth}^2 - \bar{P}\eth^3 - 2(\eth\bar{P})\eth^2](r^2\alpha)_{,1} + O(\lambda^4) .\end{aligned} \tag{A7}$$

We now investigate the asymptotic behavior of $\psi^{(3)}$, at u_0. In accord with (2.11), it satisfies

$$\nabla^2\psi^{(3)} = S^{(3)} , \tag{A8}$$

where

$$\begin{aligned}S^{(3)} = {} & \frac{2}{r^3}[r^3\psi^{(2)}_{,0}]_{,1} + \frac{1}{r^3}\left[r^4\left(\frac{J^{(3)}_\psi}{r}\right)_{,1}\right]_{,1} \\ & + \mathscr{P}_1(J^{(3)}_\beta + J^{(3)}_{Z,1}) .\end{aligned} \tag{A9}$$

A theorem due to Persides[16] will be helpful in that regard.

Theorem A: Let $f(\bar{r})$ be a continuous and bounded function. The necessary and sufficient conditions for $\nabla^2\phi = f$ to have a solution of the form

$$\Phi(\bar{r}) = \sum_{k\geq 0}\frac{\Phi_k(\theta,\phi)}{r^{1+k}} \tag{A10}$$

outside some radius $r \geqslant r_0$ are

$$f(\bar{r}) = \sum_{k\geq 0}\frac{f_k(\theta,\phi)}{r^{3+k}} , \tag{A11}$$

$$\oint f_l(\theta,\phi)\,Y_{lm}(\theta,\phi)\,d\Omega = 0\,. \tag{A12}$$

for $r \geqslant r_0$.

Before applying this theorem to the source $S^{(3)}$, we list some lower-order results which were established in Ref. 6 at any time u for which the Newtonian limit condition (2.6) holds. In the region $r > r_0$, exterior to the matter source,

$$\alpha^{(n)} = -c_{(n)}/r^2 + O(1/r^4)\,, \tag{A13}$$

$$P^{(n)} = -\eth^2\left(\frac{K^{(n)}}{2} + \frac{c^{(n)}}{2r}\right) + O\left(\frac{1}{r^3}\right), \tag{A14}$$

$$\beta^{(n)} = H^{(n)}\,, \tag{A15}$$

$$Z^{(n)} = L^{(n)} + \frac{2H^{(n)}}{r} + \frac{(2+\eth\bar{\eth})c^{(n)}}{2r^2} + O\left(\frac{1}{r^3}\right) \tag{A16}$$

for $0 \leqslant n \leqslant 1$. Here $c^{(n)}$, $H^{(n)}$, and $L^{(n)}$ are independent of r and satisfy

$$c^{(0)} - \bar{c}^{(0)} = 0, \quad A^{(0)} = 0, \quad A^{(1)} - \bar{A}^{(1)} = 0\,. \tag{A17}$$

In addition, all the remainder terms in (A13)–(A16) are analytic in $1/r$. Also, for $r \geqslant r_0$, $\psi^{(2)}_{,0}$ has the form

$$\psi^{(2)}_{,0} = \sum_{k\geqslant 2} \frac{a_k(\theta,\phi)}{r^{1+k}} + \sum_{l\geqslant 3} \frac{b_{lm}\,Y_{lm}\,\ln r}{r^{l+1}}\,. \tag{A18}$$

(Recall that neither ψ nor S have $l=0$ or $l=1$ parts.) The critical result is that there is no quadrupole contribution to the $\ln r$ series in (A18).

Using (A13)–(A17), direct substitution into (A2)–(A9) leads to a $S^{(3)}$ of the form, for $r \geqslant r_0$,

$$S^{(3)} = (2/r^3)\left[r^3\psi^{(2)}_{,0}\right]_{,1} + O(1/r^5)\,,$$

where the remainder is analytic in $1/r$. Using (A18), we may then set

$$S^{(3)} = f + \sum_{l\geqslant 2} \frac{A_{lm}}{r^{l+3}} + \sum_{l\geqslant 3} \frac{b_{lm}\,Y_{lm}\,\ln r}{r^{l+2}}\,, \tag{A19}$$

for $r \geqslant r_0$ (where r_0 is chosen to be sufficiently large to guarantee convergence). Here f satisfies the conditions of Theorem A and, furthermore, $f = O(1/r^5)$. Thus, for $r \geqslant r_0$, a solution of $\nabla^2\Phi = f$, which vanishes at infinity, has the analytic form (A10), with $k \geqslant 2$. Also, for $r \geqslant r_0$, a solution of the Poisson equation, which vanishes at infinity and whose source is the A_{lm} series in (A19), is given by

$$-\sum_{l\geqslant 2} \frac{A_{lm}\,Y_{lm}\,\ln r}{(2l+1)r^{l+1}}\,,$$

so that it is $O(\ln r/r^3)$ and uniformly smooth with respect to r. A corresponding solution for the B_{lm} series in (A19) is

$$\sum_{l\geqslant 3} \frac{B_{lm}\,Y_{lm}}{2lr^l}\left[\frac{(2l-1)}{2l} - \ln r\right],$$

which, again, is $O(\ln r/r^3)$ and uniformly smooth.

Summing these individual results, the contribution to $\psi^{(3)}$ for the source (A19), in the region $r \geqslant r_0$, is $O(\ln r/r^3)$ and uniformly smooth. Since the contribution from the compact region $r \leqslant r_0$ is analytic and $O(1/r^3)$, $\psi^{(3)}$ has the asymptotic behavior necessary to establish (2.17).

APPENDIX B: λ-DEPENDENT CONFORMAL BONDI FRAME

We describe the λ-dependent solutions of Einstein's vacuum equations in terms of the conformally rescaled Bondi metric $\tilde{g}_{\mu\nu} = \Omega^2 g_{\mu\nu}$ given by (3.7), with $h_{AB} = q_{AB} + \lambda^2\gamma_{AB}$, in the (v,l,x^a) Bondi coordinates. We need only terms up to the orders in λ and l corresponding to the remainder terms in the logarithmically asymptotically flat[12] null data

$$\gamma_{AB} = lc_{AB} + \frac{\lambda^2 l^2}{4} q_{AB}c^{DE}c_{DE} + \frac{l^3}{6}\left(k_{AB} - \frac{5}{6}\lambda^3 j_{AB}\right) + O(l^4) + \lambda^3\left[(l^3\ln l/6)j_{AB} + O(l^4\ln l)\right] \tag{B1}$$

(where indices are raised with respect to q^{AB}; i.e., $c^{AB} = q^{AD}q^{BE}c_{DE}$). Here c_{AB}, k_{AB}, and j_{AB} are l independent but λ dependent, the remainders are uniformly smooth and j_{AB} represents the leading order logarithmic dependence. The condition that $h^{AB}\gamma_{AB,1} = 0$ dictates the form of the $O(l^2)$ term in (B1) and also leads to the trace conditions $q^{AB}c_{AB} = q^{AB}k_{AB} = q^{AB}j_{AB} = 0$. Introducing potentials $q^Aq^Bc_{AB} = \eth^2 c$, $q^Aq^Bk_{AB} = \eth^2 k$, and $q^Aq^Bj_{AB} = \eth^2 j$, the data (B1) leads to (4.2).

The radial integrals of the hypersurface equations[14] give, for the data (B1),

$$\beta = -(\lambda^2 l^2/32)c^{AB}c_{AB} + O(l^4) + \lambda^5 O(l^3\ln l)\,, \tag{B2}$$

$$U_A = -(r^2/2)c_{AB}{}^{:B} + (l^3/3)N_A + O(l^4) + \lambda^3 O(l^4\ln l)\,, \tag{B3}$$

$$W = -2M + O(l) + \lambda^3 O(l\ln l)\,, \tag{B4}$$

where M and N_A are the mass and angular momentum aspects, respectively. The evolution equation and supplementary conditions give

$$q^Aq^B(l^{-1}\gamma_{AB,v})_{,l} = -(l/3)q^Aq^BN_{A:B} + O(l^2) + l^3O(l^2\ln l)\,, \tag{B5}$$

$$M_{,v} = \tfrac{1}{4}c^{AB}{}_{,v:AB} - (\lambda^2/8)c^{AB}{}_{,v}c_{AB,v}\,, \tag{B6}$$

$$N_{A,v} = -2M_{,A} - (c_{BD}{}^{:D}{}_A - c_{AD}{}^{:D}{}_B)^{:B} + \lambda^2\left[-\tfrac{5}{16}(c^{DE}c_{DE})_{,vA} - \tfrac{1}{2}c^{BD}{}_{:A}c_{BD,v} + c_{AB}c^{BD}{}_{,v:D} - c^{BD}c_{AD,v:B}\right]. \tag{B7}$$

After setting $\eth\eta = \sqrt{2}q^AN_A$ and rewriting in terms of spin-weighted quantities, substitution of the null data (B1) into the evolution equation (B5) leads to $j_{,v} = 0$ and (4.3). Similarly, the supplementary conditions (B6) and (B7) lead to (4.4) and (4.5).

[1] Y. C. Bruhat and J. York, *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. 1, p. 23; A. E. Fischer and J. E. Marsden, *General Relativity*, edited by S. W. Hawking and W. Israel (Cambridge U. P., Cambridge, 1979), p. 138.

[2] H. Friedrich, Commun. Math. Phys. **103**, 35 (1986); H. Müller zum Hagen and H. J. Seifert, Gen. Relativ. Gravit. **8**, 259 (1977).

[3] H. Freidrich and J. Stewart, Proc. R. Soc. London Ser. A **385**, 345 (1983).

[4] M. Walker and C. M. Will, Phys. Rev. Lett. **45**, 1741 (1980); J. L. Anderson, *ibid.* **45**, 1745 (1980); T. Damour, *ibid.* **51**, 1019 (1983); F. I. Cooperstock and P. H. Lim, *ibid.* **55**, 265 (1985); T. Futamase and B. F.

Schutz, Phys. Rev. D **32**, 2557 (1985); T. Persides, "The gravitational field in the wave zone. II. Consequences of the asymptotic structure," Gen. Relativ. Gravit. (to appear).
[5]J. Winicour, J. Math. Phys. **24**, 1193 (1983).
[6]J. Winicour, J. Math. Phys. **25**, 2506 (1984).
[7]J. Winicour, "The quadrupole radiation formula," Gen. Relativ. Gravit. (to appear).
[8]R. A. Isaacson, J. S. Welling, and J. Winicour, J. Math. Phys. **26**, 2859 (1985).
[9]H. Bondi, M. G. J. van der Burgh, and A. W. K. Metzner, Proc. Roy. Soc. Ser. A **269**, 21 (1962); R. K. Sachs, *ibid.* **270**, 103 (1962); E. T. Newman and T. W. J. Unti, J. Math. Phys. **3**, 891 (1962).
[10]R. A. Isaacson, J. S. Welling, and J. Winicour, J. Math. Phys. **24**, 1193 (1983).
[11]J. Ehlers, *Grundlagen Probleme der Physik*, edited by J. Nitsch *et al.* (Bl Hochschultaschenbücher, Mannheim, 1981), p. 65.
[12]J. Winicour, Found. Phys. **15**, 605 (1985).
[13]R. Penrose, Proc. R. Soc. London Ser. A **284**, 159 (1965).
[14]L. Tamburino and J. Winicour, Phys. Rev. **150**, 1039 (1966).
[15]T. Damour (private communication).
[16]S. Persides, J. Phys. A: Math. Gen. **19**, 485 (1986).

On classical action at a distance theories which contain a cutoff

G. J. H. Burgers[a)]
Instituut Lorentz, Postbus 9506, 2300 RA Leiden, The Netherlands

H. Van Dam[b)]
Instituut Voor Theoretische Fysica, Rijks Universiteit Utrecht, P. O. Box 80.006, 3508 TA Utrecht, The Netherlands

(Received 28 August 1986; accepted for publication 12 November 1986)

It is pointed out that a certain class of classical relativistic action at a distance theories contains a cutoff which eliminates self-interaction of the particles. This cutoff is put in by hand, but one might hope that eventually it may be produced by considerations of space-time in the small. This class of theories is extended from models that closely mimic classical electrodynamics to models that resemble Yang–Mills and gravity.

I. INTRODUCTION AND SUMMARY

Classical relativistic theories that describe the interaction between point particles have been discussed by several authors.[1] Here we shall concentrate on those theories which are of the type where the interaction is given by manifestly covariant integrodifferential equations.[1–10]

First, in Sec. II, we discuss the familiar vector case, which, loosely speaking, corresponds to the exchange of vector particles in field theory. This case closely mimics electrodynamics, as is familiar from the work of Tetrode and Fokker[2] and Wheeler and Feynman.[3] We show that from our point of view[5] a cutoff at short distances, which eliminates self-interactions, is contained in a natural way. The precise details of the cutoff are put in by hand (for instance that it is related to the Planck length); it would be desirable to derive these properties from those of space-time in the small. The cutoff may be described briefly in terms of classical field theory by saying that the particles react to the field locally, but do not produce the field locally.

Section III contains a model that mimics the Yang–Mills theory. This model is mainly used to facilitate the discussion of the model of gravity of Sec. IV, which eventually turns out to be simpler. Just like the model of Sec. II, the Yang–Mills model is defined in Minkowski space, and it also contains a similar cutoff. The difference with the model of Sec. II is in the "quanta" that are being exchanged. In the model of Sec. II the quanta are rather different from the particles, thus one has an "action at a distance" theory of interactions between particles. The quanta of Sec. III still differ from particles: the particles have infinite timelike world lines, whereas the quanta have finite spacelike world lines and carry infinitesimal four-momentum. The equations for those world lines are, however, rather similar and both particles and quanta exchange quanta. Thus instead of an action at a distance theory one obtains a model where "particles" are exchanged.

Section IV contains a proposal for a model of gravity which has some similarity to the model of Sec. III. It differs from the model of Ref. 8 in several ways, one of which is that it contains a cutoff. The difference with general relativity is small for large distances, but quite striking at short distances. The cutoff is natural at the Planck length now. Particles within that length do not interact. In particular there is no self-interaction for particles. Thus one expects no real singularities, no infinite tidal forces as occur in general relativity (making classical relativity inconsistent within itself). The model of Sec. IV is therefore perhaps of interest in the study of black holes. Also one wonders, in view of superstring theory, which claims it provides a finite quantum theory of gravity,[11] what the classical limit of that theory might be. The quanta of Sec. IV, just like those of Sec. III, take on more and more particle properties. One might say that instead of an action at a distance model one has a particle model.

II. VECTOR INTERACTION

We begin by giving the familiar classical action which describes the electromagnetic field in interaction with a relativistic point particle. The action is[12]

$$S = -m\int ds\left[\eta_{\mu\nu}\frac{dx^\mu}{ds}\frac{dx^\nu}{ds}\right]^{1/2} - \frac{1}{4}\int dx\, F^{\mu\nu}F_{\mu\nu} - e\int ds\, A_\mu(x)\frac{dx^\mu}{ds}. \tag{2.1}$$

Here $\eta_{\mu\nu}$ is the metric of Minkowski space with diagonal elements $+1, -1, -1, -1$;

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{2.2}$$

The parameter s is an arbitrary but monotonously increasing parameter along the world line of the particle. Notice that the action (2.1) is invariant for chronometric transformations[13]

$$s' = g(s), \tag{2.3}$$

as well as for gauge transformations

$$A_\mu \to A_\mu + \partial_\mu \Lambda. \tag{2.4}$$

The latter transformations lead to

$$\delta S = -\int ds\, \partial_\mu \Lambda \frac{dx^\mu}{ds} = -\int ds \frac{d}{ds}\Lambda(x(s)) = 0,$$

where the last equal sign follows if one assumes $\Lambda(x)$ to have finite support.

The equations of motion are familiar,

[a)] Present address: CERN, CH-1211, Geneva 23, Switzerland.
[b)] On leave from the Department of Physics and Astronomy, University of North Carolina, Chapel Hill, North Carolina 27514.

 0022-2488/87/030677-08$02.50

$$m\ddot{x}^\mu = eF^{\mu\nu}\dot{x}_\nu, \tag{2.5a}$$

$$F_{\mu\nu,\sigma} + F_{\sigma\mu,\nu} + F_{\nu\sigma,\mu} = 0, \tag{2.5b}$$

$$F^{\mu\nu}{}_{,\nu} = j^\mu, \tag{2.5c}$$

with

$$j^\mu = e\int d\tau\, \delta^4(x - x(\tau))\dot{x}^\mu(\tau), \tag{2.5d}$$

where we choose s to be the proper time τ. In classical electrodynamics one limits oneself to retarded wave solutions of (2.5c).

The action at a distance approach[2–5] does not attempt to treat the electromagnetic field as an independent entity. Photons are not emitted but rather exchanged between particles or between particles and an "absorber."[3,7] This approach is remarkably successful although there are some questions about initial data to which we shall refer briefly at the end of this section.

One may write a formal action for the action at a distance theory which corresponds to (2.1).[6] Limiting oneself to two particles it is

$$S^1 = m_1\int ds_1\sqrt{\eta_{\mu\nu}\dot{x}_1^\mu\dot{x}_1^\nu} - m_2\int ds_2\sqrt{\eta_{\mu\nu}\dot{x}_2^\mu\dot{x}_2^\nu}$$

$$- 2e_1e_2\int ds_1\int ds_2\,\eta_{\mu\nu}\dot{x}_1^\mu\dot{x}_2^\nu f(\rho^2). \tag{2.6}$$

Here $\dot{x}_1^\mu$ stands for $(d/ds_1)x_1^\mu(s_1)$;

$$\rho^2 = (x_1 - x_2)^\mu(x_1 - x_2)^\nu\eta_{\mu\nu}; \tag{2.7}$$

the function f, which has dimension $(\text{length})^{-2}$, is arbitrary except that we take it as nonzero only for spacelike values of $(x_1 - x_2)$, i.e., for $\rho^2 < 0$.

What is to be varied in (2.6) consists of the two world lines $x_1^\mu(s_1)$ and $x_2^\mu(s_2)$ of the particles. If $f(\rho^2) = 0$ except for $\rho^2 < 0$, then there is no self-interaction as we shall discuss shortly, and the equations of motion may be written with the aid of the vector field

$$A^\mu(x) = e_1\int ds_1\,\dot{x}_1^\mu(s_1)\,f((x - x_1(s_1))^2)$$

$$+ e_2\int ds_2\,\dot{x}_2^\mu(s_2)\,f((x - x_2(s_2))^2). \tag{2.8}$$

Using (2.8) and using for s_1 and s_2 the proper times τ_1 and τ_2 one obtains the equations

$$m_1\ddot{x}_1^\mu = e_1F^{\mu\nu}(x_1)\dot{x}_{1\nu}, \tag{2.9a}$$

$$m_2\ddot{x}_2^\mu = e_2F^{\mu\nu}(x_2)\dot{x}_{2\nu}, \tag{2.9a'}$$

$$F_{\mu\nu,\sigma} + F_{\sigma\mu,\nu} + F_{\nu\sigma,\mu} = 0. \tag{2.9b}$$

Comparing (2.9a), (2.9b), (2.8) with (2.5a), (2.5b), (2.5c), (2.5d) we see that the two first pairs agree. The particles in (2.9a) and (2.9b) react locally to the field just as in (2.5a) and (2.5b). However, (2.8) replaces (2.5c) and (2.5d); the particles no longer produce the field locally, as we shall discuss.

Returning to the function $f(\rho^2)$ in (2.6) we will assume that it has a cutoff at short distances D:

$$f(\rho^2) = 0, \quad \text{for } \rho^2 > -D^2. \tag{2.10}$$

There is an additional condition on f which comes from the demand that as the particles separate the forces approach zero.[5] It is

$$\int d\rho^2\,\frac{d}{d\rho^2}f(\rho^2) = 0. \tag{2.11}$$

Within the conditions (2.10) and (2.11) the function f is arbitrary. In fact there is a mapping between the function f and the nonrelativistic potential function to which it leads.[6]

Besides including a rather arbitrary function f the type of interaction [(2.10) and (2.11)] has two advantages which we illustrate for

$$f(\rho^2) = \delta(\rho^2 + D^2). \tag{2.12}$$

The first advantage is obvious from Fig. 1. The world line x_2 is closer to $x_1(\tau)$ than D; therefore, there is no interaction (at least for a segment of x_1 around τ_1), and we have "asymptotic freedom." In particular there is no self-interaction and one may write (2.8) for all x. Wheeler–Feynman electrodynamics[3] also has no self-interaction, but it is by definition, not built in as with (2.12). In the case of Wheeler–Feynman electrodynamics this lack of self-interaction leads to trouble when quantizing and kinking a world line backwards to represent a pair; one obtains a noninteracting electron–position pair.[14]

The Fokker–Tetrode action which leads to Wheeler–Feynman electrodynamics is obtained from (2.6) by setting

$$f(\rho^2) = \delta(\rho^2). \tag{2.13}$$

The equation of motion may not be written as (2.9) with (2.8), as the self-field must be explicitly excluded. This self-field is excluded in (2.6), but not in (2.8) and (2.9), which may only be written provided (2.10) is satisfied. The choice (2.13) thus leads to Wheeler–Feynman (WF) electrodynamics which is a form of electrodynamics with half-retarded half-advanced Green's functions. This half-advanced half-retarded electrodynamics turns into the observed retarded form of electrodynamics if one puts in an absorber.[3,7] This elegant device ascribes the appearance of retarded Green's functions to the increase of entropy in the large.

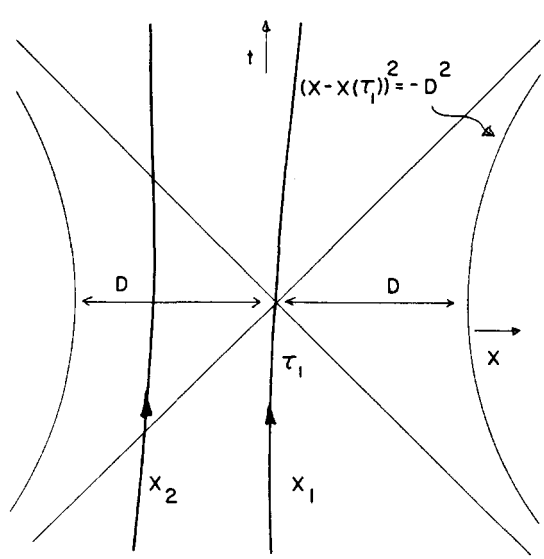


FIG. 1. No interaction for world lines close together.

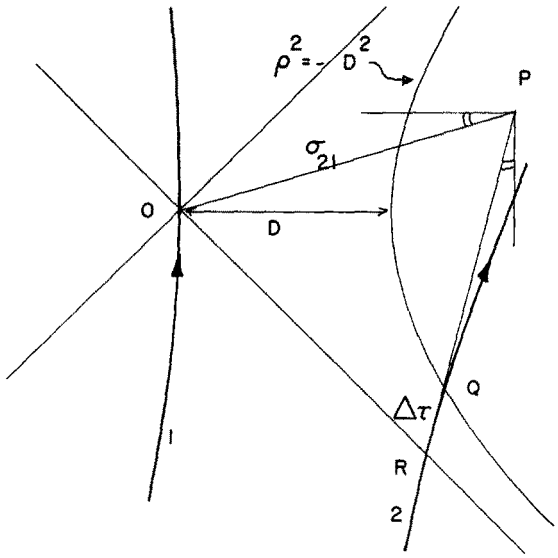


FIG. 2. For large distance the interaction approaches the WF interaction.

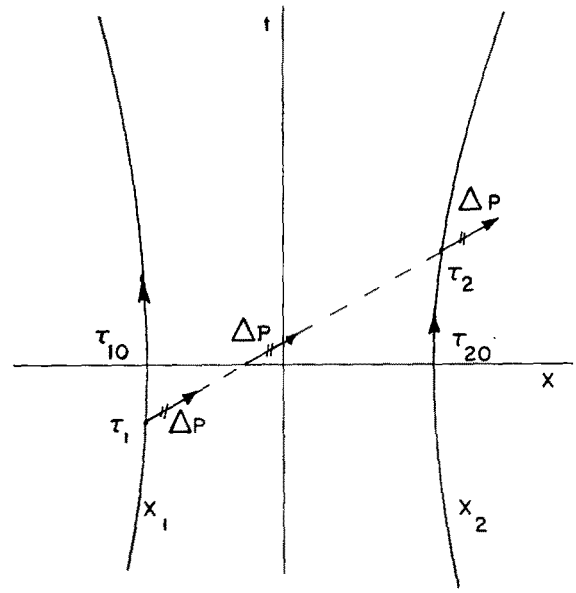


FIG. 3. Exchange of linear four-momentum between world lines.

Notice that f in (2.6) and (2.8) is symmetric and classical electrodynamics with $f=\delta(\rho^2)\theta(x^0)$ is not possible. However, if one is willing to give up the action principle one might proceed with this choice in (2.8).

The second advantage has to do with the fact that obviously (2.12) is close to (2.13) and thereupon to electrodynamics. For distances σ large compared to D of (2.12), the interaction rapidly approaches the WF interaction (2.13). In Fig. 2 an estimate is made for the length of proper time $\Delta\tau$ between passing through the lightcone based on 0 and through the spatial hyperboloid $\rho^2=-D^2$ for a particle with world line x_2 (Ref. 5). In the triangle OPR, $\mathrm{OP}=\sigma_{21}$ is the spatial distance of O as seen in the frame of x_2 as it crosses the advanced lightcone from O. We have $\mathrm{RP}=\mathrm{OP}=\sigma_{21}$, where RP is along the time axis of that frame. We assume $\Delta\tau$ is so short that we may measure it along the straight RQ. Then one has

$$\mathrm{PQ}^2-\mathrm{OP}^2=-D^2$$

or

$$(\mathrm{OP}-\Delta\tau)^2-\mathrm{OP}^2=-D^2,$$

hence

$$2\sigma_{21}\Delta\tau=D^2+\Delta\tau^2$$

or

$$\Delta\tau\sim D^2/2\sigma_{21}. \tag{2.14}$$

Taking D small, say the Planck length, it is obvious that $\Delta\tau$ approaches zero rapidly as σ_{21} increases, justifying the assumption made in establishing (2.14).

Next we wish to discuss the conservation laws of linear and angular momentum in a way similar to that of Refs. 5 and 6. We shall need these results in the next sections. Equations (2.9) may be written

$$\dot{p}_1^\mu=2e_1e_2\int d\tau_2\,\dot{x}_1\cdot\dot{x}_2(x_1-x_2)^\mu f', \tag{2.15a}$$

$$\dot{p}_2^\mu=2e_1e_2\int d\tau_1\,\dot{x}_1\cdot\dot{x}_2(x_2-x_1)^\mu f', \tag{2.15b}$$

where $f'(\rho^2)=(d/d\rho^2)f(\rho^2)$, and where

$$p_1^\mu=m\dot{x}_1^\mu+2e_1e_2\int d\tau_2 f\dot{x}_2^\mu, \tag{2.16a}$$

$$p_2^\mu=m\dot{x}_2^\mu+2e_1e_2\int d\tau_1 f\dot{x}_1^\mu. \tag{2.16b}$$

Thus, as illustrated in Fig. 3, the interaction may be seen as an exchange of four-momentum between pairs of spacelike events $x_1(\tau_1)$, $x_2(\tau_2)$. With (2.15) the amount of four-momentum exchanged is

$$\Delta p^\mu=-2(x_1-x_2)^\mu e_1e_2\,d\tau_1\,d\tau_2(\dot{x}_1\cdot x_2)f'. \tag{2.17}$$

As this four-momentum is in the direction of x_1-x_2, there will be conservation of six angular momentum as well as of four-momentum. This is quite analogous to Newtonian mechanics, which may be described by the (instantaneous) transfer of linear momentum between points in Euclidean space. If the exchange in Newton's theory has the direction of the vector between the points, one has three angular momentum conservation laws as well as three linear momentum conservation laws. The difference between the relativistic case and the Newtonian one is twofold. First the relativistic case has four more conservation laws. Second, however, the exchange is not instantaneous. Thus, in Fig. 3, when computing the total linear (or angular) momentum of the system at $t=0$ (intersecting the orbits at τ_{10} and τ_{20}) one must take into account the linear (or angular) momentum in transit at that time. For instance, the total linear momentum at $t=0$ (given by τ_{10} and τ_{20}) is[5,6]

$$\begin{aligned}P^\mu(\tau_{10},\tau_{20})=&p_1^\mu(\tau_{10})+p_2^\mu(\tau_{20})\\&-2\left(\int_{-\infty}^{\tau_{10}}d\tau_1\int_{\tau_{20}}^{\infty}d\tau_2-\int_{\tau_{10}}^{\infty}d\tau_1\int_{-\infty}^{\tau_{20}}d\tau_2\right)\\&\cdot e_1e_2\dot{x}_1^\alpha(\tau_1)\dot{x}_2(\tau_2)_\alpha f'(x_1^\mu(\tau_1)-x_2^\mu(\tau_2)).\end{aligned} \tag{2.18}$$

With (2.17) one may check that $P^\mu(\tau_{10},\tau_{20})$ is conserved as

$$\frac{d}{d\tau_{10}}P^\mu(\tau_{10},\tau_{20})=\frac{d}{d\tau_{20}}P^\mu(\tau_{10},\tau_{20})=0. \tag{2.19}$$

Before summarizing this section we shall briefly mention two topics: causality and initial conditions.

If one defines causality to be the condition that small disturbances do not propagate with a speed faster than light, then the present proposal looks acausal. However, at first sight also Wheeler–Feynman electrodynamics looks acausal; actually it is causal with suitable boundary conditions (absorber).[3,7] For (2.12) an absorber can be introduced and then for D small one would expect that the acausality will be hard to observe in a realistic case.

As for initial condition (2.8), (2.9) suggest a perturbative solution starting with two straight world lines. For a limited class of functions f one can show that this perturbation series converges, suggesting that the appropriate initial conditions for these equations for N particles are the $6N$ spatial positions and velocities of the particles at one instant of time. Wheeler–Feynman electrodynamics has an f outside this class of functions mentioned. The situation is thus less clear and it has often been argued that one needs an infinite set of initial conditions at one instant of time[3,4,15] (representing the degrees of freedom of the suppressed A_μ field).

To summarize, let us assume f to satisfy (2.10); for instance let it be given by (2.12) with D small. Then, Eqs. (2.9) show particles which react locally to a field which resembles the electromagnetic field at distances large compared to D [see (2.14)]. Equation (2.8) shows that the field is not produced locally. The field (2.8) is zero along the world line of the particle which produces it. Actually it is zero inside a tube of radius D around the world line of that particle. It is as if in producing the field the charge has been pushed out a distance D, whereas in reacting to the field it is centered on the world line of the particle.

III. "YANG–MILLS" INTERACTION

First, let us review the classical action describing a relativistic point particle carrying an isotopic spin $I^a(s)$ along its world line $x^\mu(s)$ in interaction with a Yang–Mills field A^a_μ. It is given by[16]

$$S_Y = -m\int ds\left[\eta_{\mu\nu}\frac{dx^\mu}{ds}\frac{dx^\nu}{ds}\right]^{1/2} - \int dx\,\frac{1}{4g^2}G^a_{\mu\nu}G^{a\mu\nu} - \int ds\,\frac{dx^\mu}{ds}(s)I^a(s)A^a_\mu(x(s)). \tag{3.1a}$$

Here

$$G^a_{\mu\nu} = gF^a_{\mu\nu} - g^2\epsilon^{abc}A^b_\mu A^c_\nu, \tag{3.1b}$$

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu. \tag{3.1c}$$

The action (3.1a) is invariant under gauge transformation of the Yang–Mills field A^a_μ

$$A^a_\mu \to A^a_\mu + \partial_\mu\Lambda^a + g\epsilon^{abc}A^b_\mu\Lambda^c, \tag{3.2}$$

provided

$$0 = \int ds\,\frac{dx^\mu}{ds}I^a(\partial_\mu\Lambda^a + g\epsilon^{abc}A^b_\mu\Lambda^c)$$

or

$$0 = \int ds\,I^a\frac{d}{ds}\Lambda^a + \int ds\,\frac{dx^\mu}{ds}I^a g\epsilon^{abc}A^b_\mu\Lambda^c.$$

Taking the function Λ^a_μ of finite support and otherwise arbitrary, this implies the constraint

$$\frac{d}{ds}I^a(s) = g\epsilon^{abc}I^b(s)A^c_\mu(x(s))\dot{x}^\mu(s). \tag{3.3}$$

In other words the equation of motion for the isotopic spin needs no separate derivation. The other equations of motion are, choosing for s the proper time τ,

$$m\ddot{x}^\mu = G^{a\mu\nu}\dot{x}_\nu I^a, \tag{3.4}$$

$$D^{ab}_\mu G^{b\mu\nu}(x) = \int d\tau\,\delta(x - x(\tau))\dot{x}^\mu(\tau)I^a(\tau), \tag{3.5}$$

where

$$D^{ab}_\mu G^{b\mu\nu} = \partial_\mu G^{a\mu\nu} + g\epsilon^{abc}A^b_\mu G^{c\mu\nu}. \tag{3.6}$$

The fact that D^{ca} on the left-hand side of (3.5) gives zero again implies the constraint (3.3).

Next we attempt to set up a somewhat similar action at a distance theory. First, we set up an exchange of linear momentum and isotopic spin which mimics the structure of the preceding section with the additional global conservation law of isospin. In analogy with (2.17) and Fig. 3 one would suggest as the simplest exchange which conserves total isospin as well as the length of each isospin:

$$\Delta p^\mu = -2(x_1 - x_2)^\mu g^2\dot{x}^\alpha_1\dot{x}_{2\alpha}I^a_1I^a_2f'(\rho^2)d\tau_1\,d\tau_2, \tag{3.7}$$

$$\Delta I^a = g^2\dot{x}^\alpha_1\dot{x}_{2\alpha}I^b_2I^c_1\epsilon^{abc}f(\rho^2)d\tau_1\,d\tau_2. \tag{3.8}$$

The conservation laws of linear four- and angular six-momentum follow from an argument similar to that given in Sec. II near Eq. (2.18).

Using proper times the equations of motion for the particles are

$$m_1\ddot{x}^\mu_1 = gF^{a\mu\nu}(x_1(\tau))I^a(\tau_1)\dot{x}_{1\nu}(\tau_1), \tag{3.9}$$

$$\dot{I}^a(\tau_1) = g\epsilon^{abc}A^c_\mu\dot{x}^\mu_1I^b_1(\tau_1), \tag{3.10}$$

where

$$A^b_\mu(x) = g\int ds_1\,\dot{x}_{1\mu}(\tau_1)I^b_1(\tau_1)f((x - x_1)^2) + g\int ds_2\,\dot{x}_{2\mu}(\tau_2)I^b_2(\tau_2)f((x - x_2)^2), \tag{3.11}$$

the equations for $\ddot{x}_2$ and I^a_2 being similar to (3.9).

Equation (3.10) is identical to (3.3), and (3.9) lacks the A^2 term of (3.4). As there is conservation of the quantities of isospin and linear and angular momentum one might stop at this point. The equations lack the nice geometric structure of the equations that follow from (3.1).

Equations (3.7)–(3.11) imply a sharp difference between particles and quanta. Along the quantum lines Δp^μ and ΔI^a are transported without change. Along the world lines of the particles the linear momentum $m\dot{x}^\mu$ changes, keeping its length, and the isospin is parallel transported with a connection

$$\epsilon^{abc}A^c_\mu(x). \tag{3.12}$$

It is tempting to remove this particular difference between

particles and quanta. The removal of the difference leads to a proliferation of quanta, as we shall see.

An equivalent approach is to restore the invariance for local gauge transformation (3.2). This implies replacing $F^{a\mu\nu}$ in (3.9) with the properly transforming $G^{a\mu\nu}$ of (3.1a);

$$m_1\ddot{x}_1^\mu = gG^{a\mu\nu}(x_1(\tau_1))I^a(\tau_1)\dot{x}_{1\nu}(\tau_1), \tag{3.13}$$

as well as demanding that the quanta also transport linear momentum and isospin as in (3.13), (3.10). The quanta have spacelike world lines, which may be parametrized by proper length. One has

$$(\dot{I}_Q^a) = g\epsilon^{abc}A_\mu^c(x_Q)I_Q^b\dot{x}_{Q\mu}, \tag{3.14}$$

$$m_Q\ddot{x}_Q^\mu = gG^{a\mu\nu}(x_Q)\dot{x}_{Q\nu}I_Q^a. \tag{3.15}$$

Here m_Q is the inertia of the quantum and I_Q^a is the isospin which it carries:

$$m_Q = 2d(1,2)g^2(\dot{x}_1^\mu P_{\mu,\beta}(1,2)\dot{x}_2^\beta) \times(I_1^a P_{a,b}(1,2)I_2^b)f'(d^2(1,2))d\tau_1\,d\tau_2; \tag{3.16}$$

$$I_Q^a = g^2(\dot{x}_1^\mu P_{\mu,\beta}(1,2)\dot{x}_2^\beta)\epsilon^{abc}P_{c,d}(Q,1)I_1^d P_{b,e}(Q,2) \times I_2^e f(d^2(1,2))d\tau_1\,d\tau_2. \tag{3.17}$$

In the last pair of formulas $d(1,2)$ is the proper distance along the world line from $x_1(\tau_1)$ to $x_2(\tau_2)$; $P_{\mu,\beta}(1,2)$ stands for Fermi–Walker transport along that world line of the following $\dot{x}_2^\beta$; $P_{a,b}(Q,1)$ stands for parallel transport using $A_\mu^a(x)$ along the same world line from $x_1(\tau_1)$ to x_Q. As lowest order approximation one has $d^2(1,2) = (x_1 - x_2)^2$; $P_{a,b} = \delta_{ab}$; $P_{\mu,\beta} = \eta_{\mu\beta}$, describing the properties of the quantum (3.7) and (3.8). The generalization of (3.7) is

$$\Delta p^\mu = m_Q\dot{x}_Q^\mu + I_Q^a A^{a\mu}. \tag{3.7'}$$

But now, (3.14) and (3.15) imply that other quanta are being exchanged between the quanta and the particles, and between quanta and quanta, etc. The quanta react to and become sources of the field and one has a proliferation of quanta. Note that this proliferation does not happen for "electrodynamics" as the quanta do not carry charge; I_Q^a is then zero in (3.14) and (3.15).

Before proceeding with this let us point out another complication: $G^{a\mu\nu}$ of (3.14) contains an A^2 term [see (3.1a)]. This implies a simultaneous exchange of linear momentum between an event on one world line with two events on other world lines (triangular exchange). Does this violate the conservation laws? The extra contribution to $m_1\ddot{x}_1^\mu$ is

$$g^3\int d\tau_2\,d\tau_2'\,\dot{x}_2^\mu(\tau_2')\dot{x}_1\cdot\dot{x}_2\,I_1^aI_2^bI_{2'}^c\epsilon^{abc}f((1,2)^2)f((1,2')^2).$$

By symmetry this is perpendicular to $\dot{x}_1^\mu$ and via a partial integration over $d\tau_2'$ (which is allowed as the support of f is nonzero only outside the lightcone and as the world line 2 is timelike) this can be put in the form

$$\Delta' p^\mu = -g^3 d\tau_1\,d\tau_2\,d\tau_2'(x_1 - x_2')^\mu\dot{x}_1\cdot\dot{x}_2 f(1,2) \times I_2^bI_1^a\epsilon^{abc}\frac{d}{d\tau_2'}f(1,2')I_{2'}^c,$$

which shows that the ten kinematic conservation laws are satisfied.

For particles, the motion in the field $A_\mu^a(x)$ is described by (3.10), (3.13), and for quanta this motion is given by (3.14)–(3.17). What we need next is an expression for $A_\mu^a(x)$; we propose

$$A_\mu^a(x) = A_{\mu P}^a(x) + \sum_Q A_{\mu Q}^a(x). \tag{3.18}$$

Here $A_{\mu P}^a(x)$ is the particle contribution:

$$A_{\mu P}^a(x) = g\int d\tau_1\,P_{\mu,\nu}{}^{a,b}(x,1)\dot{x}_1^\nu(1)I^b(1)f(d^2(x,1)) + g\int d\tau_2\,P_{\mu,\nu}{}^{a,b}(x,2)\dot{x}_2^\nu(2)I^b(2)f(d^2(x,1)). \tag{3.19}$$

Here $P_{\mu,\nu}{}^{a,b}(x,1)$ combines two kinds of parallel transport along the curve from x_1 to x, described by (3.15). The first kind is Fermi–Walker transport of $\dot{x}_1^\nu(x_1)$ along that curve. The second kind is transport using the connection $A_\mu^a(x)$ of $I^a(x_1)$; $d^2(x,1)$ stands for the proper distance along the curve mentioned between x and x_1.

The quantum contribution is given by an infinite set of terms. To $\Sigma_Q A_{\mu Q}^a$ each quantum line gives a contribution similar to (3.19), where the integral is over the proper length of the quantum line and where $I^b(1)$ is replaced by I_Q^b. For the quantum line which connects the particle world lines $x_1(\tau_1)$ and $x_2(\tau_2)$ we use the symbol $[1,2]$; $\tau_{[1,2]}$ is the proper length along that world line. In a similar way we use $[1,12]$ for the quantum line connecting the particle line 1 with the quantum line $[1,2]$. Here $[12,12]$ will stand for a self-interaction of $[1,2]$ with $[1,2]$, etc. This is illustrated in Fig. 4. The contribution of $[1,2]$ to $\Sigma_Q A_{\mu Q}^a$ is

$$A_{\mu_{[1,2]}}^a = g\int d\tau_1\int d\tau_2\int d\tau_{[1,2]}P_{\mu\nu}{}^{a,b}(x,[1,2])\dot{x}_{[1,2]}^\nu \times f(d^2(x,[1,2])) \times g^2(\dot{x}_1^\alpha P_{\alpha,\beta}(1,2)\dot{x}^\beta(2))\cdot\epsilon^{abc}P_{c,d}([1,2],1) \times I_1^aP_{b,e}([1,2],2)I_2^e f(d^2(1,2)). \tag{3.20}$$

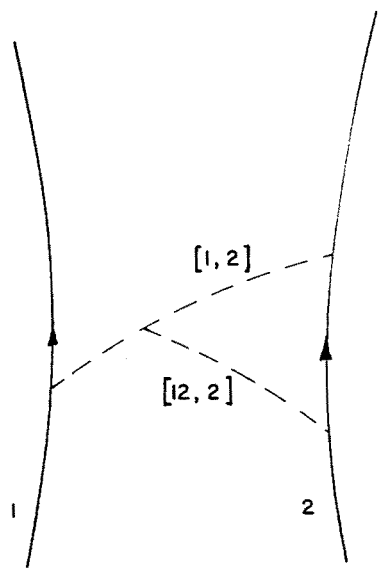


FIG. 4. Proliferation of interaction lines between world lines and interaction lines.

Here we used (3.17); it is straightforward to write any of the following terms, such as the contribution of [1,12]. The last term on the right-hand side of (3.18) stands for the infinite sum over all possible interaction quanta.

Equations (3.10), (3.13), (3.14)–(3.16), (3.18), (3.19), (3.20),...may be solved order by order starting with just two straight world lines for the particles 1 and 2. The question of initial conditions is, however, even more complicated than it was in Sec. II.

We have three remarks to end this section. First, the proliferation of quanta is kept somewhat under control since as the number of quanta considered grows, so does the order in g. Also, the conservation laws will be satisfied as they are at every stage of refinement.

Second, a cutoff at a distance D is included, just as in Sec. II through the function f, which is taken to satisfy (2.10), and for which one may take (2.12). We feel that this cutoff reflects some underlying property of space in the small, but have not made progress along that line.

Third, notice that the distinction between particles and quanta is disappearing. Both particles and quanta produce and react to the field which curves their world lines and parallel transports their isospin. There are still differences: the world lines of the particles are timelike and infinite in length, and the world lines of the quanta are spacelike and finite in length. The number of particles is finite. The number of quanta is infinite, and the amount of charge and linear momentum carried by the quanta is infinitesimal. Nevertheless, the proposal of this section is not really any longer an action at a distance theory, but rather a description in terms of particles, the quanta becoming particles.

IV. A PROPOSAL FOR A MODEL OF GRAVITY WITH A CUTOFF

Section II describes a vector interaction. A scalar interaction was given in Ref. 6, and it is not hard to write a general tensor interaction, which is a straightforward generalization of the vector interaction of Sec. II.[17] In writing this generalization, the important step is to put in the correct powers of $\dot{x}_1^2$ and $\dot{x}_2^2$ to guarantee chronometric invariance.[13,17] This chronometric invariance guarantees that one may use proper time for the description of the world lines and that using proper time $\dot{x}^\mu \ddot{x}_\mu = 0$.

To get a first approximation to a theory of gravity which comes as close as possible to the experimental results of Einstein's theory one must take what looks in the present framework like a mixture of symmetric two-tensor interaction and a scalar interaction. In field theory this particular mixture does not look like a mixture. It follows from unitarity and from the fact that the graviton has zero rest mass and helicity ± 2.[18,19] Thus, a first-order proposal that is close to the linearized theory of gravity is given by the action

$$S = -m_1 \int ds_1 (\dot{x}_1^2)^{1/2} - m_2 \int ds_2 (\dot{x}_2^2)^{1/2} - 2m_1 m_2 G \times \int ds_1 \int ds_2 \frac{[(\dot{x}_1^\alpha \dot{x}_2^\beta \eta_{\alpha\beta})^2 - \frac{1}{2}\dot{x}_1^2 \dot{x}_2^2]}{\sqrt{\dot{x}_1^2 \dot{x}_2^2}} f((x_1 - x_2)^2), \tag{4.1}$$

where all $\dot{x}^2$ are defined with $\eta_{\mu\nu}$.

As G has the dimension of (length)^2 and as f has dimension (length)$^{-2}$, it is tempting to take for Gf something like

$$Gf((x_1 - x_2)^2) = G(\delta((x_1 - x_2)^2 + G)). \tag{4.2}$$

In any case, f should satisfy (2.10).

In analogy with (2.15) the equations of motion may be written as

$$\dot{p}_1^\mu = 2m_1 m_2 G \int ds_2 \left\{ (\dot{x}_1 \cdot \dot{x}_2)^2 - \frac{1}{2}\dot{x}_1^2 \dot{x}_2^2 \right\} (x_1 - x_2)^\mu f', \tag{4.3a}$$

$$\dot{p}_2^\mu = 2m_1 m_2 G \int ds_1 \left\{ (\dot{x}_1 \cdot \dot{x}_2)^2 - \frac{1}{2}\dot{x}_1^2 \dot{x}_2^2 \right\} (x_2 - x_1)^\mu f', \tag{4.3b}$$

with the obvious definitions of p_1 and p_2. The proof of the conservation laws is a repetition of that of Sec. II. As there is no self-interaction due to the cutoff in (4.2) one may introduce a field $h_{\mu\nu}(x)$ for all x by

$$h_{\mu\nu}(x) = 2m_2 G \int ds_2 \frac{(\dot{x}_{2\mu}\dot{x}_{2\nu} - \frac{1}{2}\eta_{\mu\nu}\dot{x}_2^2)}{(\dot{x}_2^2)^{1/2}} f((x - x_2)^2) + 2m_1 G \int ds_1 \frac{(\dot{x}_{1\mu}\dot{x}_{2\nu} - \frac{1}{2}\eta_{\mu\nu}\dot{x}_1^2)}{(\dot{x}_1^2)^{1/2}} f((x - x_1)^2). \tag{4.4}$$

With this one may write the equations of motion for the particles as

$$\ddot{x}_1^\mu = \Gamma_{\alpha\beta}^\mu(x_1)\dot{x}_1^\alpha \dot{x}_1^\beta, \tag{4.5a}$$

$$\ddot{x}_2^\mu = \Gamma_{\alpha\beta}^\mu(x_2)\dot{x}_2^\alpha \dot{x}_2^\beta, \tag{4.5b}$$

where we must choose for the arbitrary parameters s_1 and s_2 the proper times of the two particles, and where $\Gamma_{\alpha\beta}^\mu(x)$ is formed in the usual way from first derivatives of $h_{\mu\nu}(x)$.

If one replaces the function f, which is defined by (4.2) by $\delta((x_1 - x_2)^2)$, then one obtains, except for self-interaction, the half-advanced half-retarded version of the usual linearized theory. As the tests of general relativity (bending of light, gravitational red shift, delay of radar echoes, perihelion precession) involve time translation invariant situations, it makes no difference for these predictions to have half-advanced half-retarded Green's functions. Furthermore, with the arguments given in Sec. II near Eq. (2.14), it is obvious that with (4.1) and (4.2) one obtains the usual results of the linearized theory.

Thus, the bending of light comes out right (approximating light by a very fast point particle). The perihelion shift comes out $\frac{1}{6}$ too large[20]; in the linearized approach one may correct this by including the interaction of the planet with the gravitational field between planet and star.[20] This correction is not yet contained in (4.1), where one only has an interaction between particles, not between particles and quanta. The required $1/r^2$ potential could be put in by hand via f. A similar problem is that in the present model potential gravitational energy is represented as energy on its way as in Fig. 3. A third particle interacts directly with two constituents of a bound system, but not with this energy on its way. Thus there is a violation of inertial mass = gravitational mass. This leads to trouble; as in stars the contributions of

the potential energy can be considerable indeed.

The gravitational red shift is predicted correctly,[20] but the time translation invariance of Minkowski space leads to a familiar contradiction.[21] Either Planck's constant h becomes dependent on the gravitational field or one must include $h_{\mu\nu}$ and $g_{\mu\nu}$ together in a metric that changes the proper time. The radar delay from interior planets may also be explained,[20] but again suggests inclusion of $h_{\mu\nu}$ into the metric.

The solution to these problems appears to be the following.

For the action describing the world line of a point particle in a given Riemannian manifold one has

$$S = -\int m\,ds\left[g_{\mu\nu}(x(s))\frac{dx^\mu}{ds}\frac{dx^\nu}{ds}\right]^{1/2}, \tag{4.6}$$

where $g_{\mu\nu}(x)$ is the given metric. One obtains, writing $\dot{x}^\mu$ for $(d/ds)x^\mu(s)$,

$$(g_{\mu\nu}\dot{x}^2 - g_{\mu\alpha}g_{\beta\nu}\dot{x}^\alpha\dot{x}^\beta)\ddot{x}^\nu = \tfrac{1}{2}g_{\alpha\beta,\lambda}\dot{x}^\alpha\dot{x}^\beta\dot{x}^\lambda g_{\mu\nu}\dot{x}^\nu + \dot{x}_1^2(g_{\mu\nu,\alpha}\dot{x}^\alpha\dot{x}^\nu - \tfrac{1}{2}g_{\alpha\beta,\mu}\dot{x}^\alpha\dot{x}^\beta). \tag{4.7}$$

This equation reflects the chronometric invariance of (4.6): $\dot{x}^\mu$ into either the left-hand side or right-hand side of (4.7) gives zero. Choosing the gauge of proper time, $(d/d\tau)g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0$, (4.7) takes as the familiar form

$$\ddot{x}^\mu = \Gamma^\mu_{\alpha\beta}\dot{x}^\alpha\dot{x}^\beta, \tag{4.8}$$

with

$$\Gamma^\mu_{\alpha\beta} = \tfrac{1}{2}g^{\mu\nu}\{g_{\nu\beta,\alpha} + g_{\nu\alpha,\beta} - g_{\alpha\beta,\nu}\} \tag{4.9}$$

the connection of Riemannian geometry.

We shall assume (4.8), (4.9), or (4.7) not only for the world lines of the particles, but also for the spacelike world lines of the quanta. The metric is determined by all these world lines in analogy to the Yang–Mills potentials A^a_μ of Sec. III. The form (4.8) may be maintained for the quantum lines by using proper length instead of proper time. We shall use the symbol τ also along those lines. Notice, however, that $\dot{x}^2 = -1$ along quantum lines. The expression for $g_{\mu\nu}(x)$ contains three parts: the first part is $\eta_{\mu\nu}$, the second part $g_{\mu\nu P}$ is due to the particles, and the third part $g_{\mu\nu Q}$ due to the quanta

$$g_{\mu\nu}(x) = \eta_{\mu\nu} + g_{\mu\nu P} + \sum_Q g_{\mu\nu Q}, \tag{4.10}$$

$$g_{\mu\nu P}(x) = 2\int d\tau_1\, P_{\mu\lambda,\alpha\beta}(x,1)\left(\dot{x}_1^\alpha\dot{x}_1^\beta - \frac{1}{2}g^{\alpha\beta}\right) \times m_1 G f(d^2(x,1)) + 2\int d\tau_2\, P_{\mu\nu,\alpha\beta}(x,2)\left(\dot{x}_2^\alpha\dot{x}_2^\beta - \frac{1}{2}g^{\alpha\beta}\right) \times m_2 G f(d^2(x,2)), \tag{4.11}$$

where G is Newton's constant; $d^2(x,1)$ is the proper distance along the geodesic of the metric (4.10) which quanta follow from x_1 to x_2; $P_{\mu\nu,\alpha\beta}(x,1)$ and $P_{\mu,\alpha}(x,1)$ refer to Fermi–Walker transport along this geodesic.[22] (This is for a geodesic equivalent to covariant transport.) If there is more than one geodesic between x_1 and x then contributions from various geodesics are added.

For a general quantum line x_Q one may write, similarly,

$$g_{\mu\nu Q}(x) = 2\int d\tau_Q\, P_{\mu\nu,\alpha\beta}(x,Q)\left(\dot{x}_Q^\alpha\dot{x}_Q^\beta + \frac{1}{2}g^{\alpha\beta}\right) \times m_Q G f(d^2(x,Q)), \tag{4.12}$$

where the $+\frac{1}{2}g^{\alpha\beta}$ in the term on the right follows from (4.4) and differs in sign from (4.11) as the world lines of the quanta are spacelike. The m_Q is the inertia carried by the world line of the quantum considered. Using the labels 1 and 2 for the world lines $x_1^\mu(\tau_1)$ and $x_2^\mu(\tau_2)$, [1,2] is short for the quantum line connecting 1 and 2. Similarly, as illustrated in Fig. 4, [1,12] is a quantum line connecting [1,2] and 1, etc. For the world line [1,2], m_Q is given by

$$m_{[1,2]} = d\tau_1\, d\tau_2\, m_1 m_2 G f'(d^2(1,2)) \times\left[(\dot{x}_1^\alpha P_{\alpha,\beta}(1,2)\dot{x}_2^\beta)^2 - \tfrac{1}{2}\right] d(1,2), \tag{4.13}$$

which follows from a generalization of (4.3), with $P_{\alpha,\beta}(1,2)$ standing for Fermi–Walker (or parallel) transport along the geodesic [1,2]. The quantum lines of type [1,2] give a total contribution to $\Sigma_Q g_{\mu\nu Q}$ of

$$g_{\mu\nu[1,2]} = \int d\tau_1 \int d\tau_2 \int d\tau_{[1,2]}\, P_{\mu\nu,\alpha\beta}(x,[12])(\dot{x}^\alpha_{[12]}\dot{x}^\beta_{[12]} + g^{\alpha\beta})Gf(d^2(x,[12]))\cdot m_1 m_2 G f'(d^2(1,2)) \times\left[(\dot{x}_1^\alpha P_{\alpha,\beta}(1,2)\dot{x}_2^\beta)^2 - \frac{1}{2}\right] d(1,2). \tag{4.14}$$

For a quantum line [1,1'2],

$$m_{[1,1'2]} = d\tau_1\, d\tau_{[12]}\, m_1\, d\tau_1'\, d\tau_2\, m_1 m_2 g f'(d^2(1,2)) \times\left[(\dot{x}_1^\alpha P_{\alpha,\beta}(1',2)\dot{x}_2^\beta)^2 - \tfrac{1}{2}\right] \times d(1',2) G f'(d^2(1,[1'2])) \times\left[(\dot{x}_1^\alpha P_{\alpha,\beta}(1,[1'2])\dot{x}^\beta_{[12]})^2 - \tfrac{1}{2}\right] d(1,[1',2]),$$

leading to an order G^3 contribution of type $\int d\tau_1' \int d\tau_2 \int d\tau_1 \int d\tau_{[1'2]}$, called $g_{\mu\nu[1,1'2]}$. In this way it is straightfoward to construct the higher-order terms.

Equations (4.7)–(4.14),...suggest a solution in terms of a perturbation series. One starts with the zeroth approximation that both world lines 1 and 2 are straight, and computes $g_{\mu\nu}$ on that assumption, etc. Such a solution is probably similar to that recently proposed by Turygin[23] for a half-retarded half-advanced potential, and similar to, but less far reaching than that of Ref. 8.

For the choice (2.13) of f, or in Turygin's formulation, one expects to find back the eternal Kruskal–Schwarzschild black hole solution. This is because that solution is time

translation invariant so that a choice of Green's function should not matter. Taking the mass of that black hole to be one solar mass and taking a spacelike function with cutoff $\sqrt{G}$ as in (4.2) one expects no difference at the Schwarzschild radius $r = r^*$ of one mile. This is because of arguments of the type illustrated in Fig. 2 and (2.14). Taking the cutoff at the Planck length, $\sqrt{G}$ as in (4.2), it will be completely unimportant at $r = r^*$ and one will have the usual horizon. The difference will be inside the horizon; there will be no singularity there because of the cutoff which is introduced.

To summarize: one has a Riemannian metric and test particles follow geodesics in that metric (4.8). The Bianchi identities will be satisfied, but not Einstein's equations. The particles react to the field locally but they do not produce it locally if the function f is chosen as in (4.2). The hope is to ascribe this property to space-time itself.[24]

Some final remarks: First, notice that one could have started with an arbitrary background $g_{\mu\nu}^{(0)}$ instead of with $\eta_{\mu\nu}$. The solutions will then be close to those of Einstein's theory only if $g_{\mu\nu}^{(0)}$ satisfies Einstein's equations.

Second, notice that the quanta have become like particles. The only distinctions are that quanta carry infinitesimal "mass," have finite spacelike world lines, and are unlimited in number. Thus in a sense one no longer has an action at a distance theory, but rather a pure particle theory.

Third, the split between background and quanta seems to depend on the coordinate system. Thus the quanta seem to depend on the coordinate system, a phenomenon which makes one think of Hawking radiation.

For explanations of the appearance of the background term $\eta_{\mu\nu}$ in (4.10) we refer to Refs. 8 and 19.

ACKNOWLEDGMENTS

The authors are grateful for discussions with Professor F. A. Berends, Professor L. C. Biedenharn, Professor Th. W. Ruygrok, Professor E. P. Wigner, and Professor J. W. York.

This research was supported in part by the U.S. Department of Energy under Grant No. De-FG05-85ER40219.

[1]L. H. Thomas, Phys. Rev. **92**, 1300 (1953); L. Bel, Ann. Inst. H. Poincaré A **18**, 57 (1973); F. Rohrlich, Ann. Phys. (NY) **117**, 292; V. V. Molotkov and I. T. Todorov, Commun. Math. Phys. **79**, 111 (1981); A. Komar, Phys. Rev. D **18**, 3617 (1978).
[2]H. Tetrode, Z. Phys. **10**, 317 (1922); A. D. Fokker, *ibid.* **58**, 386 (1929); Physica **12**, 145 (1932).
[3]J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. **17**, 156 (1945); **21**, 424 (1949).
[4]J. W. Dettman and A. Schild, Phys. Rev. **95**, 1059 (1954).
[5]H. Van Dam and E. P. Wigner, Phys. Rev. B **138**, 1576 (1965); **142**, 838 (1966); E. P. Wigner, in *Coral Gables Conference*, edited by T. Gudehus (Gordon and Breach, New York, 1969), p. 344.
[6]A. Katz, J. Math. Phys. **10**, 1929 (1969).
[7]J. E. Hogarth, Proc. R. Soc. London Ser. A **267**, 365 (1962).
[8]F. Hoyle and J. V. Narlikar, Proc. R. Soc. London Ser. A **277**, 1 (1964); **282**, 178, 184, 191 (1964); for a criticism see S. Deser and F. A. Pirani, Proc. R. Soc. London Ser. A **288**, 133 (1965).
[9]R. Marnelius, Phys. Rev. D **16**, 2535 (1974).
[10]J. Weiss, J. Math. Phys. **27**, 1015, 1021 (1986).
[11]J. Schwartz, Phys. Rep. **89**, 223 (1982); Phys. Lett. B **151**, 21 (1985).
[12]L. D. Landau and E. Lifschitz, *The Classical Theory of Fields* (Addison–Wesley, Cambridge, MA, 1951).
[13]N. Mukunda, H. Van Dam, and L. C. Biedenharn, Phys. Rev. D **22**, 1938 (1980).
[14]J. A. Wheeler (private communication).
[15]C. M. Anderson and H. C. von Bayer, Ann. Phys. (NY) **60**, 67 (1970); R. D. Driver, Phys. Rev. D **19**, 1098 (1979).
[16]S. K. Wong, Nuovo Cimento A **65**, 689 (1970).
[17]B. de Wit and D. Friedman, Phys. Rev. D **21**, 358 (1980).
[18]See, for instance, J. Schwinger, *Particles, Sources and Fields* (Addison–Wesley, Reading, MA, 1970), Sec. 2.4.
[19]H. Van Dam and M. Veltman, Nucl. Phys. B **22**, 397 (1970); Gen Relativ. Gravit. **3**, 215 (1972).
[20]See Sec. 3.17 of Ref. 18.
[21]C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), p. 187; A. Schild, Texas Q. **3**, 42 (1960); in *Evidence for Gravitational Theories*, edited by C. Möller (Academic, New York, 1963).
[22]J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1960), pp. 13, 150.
[23]A. Yu. Turygin, Gen. Relativ. Gravit. **18**, 333 (1986).
[24]E. P. Wigner (private communication).

Quotient of manifolds by discrete groups

F. Ardalan and H. Arfaei
International Center for Theoretical Physics, Trieste, Italy and Department of Physics, Sharif University of Technology, P. O. Box 11365-8639, Tehran, Iran[a)]

(Received 24 October 1985; accepted for publication 12 November 1986)

The quotient of manifolds by discrete subgroups of their isometry group are considered. In particular, symmetry breaking due to the quotient structure, topological properties, and harmonic analysis of the resultant manifolds are discussed and illustrated by two-dimensional examples. New solutions of $d = 11$ supergravity and the $d = 6$ Einstein–Yang–Mills theory are thus obtained, for which alterations in their spectrum and symmetry breaking are discussed.

I. INTRODUCTION

Solutions of field equations of theories involving gravity are of considerable interest. In particular, solutions of 11-dimensional and ten-dimensional supergravity have been studied extensively. The former for its uniqueness properties and the latter because it is presumably the limit of superstring theories. Solving these equations in general results in the description of the local properties of the manifold and, in particular, its Riemann curvature tensor. However, the determination of the curvature tensor does not uniquely determine the global structure of the manifold solution. In general there exist many manifolds that have the same curvature but different global structures often obtained by identifying certain points of a simply connected manifold.[1] As this procedure is unfamiliar to most physicists we will outline the construction of these non-simply connected manifolds and describe how, given a simply connected one, other manifolds with the same curvature may be obtained. In the case of symmetric spaces *all* manifolds with the same curvature are thus obtained. We will then consider symmetries of these manifolds and find how their harmonic analysis depends on the manner of identification and how mass spectra of related Kaluza–Klein theories are affected. These are discussed in the next section. In Sec. III two illustrative simple examples of zero curvature and positive curvature in two dimensions are treated in some detail. In Sec. IV we apply these ideas to S^7 solution of 11-dimensional supergravity and S^2 compactification of the six-dimensional Einstein–Yang–Mills theory. It is found that the spectrum in both cases are dramatically changed and the symmetry of the resultant S^7 compactification of $d = 11$ supergravity is reduced from SO(8) to U(1)[4] without introduction of Higgs particles.

II. GENERAL PROPERTIES

To being with, we note that, given a Riemannian manifold M and a discrete subgroup Γ of isometries of M, the set M/Γ of orbits of M under the action of Γ is itself a manifold provided no points of M are left invariant by any element of Γ, i.e., Γ acts freely on M.

To see what goes wrong when Γ does not act freely on M, consider $M = R^2$ and Γ the two-element group generated by a rotation of π about the origin. It can easily be seen that the resulting M/Γ is the two-dimensional cone which is not a manifold. The singularity at the origin is in fact due to the invariance of the origin under Γ.

In general, it is not difficult to see that the local properties of M and M/Γ, such as curvature, torsion, and metric, are the same; however, it is not clear that manifolds with the same local properties can be obtained from a single simply connected one by dividing by discrete freely acting groups. In the case of symmetric spaces this is the case and *all* the manifolds with the same curvature are thus obtained[2]: one takes the simply connected manifold M with the given Riemann curvature tensor and divides it by a subgroup Γ of isometries of M that are discrete and act freely on M. Choosing all such possible nonequivalent subgroups one recovers all the possible manifolds with the same curvature as M. Two subgroups are considered equivalent if they are conjugate. The reason symmetric spaces are more manageable in this respect is that, for a symmetric space, the symmetry allows one to construct a covering mapping between any two manifolds of the same curvature; and enumeration of possible manifolds covered by the simply connected manifold of the given Riemann curvature leads to enumeration of the possible nonequivalent discrete subgroups of the isometry group of the manifold which act freely on it. The relation between Riemannian manifolds of the same curvature tensor have been considered for some other cases also and similar results obtained[3]; however, for the rest of the article we will confine ourselves to the case of a simply connected Riemannian manifold M and its relation to various manifolds M/Γ, where Γ is necessarily a discrete isometry group of M acting on it freely.

An important question is the relation between the symmetries of M and M/Γ. It is easily seen that the group of isometries of M/Γ, denoted by $I(M/\Gamma)$, is the normalizer of Γ in $I(M)$, the group of isometries of M, i.e., the subgroup of $I(M)$ consisting of elements which commute with every element of Γ. To see this, we note that if an element of the isometry group $I(M)$ did not commute with some element γ of Γ, then for some $x \in M$, we would have $g(\gamma x) \neq \gamma(gx)$, i.e., on the space M/Γ, where γx and $\gamma(gx)$ are, respectively, identified with x and gx, the group action g would not be well defined. Consequently, the only subgroup of $I(M)$ defined on M/Γ is the normalizer of Γ in $I(M)$ and

[a)] Permanent address.

 0022-2488/87/030685-05$02.50

$I(M/\Gamma)$ = normalizer of Γ in $I(M)$;

when additional gauge symmetries are also considered, a similar argument[4] leads to the same conclusion that the remaining symmetry on M/Γ is the normalizer of Γ in the gauge symmetry of M.

This reduction of $I(M)$ to $I(M/\Gamma)$ allows for a mechanism of symmetry breaking in pure Kaluza–Klein theories, which, to our knowledge, has not been considered previously, although the case of Kaluza–Klein–Yang–Mills symmetry breaking has recently been discussed with the breaking of the gauge symmetry.[4]

It is clear that when $I(M)$ is transitive on M, i.e., M is homogeneous, then M/Γ is also homogeneous, i.e., $I(M/\Gamma)$ is transitive on M/Γ, if Γ commutes with $I(M)$. However, $I(M/\Gamma)$ may still be transitive on M and consequently on M/Γ, making M/Γ homogeneous even if Γ does not commute with $I(M)$. In general $I(M/\Gamma)$ is not transitive on M and M/Γ is not homogeneous.

The next interesting aspect of the quotient manifold M/Γ is the relation between the topological invariants of M and M/Γ. When M is simply connected, then the first homotopy group of M/Γ is clearly isomorphic with Γ, which could be a non-Abelian group. However, the higher homotopy groups of M/Γ are identical to those of M (see Ref. 5), which are necessarily Abelian. The homology groups of M and M/Γ are not so simply related, as can be seen from simple examples. We will list them for our examples in the next section, but will not pursue it any further here other than mentioning a simple relation between their characteristic classes when they are expressible in terms of geometric quantities. For example, due to the Gauss–Bonnet theorem, which relates the Euler characteristic to an integral of the curvature over the compact manifold, it is easily seen that the Euler characteristic of M/Γ is smaller than that of M by a factor equal to the number of elements of Γ (the volume of M/Γ is that much "smaller").

To end the discussion of the general properties of M/Γ we will consider harmonic analysis over the manifold M/Γ. For simplicity we only consider scalar functions over M and M/Γ. When M is a homogeneous space, f can be decomposed into a direct sum of irreducible pieces under the action of the isometry of M, and the decomposition is well known.[6] However, when the same function is considered over M/Γ, one has to make sure that it is well defined, i.e., it must be guaranteed that $f(\gamma x) = f(x)$, $\forall x \in M$ and $\forall \gamma \in \Gamma$ (more general transformation properties with some weight associated with Γ have been considered in the literature[7]). Then the decomposition of f under the residual symmetry of M/Γ is to be investigated anew. In general, the decomposition changes when we go from M to M/Γ resulting in a significant change in the spectrum of the related Kaluza–Klein-type theories. To illustrate this, we limit ourselves to the case of the manifold M being a Lie group G and consider a scalar function f under left multiplication of G on itself.

It is known that the space of scalar functions is decomposed into irreducible representations of G with multiplicities equal to the dimension of the irreducible representation spaces. When dividing M by Γ, the space of scalar functions on M/Γ (which are therefore those functions on M which remain fixed under Γ) decompose under G with a different set of multiplicities; each multiplicity now being equal to the number of linearly independent vectors in the irreducible representation space (of G).[7] Thus not only the number of states with the same quantum numbers appearing in a Kaluza–Klein theory changes thereby, but some states may even disappear altogether from the spectrum of the theory.

III. EXAMPLES

To illustrate some of the aspects of the above discussion we will go into some detail for the two simple cases of S^2 and R^2. In the case of S^2, the simply connected two-dimensional manifold of constant positive curvature (Gauss curvature), the classification of manifolds covered by it is quite simple. There is only one other manifold: RP^2, obtained by identifying the antipodal points, i.e., by dividing S^2 by the subgroup of O_3 consisting of the identity element and the reflection about the origin, $\Gamma = \{1, -1\}$. In fact it is possible to prove that for all even-dimensional spheres this group Γ is the only discrete freely acting isometry of S^{2n} and thus RP^{2n} is the only manifold covered by S^{2n}. This is because $SO(2n+1)$ has at least one eigenvalue equal to unity. Moreover, it may be shown[2] that S^{2n} is the only simply connected $2n$-dimensional Riemannian manifold of constant positive sectional curvature, thus completing the classification of constant positive curvature manifolds. Returning to S^2, we note that $\Gamma = \{1, -1\}$ commutes with O_3 and therefore Rp^2 has the same isometry group as S^2, i.e., O_3; and there is no way therefore of reducing the symmetry of S^2 by identification of its points. As far as their topology is concerned, we have for the first homotopy group $\pi_1(S^2) = 0$ and $\pi_1(RP^2) = Z_2$ reflecting the double connectedness of RP^2 and $\pi_2(S^2) = \pi_2(RP^2)$ reflecting the identity of higher homotopy groups discussed above. For purposes of comparison we list their homology groups as well: $H_1(S^2) = 0$, $H_2(S^2) = 0$; $H_1(RP^2) = Z_2$, $H_2(RP^2) = 0$. Note that the first integral homology group of RP^2 is only torsion. The Euler characteristics of S^2 and RP^2 are 2 and 1, respectively, in agreement with the general discussion in Sec. II. The harmonic expansion for a scalar function on S^2 is the familiar spherical harmonic expansion

$$f(\theta,\varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\theta,\varphi) , \tag{1}$$

where (θ,φ) are the spherical coordinates of S^2 and the Y_{lm} are the spherical harmonics. Reduction of S^2 to RP^2 eliminates odd-l spherical harmonics from the decomposition

$$f(\theta,\varphi) = \sum_{l \text{ even}} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\theta,\varphi) , \tag{2}$$

thus $(4n+3)$-dimensional representations of O_3 do not occur in the decomposition.

For the second example we take R^2. Its isometry is E_2, which is the semidirect product of O_2 with the group of translations in two dimensions. It is clear that translations act freely on R^2 and consequently any of its discrete subgroups will do. It is not hard to see that the most general discrete subgroup Γ of E_2 acting freely on R^2 is a combination of a translation and a reflection about an axis, say the x axis.[2] Thus it is straightforward to deduce the possible mani-

folds covered by R^2: (1) the cylinder, where Γ is generated by a single translation; (2) the torus, where Γ is generated by two independent translations; (3) the Möbius strip, where Γ is generated by the reflection about the x axis together with a translation along the x axis; and finally (4) the Klein bottle, where Γ is generated by the reflection about the x axis together with a translation along the x axis and a translation along the y axis. Of course the length of various translations here generate different manifolds with different "sizes." The above four categories may be shown to be the only flat manifolds of dimension 2 (see Ref. 2). For brevity, we will limit the rest of the discussion to the compact manifolds torus T^2 and Klein bottle K.B. For T^2 the normalizer of Γ in E_2 is a subgroup of translations which is the two-dimensional Abelian torus group $U(1)\times U(1)$. For K.B. the isometry group is further reduced; the normalizer of Γ in E_2 being just translations along the x axis modulo its size which is a single $U(1)$ group. It is noteworthy that the K.B. can also be obtained from the torus T^2 by cutting T^2 in half and identifying opposite corners, i.e., by dividing T^2 by the group generated by reflection about the x axis together with a translation along that axis. Then the symmetry of T^2, i.e., $U(1)\times U(1)$ is reduced to $U(1)$, the symmetry of the Klein bottle via the above arguments. It is therefore interesting to compare their homotopy and homology groups. They are as follows: $\pi_1(T^2)$ is an Abelian infinite group generated by two elements; $\pi_1(\text{K.B.})$ is the non-Abelian group of the semidirect product of the infinite group generated by two translations and a reflection about an axis. Note that the Klein bottle is therefore an example of a manifold with non-Abelian fundamental group. Here π_2 of T^2 and K.B. are trivial. Also, $H_1(T^2)=Z\times Z$, $H_1(\text{K.B.})=Z\times Z_2$, $H_2(T^2)=Z$, $H_2(\text{K.B.})=0$. The Euler characteristics of both T^2 and K.B. vanish. Harmonic analysis on T^2 is the usual Fourier series decomposition in two variables; when going to the Klein bottle, the odd functions in y, $\sin ny$, are dropped since they are not invariant under the reflection part of Γ.

IV. PHYSICAL APPLICATIONS

In this section we apply the procedure of Sec. II to the S^2-monopole-type compactification[8] of the six-dimensional Einstein–Yang–Mills theory by considering the quotient space $RP^2=S^2\{1,-1\}$, which was considered in Sec. III; and to the S^7 compactification[9] of 11-dimensional supergravity by dividing S^7 over one of the many appropriate discrete subgroups of SO(8). In both cases the spectrum is altered drastically. Among other things the "photon" is removed from the spectrum of the former theory and in the latter theory the 35 massless vector particles is reduced to four particles. Moreover the SO(8) symmetry of S^7 is reduced to $U(1)$[4].

In the six-dimensional model of Ref. 8 the manifold $M_4\times S^2$ is found as a solution of the coupled Einstein–Yang–Mills equations. The metric is decomposed into two parts, $g_{\mu\nu}(x)$ on the M_4 and $g_{mn}(y)$ on the S^2, where x and y parametrize M_4 and S^2, respectively,

$$g_{ij}\,dz^i\,dz^j=g_{\mu\nu}(x)dx^\mu\,dx^\nu+g_{mn}(y)dy^m\,dy^n\,. \tag{3}$$

The two-sphere metric has the standard form

$$g_{mn}(y)dy^m\,dy^n=a^2(d\theta^2+\sin^2\theta\,d\varphi^2)\,, \tag{4}$$

where θ and φ are the usual spherical coordinates while M_4 is found to be an anti-de Sitter space. The U(1) field has nonzero components only on the S^2 part where it has a monopolelike configuration of topological charge n,

$$A_i(z)dz^i=A_m(y)dy^m=(n/2e)(\cos\theta\pm 1)d\varphi\,, \tag{5}$$

the plus (minus) sign referring to the coordinate patch excluding the south (north) pole of the sphere. Note that in the overlap region the two expressions can be transformed to one another by the gauge transformation

$$\Lambda=n\varphi \tag{6}$$

on the lower half of the sphere.

The identification we apply to S^2 as illustrated in the previous section is the identification of antipodal points, (θ,φ) and $(\pi-\theta,\varphi+\pi)$. Hence a scalar field $\Phi(\theta,\varphi)$ on S^2 is well defined on RP^2 provided

$$\Phi(\theta,\varphi)=\Phi(\pi-\theta,\varphi+\pi) \tag{7a}$$

and a vector function is well defined provided

$$A_m=(\theta,\varphi)=-A_m(\pi-\theta,\varphi+\pi)\,. \tag{7b}$$

For a gauge field A_m on RP^2 this condition can be imposed after making a gauge transformation

$$A_m(\theta,\varphi)=-A_m(\pi-\theta,\varphi+\pi)-\nabla_m\Lambda\,. \tag{7c}$$

Similarly for a second-rank tensor we must have

$$g_{mn}(\theta,\varphi)=g_{mn}(\pi-\theta,\varphi+\pi)\,. \tag{7d}$$

Consequently the S^2 metric is well defined on RP^2, while the gauge field (5) transforms properly after making a gauge transformation,

$$\begin{aligned}A_\varphi(\pi-\theta,\varphi+\pi)&=(-n/2e)(\cos\theta\mp 1)\\&=(-n/2e)(\cos\theta\pm 1)+(1/e)\partial_\varphi\Lambda\,,\end{aligned}$$

where $\Lambda=n\varphi$ in accordance with (7c). The field strength $E_{\theta\varphi}=-n/2e$ is also well defined and agrees with (7d).

To obtain the spectrum of fluctuations about these RP^2 solutions to the Einstein–Yang–Mills theory one has to start from the S^2 solutions and impose the above constraints. These constraints will remove the scalars and second-rank tensors with odd l's and the vectors with even l's. As a result the zero mass photon which has $l=0$ is removed from the spectrum. The spectrum of massive particles similarly changes. The S^2 solution has six towers of massive particles as given in Table I. In the last column of the table, the changes occurring in the RP^2 solutions are specified. To see how these changes take place we have to examine the gauge constraints that the external sources T_{ij} and V_j coupling to g_{ij} and A_j satisfy. (For details of the S^2 solution we refer the reader to Ref. 8.) These constraints are J^i, $i=0$, and $T_{ij,j}=KF_{ij}J_j$, which, after combination with (7a)–(7d), imply $T_{++}=T_{--}$ and $T_{i+}=T_{i-}$ ($\pm$ refer to $y_5\pm iy_6$ coordinate on the sphere). This result removes all the $\lambda=2$ scalars, deleting the first tower of massive particles. The next two towers of scalars for $\lambda=+1$ and -1 are restricted to even l's and therefore half of them are absent in the case of RP^2. To obtain this last result we have used $J_+ + J_-=0$, which is again a consequence of the constraints. The vector

TABLE I. Massive particle spectrum for S^2 compactification of the six-dimensional Einstein–Yang–Mills theory and its RP^2 modification.

Space-time type	(Mass)^2	$\lvert\lambda\rvert$	l	RP^2 modification
Scalar	$M_0^2 = (l-1)((l+2)/a^2$	2	$l>2$	Removed from the spectrum
Scalar	$M_{0\pm}^2 = [2l(l+1)+1 \pm \sqrt{1+2l(l+1)}\,]/a^2$	0	$l>0$	l = even
Vector	$M_{1\pm}^2 = [l(l+1) \pm \sqrt{2l(l+1)}\,]/a^2$	1	$l>1$	l = odd
Second-rank tensor	$M_2^2 = l(l+1)/a^2$	0	$l>0$	l = even

particles are restricted to odd l's and the spin-2 particles to even l's.

This RP^2 solution, having a different spectrum of particles, has not been considered in the study of the solutions of six-dimensional Einstein–Yang–Mills theory.

Our second example is a solution to $d = 11$ supergravity. Compactifying solutions to this theory are classified.[9] In particular the most symmetric solution where the metric of the compact space is that of S^7 is studied in detail,[10,11] but the point that the metric does not specify the global structure has been overlooked. In Ref. 9 we pointed out that S^7 is the simply connected manifold with the standard spherical metric. Now we wish to look at other spaces with the same metric as S^7. Among several identifications of S^7 (see Ref. 2) we choose one that is implemented by the discrete subgroup Γ of SO(8) generated by $\gamma \in$ SO(8):

$$\gamma = \begin{bmatrix} R(2\pi/n) & 0 & 0 & 0 \\ 0 & R((2\pi/n)a_2) & 0 & 0 \\ 0 & 0 & R((2\pi/n)a_3) & 0 \\ 0 & 0 & 0 & R((2\pi/n)a_4) \end{bmatrix},$$

where $R(\alpha)$ is a 2×2 matrix in SO(2) signifying a rotation by angle α. We note that a_2, a_3, a_4, and n are relatively prime. The centralizer of Γ in SO(8) consists of a set of elements of the form

$$\begin{bmatrix} R(\theta_1) & & & \\ & R(\theta_2) & & \\ & & R(\theta_3) & \\ & & & R(\theta_4) \end{bmatrix}.$$

This is an Abelian group $U(1)^4$ consisting of independent rotations in the planes (1,2), (3,4), (5,6), and (7,8). The manifold S^7/Γ has the same curvature as S^7, but globally admits only the group $U(1)^4$ as its isometry group. Hence upon this identification the gauge group of the four-dimensional theory is broken from SO(8) to $U(1)^4$. Closer examination of the massless vector states shows that from 28 gauge particles in the S^7 solution only four will survive in S^7/Γ solution. Note that this "symmetry breaking" takes place without the introduction of Higgs bosons and the vector particles corresponding to the broken symmetries do not become massive, but are totally eliminated from the theory, similar to the symmetry breaking pattern of the Calabi–Yau compactification in the superstring theory.

Now we consider the change in the spectrum of scalars when we go from S^7 to S^7/Γ. Several authors[10,11] have studied the spectrum of S^7. Sezgin[11] has considered them in detail. Scalars can be expanded in functions of S^7, which are eigenfunctions of eight-dimensional angular momentum. They correspond to homogeneous polynomials in eight variables $X_1, X_2, \ldots, X_8$ (see Ref. 12). The degree of the polynomial is denoted by l. Sezgin finds that for $l = 2$ we obtain a 35 SO(8) multiplet of massless scalars, corresponding to 35 linearly independent second-order homogeneous polynomials in eight variables. (Note that $r^2 = \Sigma\, x_i^2$ is invariant under rotations and is a singlet, so the 36-dimensional space of second-order homogeneous polynomials decomposes into two parts, a singlet and a 35 multiplet.) In S^7/Γ only those polynomials that are invariant under Γ survive. To find them it is useful to define new variables

$$\xi_i^{\epsilon} = X_{2i-1} \pm iX_{2i}, \quad i = 1,2,3,4, \quad \epsilon = + \text{ or } - .$$

A suitable basis with definite transformation property under Γ for the $l = 2$ functions are monomials $P_{ij}^{\epsilon\epsilon'} = \xi_i^{\epsilon}\xi_j^{\epsilon'}$.

Under γ, $P_{ij}^{\epsilon\epsilon'}$ transforms as follows:

$$P_{ij}^{\epsilon\epsilon'} \to \exp(2\pi a_i/n)\epsilon + (2\pi a_j/n)\epsilon')P_{ij}^{\epsilon\epsilon'}$$

(where $a_1 = 1$). The monomial is invariant under Γ if and only if

$$a_i\epsilon + a_j\epsilon' = 0 \bmod(n).$$

Since $a_i < n$ and the a_i's and n are relatively prime the above condition can hold if (1) $i = j$ and $\epsilon = -\epsilon'$ or (2) $a_i + a_j = n$ and $\epsilon = \epsilon'$. From the first possibility we obtain three independent zero mass scalars. The second possibility does not happen all the time. Simple reasoning shows that (2) can give at most two massless scalars. So we will be left with three, four, or five massless scalars from the original 35 ones.

ACKNOWLEDGMENTS

We have profited from the indispensible guidance of Professor S. Shahshahani. We acknowledge discussions with

Dr. R. S. Randjbar-Daemi and Dr. H. Sarmadi. We would like to thank Professor J. Strathdee for reading the manuscript. We would like to thank Professor Abdus Salam, the International Atomic Energy Agency, and UNESCO for the hospitality extended to us at the International Center for Theoretical Physics, where part of this work was done.

[1]E. Witten, Nucl. Phys. B **186**, 412 (1981); "Symmetry breaking patterns in superstring models," preprint, February 1985; L. Dixon, J. A. Harvey, C. Vafa, and E. Witten, "Strings on orbifolds," Princeton preprint, 1985; F. Ardalan and H. Arfaei, Gen. Relativ. Gravit. **18**, 675 (1986).
[2]J. A. Wolf, *Spaces of Constant Curvatures* (Publish or Perish, Wilmington, DE, 1984), 5th ed.
[3]A. E. Fischer and J. A. Wolf, J. Diff. Geom. **10**, 277 (1975); E. A. Ruh, *ibid.* **17**, 1 (1982).
[4]See the second article by E. Witten in Ref. 1, and references cited therein.
[5]E. H. Spanier, *Algebraic Topology* (McGraw-Hill, New York, 1966).
[6]A. Salam and J. Strathdee, Ann. Phys. (NY), **141**, 316 (1982).
[7]I. M. Gel'fand and I. I. Pyateckii-Shapiro, Usp. Math. Nauk. **14**, 171 (1959).
[8]S. Randjbar-Daemi, A. Salam, and J. Strathdee Nucl. Phys. B **214**, 491 (1983).
[9]See the last reference in Ref. 1 above.
[10]B. Biran, A. Casher, F. Englert, and M. Rooman, Phys. Lett. B **134**, 179 (1984).
[11]E. Sezgin, Trieste Preprint IC/84/65.
[12]N. J. Vilenkin, *Special Functions and the Theory of Group Representations* (Am. Math. Soc., Providence, RI, 1968).

The many-component limit of the Potts lattice gauge model in the external field

Paweł Maślanka
Institute of Mathematics, University of Łódź, 90-238 Łódź, ul. St. Banacha 22, Poland

(Received 6 May 1986; accepted for publication 14 August 1986)

This paper generalizes the results obtained by Kotecky. The free energy of the gauge Potts model in the external field in the many component limit has been calculated. This is done in two cases: (a) the external field is switched on after the gauge fixing or (b) the external field is switched on before the gauge fixing. In both cases, the free energy can be calculated using the mean-field method, if, in case (b), the larger class of trial measures is allowed. The mechanism of phase transition is also discussed.

I. INTRODUCTION

One of the most useful methods of obtaining a least qualitative insight into a phase structure of various lattice models is provided by the mean-field theory. The core of this method consists of rewriting the expression for the partition function in equivalent form with the help of some trial probability measure, and then neglecting the correlations with respect to this distribution (this is achieved by use of the Jensen inequality). To make the problem tractable, the trial measure is chosen in such a way that all dynamic variables are distributed independently.

The mean-field method gives an upper bound for the free energy. The old question is whether there exist models for which the mean-field theory gives an exact expression. This was shown to be the case for the Ising,[1] n-vector,[2] anisotropic Heisenberg,[3] and Potts[4] models in the long-range limit, and for Ising,[5,6] n-vector, spherical, and quantum spin[5] as well as Potts[4] models in the high-density limit. The many-component limit of the Potts model was also considered from this point of view[7]; the conclusion again was that the mean field method gives an exact answer.

In all of the above cases, after the appropriate choice of the trial measure, all correlations that should disappear in order to provide the exactness of the mean-field expression do indeed vanish after the high-density, long-range, or many-component limits have been taken. An important point is that one has succeeded with the very simple form of trial probability measure. However, this need not always be the case. If the trial distribution is not consistent with the properties of the system implied by some large symmetry group, the mean-field method may fail. This can be illustrated using the gauge Potts model as an example. It was shown by Kotecky[8] that one can calculate the free energy of this model in the many-component limit by the mean-field method only after the gauge symmetry has been explicitly broken by the appropriate choice of gauge.

In this paper we generalize the results obtained by Kotecky. We calculate the free energy of the gauge Potts model in the external field in the many-component limit. We do this in two cases: (a) the external field is switched on after the gauge fixing or (b) the external field is switched on before the gauge fixing. We show that in both cases we can calculate the free energy using the mean field method, if, in case (b), we allow a larger class of trial measures. We can also calculate the minimal value of the external field above which the standard mean field theory is applicable and discuss the mechanism of phase transition.

The paper is organized as follows. In the rest of this section we introduce some necessary notions and present the results obtained by Kotecky. Finally, we state our main results in Theorem 1.1. The proof of this theorem is given in Sec. II.

Section III is devoted to some remarks and final conclusions. Some technicalities are relegated to the Appendix.

Let us consider a d-dimensional hypercubic lattice. We label the unit coordinate vectors by $\hat{\mu}$ ($\mu = 0,1,...,d-1$). If $i \in Z^d$ is a lattice site, we denote a nonoriented link connecting sites i and $i+\hat{\mu}$ by a pair (i,μ), and a plaquette bordered by links (i,μ), $(i+\hat{\mu},\nu)$, $(i+\hat{\nu},\mu)$, (i,ν), by a triple (i,μ,ν), $\mu<\nu$. Generically, the links will be denoted by l and the plaquettes by p. To any link l we attach the spin variable σ_l, taking values in $Z_q = \{0,1,...,q-1\}$. Given any configuration $\sigma \equiv \{\sigma_l\}$, we introduce the plaquette variable by $\sigma_p \equiv \sigma_{(i,\mu)} + \sigma_{(i+\hat{\mu},\nu)} - \sigma_{(i+\hat{\nu},\mu)} - \sigma_{(i,\nu)} \pmod q$, for $p=(i,\mu,\nu)$. Consider a hypercube Λ consisting of $|\Lambda|$ lattice sites. First we define version (b) of the model. The Hamiltonian reads

$$H_\Lambda(\sigma_\Lambda) = -\sum_p \delta_{\sigma_p,0} - h\sum_l \delta_{\sigma_l,0}. \tag{1.1}$$

The free energy is then defined by the formula

$$\exp[-\beta|\Lambda| f(\beta,h,q)] = Z_\Lambda(\beta,h,q) = \sum_{\sigma_\Lambda} \exp[-\beta H_\Lambda(\sigma_\Lambda)]. \tag{1.2}$$

To introduce the second version (a), we follow the standard method used in the perturbative theory of continuum gauge fields. To fix the gauge following Kotecky, we will use the temporal one: $\sigma_l = 0$, for all $l = (i,0)$; then integrate out the gauge group volume and finally couple the external field to the remaining variables. The modified Hamiltonian reads

$$\tilde{H}_\Lambda(\tilde{\sigma}_\Lambda) = -\sum_p \delta_{\sigma_p,0} - h\sum_{l\text{-horiz}} \delta_{\sigma_l,0}. \tag{1.3}$$

The second sum on the right-hand side runs over all horizontal links. The appropriate free energy is defined by

 0022-2488/87/030690-07$02.50

$$\exp(-\beta|\Lambda|\tilde{f}(\beta,h,q))$$

$$=\tilde{Z}_\Lambda(\beta,h,q)=\sum_{\tilde{\sigma}_\Lambda}\exp(-\beta\tilde{H}_\Lambda(\tilde{\sigma}_\Lambda)). \tag{1.4}$$

Let us remark that the following simple relation holds as a consequence of integration over the gauge group:

$$f(\beta,0,q)=\tilde{f}(\beta,0,q)-\beta^{-1}\log q. \tag{1.5}$$

The following fact was proved by Kotecky. The limiting free energy has the form

$$\tilde{f}_\infty(\beta)=\lim_{q\to\infty}\tilde{f}(\beta\log q,0,q)$$

$$=\min\left(-\frac{d(d-1)}{2},-\frac{(d-1)}{\beta}\right), \tag{1.6}$$

and can be obtained by the standard mean-field method (i.e., with trial probability distribution treating all variables as independent).

It follows from Eqs. (1.5) and (1.6) that

$$f_\infty(\beta)=\lim_{q\to\infty}f(\beta\log q,0,q)$$

$$=\min\left(-\frac{d(d-1)}{2}-\frac{1}{\beta},-\frac{d}{\beta}\right). \tag{1.7}$$

However, for the model defined by Eqs. (1.1) and (1.2), the mean-field method in its usual form gives

$$\lim_{q\to\infty}f_{\mathrm{MF}}(\beta\log q,0,q)=\min\left(-\frac{d(d-1)}{2},-\frac{d}{\beta}\right), \tag{1.8}$$

and for small temperatures one does not obtain the right answer.

Let us now state our main result.

Theorem 1.1: The following limiting relations hold:

$$\text{(a)}\ \tilde{f}_\infty(\beta,h)\equiv\lim_{q\to\infty}\tilde{f}(\beta\log q,h,q)=\min\left(-\frac{d(d-1)}{2}-h(d-1),-\frac{d(d-1)}{2},-\frac{d-1}{\beta}\right),$$

$$\text{(b)}\ f_\infty(\beta,h)\equiv\lim_{q\to\infty}f(\beta\log q,h,q)=\min\left(-\frac{d(d-1)}{2}-hd,-\frac{d(d-1)}{2}-\frac{1}{\beta},-\frac{d}{\beta}\right).$$

We prove this theorem in the next section and postpone its discussion to Sec. III.

II. THE PROOF OF THEOREM 1.1

A. An upper bound

To obtain an upper bound we follow the standard methodology of obtaining the mean field approximation. Let $\{i\}$ be the set of all states of the system under consideration, $H(i)$ the corresponding Hamiltonian, and $\rho(i)$ any probability distribution fulfilling the condition $\rho(i)>0$ for all i. Then

$$Z=\sum_i\exp(-\beta H(i))$$

$$\equiv\sum_i\rho(i)\exp[-\beta H(i)-\log\rho(i)]$$

$$\geqslant\exp\left\{\sum_i\rho(i)[-\beta H(i)-\log\rho(i)]\right\},$$

and consequently

$$f=-\beta^{-1}\log Z\leqslant\langle H\rangle_\rho+(1/\beta)\langle\log\rho\rangle_\rho. \tag{2.1}$$

Now, the right-hand side is the continuous function of ρ in the domain $\rho(i)\geqslant 0$. Consequently we obtain the following upper bound:

$$f\leqslant\min_{\rho;\rho(i)>0}[\langle H\rangle_\rho+(1/\beta)\langle\log\rho\rangle_\rho]. \tag{2.2}$$

To apply the bound (2.2) to our case we consider the two versions separately.

(a) Let us first assume $h\geqslant 0$. Putting

$$\rho(\tilde{\sigma}_\Lambda)=\prod_{l\in\Lambda}\rho_l(\tilde{\sigma}_l),$$

choosing a ρ_l that prefers $\sigma_l=0$ with probability p, and distributing the remaining values uniformly we can follow the calculations performed by Kotecky to obtain

$$\tilde{f}_\infty(\beta,h)\leqslant\min\left[-\frac{d(d-1)}{2}-h(d-1),-\frac{d-1}{\beta}\right]$$

$$=\min\left[-\frac{d(d-1)}{2}-h(d-1),-\frac{d(d-1)}{2},-\frac{d-1}{\beta}\right].$$

If $h<0$, we choose ρ_l preferring $\sigma_l=1$ with the probability p and distributing the remaining values uniformly. Then we obtain

$$\tilde{f}_\infty(\beta,h)\leqslant\min\left[-\frac{d(d-1)}{2},-\frac{d-1}{\beta}\right]$$

$$=\min\left[-\frac{d(d-1)}{2}-h(d-1),-\frac{d(d-1)}{2},-\frac{d-1}{\beta}\right]. \tag{2.3}$$

(b) Applying the same reasoning to the second model we find

$$f_\infty(\beta,h)\leqslant\min\left[-\frac{d(d-1)}{2}-hd,-\frac{d(d-1)}{2},-\frac{d}{\beta}\right]. \tag{2.4}$$

However, we can also use another trial distribution function. Namely, let $\rho(\sigma_\Lambda)=0$ for all configurations such that $\sigma_p\neq 0$ for at least one plaquette and all other configurations (i.e., those for which $\sigma_p=0$ for all plaquettes P) are distributed uniformly. The number of the latter equals the order of

the gauge group, i.e., $q^{|\Lambda|}$. The corresponding entropy is easily calculated:

$$\langle -\log\rho\rangle_\rho = |\Lambda|\log q. \tag{2.5}$$

The mean plaquette energy equals $|\Lambda|d(d-1)/2$. Consider finally the energy of interaction with the external field. Taking into account the fact that the transformation $\sigma_l \to \sigma_l + c$ with the same constant c for all links does not change the value of any plaquette, we conclude that, given any link l, to every state in which this link contributes an amount h of the energy, there correspond the $q-1$ states in which the link l does not contribute.

Consequently, the mean energy of the external coupling is $|\Lambda|dh/q$. Collecting the above values we get

$$f_\infty(\beta,h) \leqslant -d(d-1)/2 - 1/\beta.$$

The above inequality together with Eqs. (2.4) and (2.5) gives

$$f_\infty(\beta,h) \leqslant \min\left[-\frac{d(d-1)}{2} - hd,\ -\frac{d(d-1)}{2} - \frac{1}{\beta}, -\frac{d}{\beta}\right]. \tag{2.6}$$

B. A lower bound

Again we consider the two cases separately.

(a) We adopt the strategy developed by Kotecky and evaluate the partition function by collecting the terms that have the same sets P of nonfrustrated plaquettes and summing over all subsets P. We have

$$\tilde{Z}_\Lambda = \sum_P \exp(\beta|P|) \sum_{\tilde\sigma_\Lambda:\sigma_p=0 \Rightarrow p\in P} \exp\left(\beta h \sum_l \delta_{\tilde\sigma_l,0}\right)$$

$$\leqslant \sum_P \exp(\beta|P|) \sum_{\tilde\sigma_\Lambda:p\in P\Rightarrow\sigma_p=0} \exp\left(\beta h \sum_l \delta_{\tilde\sigma_l,0}\right). \tag{2.7}$$

Now one has to estimate the last sum on the right-hand side. Let us assume that $h \geqslant 0$. Following Kotecky we can choose the set $L(P)$ of horizontal links such that (i) for each configuration on $L\setminus L(P)$ there exists at most one configuration $\tilde\sigma_\Lambda$ such that $\sigma_p = 0$ whenever $p\in P$; and (ii) $|L(P)| \geqslant 2|P|/d$ with modulo boundary terms $(O(\partial\Lambda))$. Condition (i) means that in the sum

$$\sum_{\tilde\sigma_\Lambda:p\in P\Rightarrow\tilde\sigma_p=0} \exp\left(\beta h \sum_l \delta_{\tilde\sigma_l,0}\right),$$

the links from $L\setminus L(P)$ can be viewed as independent.

To any configuration on $L\setminus L(P)$, we can overestimate the contribution of the links in $L(P)$ by $\exp(\beta h|L(P)|)$; then those from $L\setminus L(P)$ act as the set of independent spins in external field. Consequently

$$\sum_{\tilde\sigma_\Lambda:p\in P\Rightarrow\tilde\sigma_p=0} \exp\left(\beta h \sum_l \delta_{\tilde\sigma_l,0}\right) \leqslant [(q-1)+\exp(\beta,h)]^{|\Lambda|(d-1)-(2|P|/d)} \times \exp[\beta h(2|P|/d)]. \tag{2.8}$$

From the inequalities (2.7) and (2.8) we finally get

$$\tilde{Z}_\Lambda(\beta,h,q) \leqslant \sum_{|P|=0}^{|\Lambda|d(d-1)/2} \binom{|\Lambda|d(d-1)/2}{|P|} \times \exp\left[\left(\beta + \frac{2\beta h}{d}\right)|P|\right] \times [q+\exp(\beta h)]^{(2/d)(|\Lambda|d(d-1)/2-|P|)} = [\exp\beta[1+(2h/d)] + (q+\exp(\beta h))^{2/d}]^{|\Lambda|d(d-1)/2}. \tag{2.9}$$

From the inequality (2.9) the needed lower bound

$$\tilde{f}_\infty(\beta,h) \geqslant \min\left[-\frac{d(d-1)}{2} - h(d-1), -\frac{d-1}{\beta}\right] = \min\left[-\frac{d(d-1)}{2} - h(d-1), -\frac{d(d-1)}{2}, -\frac{d-1}{\beta}\right]$$

follows immediately.

Finally, let us remark that for $h<0$ we can estimate the left-hand side of Eq. (2.8) by the number of configurations. Therefore

$$\tilde{f}_\infty(\beta,h) \geqslant \min\left[-\frac{d(d-1)}{2}, -\frac{d-1}{\beta}\right] = \min\left[-\frac{d(d-1)}{2} - h(d-1), -\frac{d(d-1)}{2}, -\frac{d-1}{\beta}\right].$$

This concludes the proof for case (a).

(b) This case is a bit more complicated. Again we write out the basic inequality

$$Z(\beta,h,q) = \sum_P \exp(\beta|P|) \sum_{\sigma_\Lambda:\sigma_p=0\Leftrightarrow p\in P} \exp\left(\beta h \sum_l \delta_{\sigma_l,0}\right) \leqslant \sum_P \exp(\beta|P|) \sum_{\sigma_\Lambda:p\in P\Rightarrow\sigma_p=0} \exp\left(\beta h \sum_l \delta_{\tilde\sigma_l,0}\right). \tag{2.10}$$

Again we assume first $h \geqslant 0$. In order to estimate the last sum on the right-hand side, let us call ∂P the set of all links bordering the plaquettes from P. Together with the sites being their end points, the links from ∂P form a graph, possibly nonconnected which is embedded into the d-dimensional lattice. We denote this graph also by ∂P. Any vertex belonging to this graph is the end point of at least two and at most $2d$ lines. The number of lines of ∂P can be easily estimated. To any plaquette there belong four bordering links and any link borders at most $2(d-1)$ plaquettes. Consequently $2|\partial P|(d-1) \geqslant 4|P|$ and finally

$$|\partial P| > 2|P|/(d-1). \tag{2.11}$$

Now it is obvious that evaluating the expression under consideration we may treat the spins sitting on links not belonging to ∂P as independent spins in the external field. As a result we get

$$\sum_{\sigma_\Lambda:p\in P\Rightarrow\sigma_p=0} \exp\left(\beta h \sum_l \delta_{\sigma_l,0}\right) < (q+\exp\beta h)^{|\Lambda|d-2|P|/(d-1)} (q+\exp\beta h)^{-a} \times \sum_{\sigma_{\partial P}:p\in P\Rightarrow\sigma_p=0} \exp\left[\beta h \sum_{l\in\partial\mathcal{P}} \delta_{\sigma_l,0}\right]. \tag{2.12}$$

Here $\sigma_{\partial P}$ denotes any configuration on ∂P, and $a = |\partial P| - 2|P|/(d-1) \geqslant 0$. Passing to the last expression on the right-hand side of Eq. (2.12), let us first consider the special form of the set P that corresponds to the set contractible to a point in the continuum version of Poincaré lemma. Namely, we demand that in each connected component $\partial P'$ of ∂P, there exists a Cayley tree T with the following property: for any $l \in \partial P'$, $l \notin T$, the (unique) loop in $\{l\} \cup T$ is the boundary of some orientable surface consisting solely of plaquettes from P. It is trivial to see that for such sets P the configuration $\sigma_{\partial P}$ is a pure gauge, i.e.,

$$\sigma_l = \sigma_i - \sigma_j, \tag{2.13}$$

where i and j are the end points of l.

Then we obtain the partition function for the usual (non-gauge) Potts model

$$\sum_{\sigma_{\partial P}: p \in P \Rightarrow \sigma_p = 0} \exp\left(\beta h \sum_l \delta_{\sigma_l 0}\right) = q^{-c(\partial P)} \sum_{\sigma_{V(\partial P)}} \exp\left(\beta h \sum_{l \in \partial P} \delta_{\sigma_l \sigma_j}\right). \tag{2.14}$$

Here $V(\partial P)$ is the set of all vertices of ∂P and $c(\partial P)$ is the number of connected components of ∂P. The extra factor $q^{-c(\partial P)}$ appears because, due to the global symmetry $\sigma_i \to \sigma_i + \text{const}$, Eq. (2.13) has q solutions with respect to σ_i's in each connected component of ∂P.

We prove the following inequality:

$$q^{-c(\partial P)} \sum_{\sigma_{V(\partial P)}} \exp\left[\beta h \sum_l \delta_{\sigma_l \sigma_j}\right] \leqslant [q + \exp(\beta h)]^{|\partial P| - 2|P|/(d-1)} \times (q^{1/d} + \exp(\beta h))^{2|P|/(d-1)}. \tag{2.15}$$

Taking into account that the statistical sum for any graph is the product of the corresponding sums for its connected components, we conclude that it is sufficient to prove that, for any connected ∂P,

$$q^{-1} \sum_{\sigma_{V(\partial P)}} \exp\left(\beta h \sum_l \delta_{\sigma_l \sigma_j}\right) \leqslant [q + \exp(\beta h)]^{|\partial P| - 2|P|/(d-1)} \times [q^{1/d} + \exp(\beta h)]^{2|P|/(d-1)}. \tag{2.16}$$

It is easy to verify (2.16) if the number of independent loops, $|h(\partial P)|$, is not less than $2|P|/(d-1)$. Indeed, we can then estimate the left-hand side of (2.16) by the similar method as in the case (a). We choose an arbitrary Cayley tree T in ∂P and write 1 instead of any $\delta_{\sigma_i \sigma_j}$ if i and j are the end points of the link $l \notin T$. Then the resulting statistical sum on T can be easily calculated, and we get

$$q^{-1} \sum_{\sigma_{V(\partial P)}} \exp\left(\beta h \sum_l \delta_{\sigma_l \sigma_j}\right) \leqslant [q + \exp(\beta h)]^{|\partial P| - 2|P|/(d-1)} \exp(\beta h |h(\partial P)|) \leqslant [q + \exp(\beta h)]^{|\partial P| - 2|P|/(d-1)} \times [q^{1/d} + \exp(\beta h)]^{2|P|/(d-1)}. \tag{2.17}$$

Let us now assume that $|h(\partial P)| < 2|P|/(d-1)$. It is also shown in the Appendix (Lemma A1) that $|h(\partial P)| \geqslant 2|P|/d$. For simplicity, we assume first that $2|P|/d$ and $2|P|/(d-1)$ are integers. Then

$$|h(\partial P)| = 2|P|/(d-1) - k, \quad k \text{ an integer} > 0,$$
$$2|P|/(d-1) - 2|P|/d = n, \quad n \geqslant k, \tag{2.18}$$
$$2|P|/(d-1) = nd, \quad |h(\partial P)| + k = nd.$$

Let us order the set of lattice sites in Λ lexicographically:

$$i \leqslant j \text{ whenever } j = i + \sum_{\mu=0}^{d-1} l_\mu \hat{\mu},$$

and the first nonvanishing l_μ is positive.

We choose the Cayley tree T in ∂P in the following way. For any site $i \in \partial P$, we connect with the minimal site j such that $j > i$ and the link (i, j) belongs to ∂P. Now, it is shown in the Appendix (Lemma A2) that there exist k sites such that they are the smaller with respect to the lexicographic ordering end points of exactly d links. We call these sites $i_1, \ldots, i_k$. Let L_L be the set of links having i_L as the smaller end point. We also let

$$G_2 = T \setminus \left(T \cap \left(\bigcup_{\alpha=1}^{k} L_\alpha\right)\right)$$

and $G_1 = \partial P \setminus G_2$. Then

$$G_1 = \bigcup_{\alpha=1}^{k} L_\alpha \cup L,$$

where $L \subset G_1$ is the set of all links such that their end points belong to G_2.

The L_α's and L are pairwise disjoint. To estimate the left-hand side of the formula (2.16) let us first write 1 instead of $\delta_{\sigma_i \sigma_j}$ for all links $l \in L$. Then we can estimate the resulting statistical sum on $\partial P \setminus L$ in the following way. We take the smallest site i belonging to ∂P and fix the spins on other sites. Then either there is exactly one line $l \in \partial P \setminus L \cap T$ connecting the site i to the rest of the graph, or $i = i_1$. In the former case we get the factor $q + \exp(\beta h)$, and in the latter, the factor $[q^{1/d} + \exp(\beta h)]^d$ (Lemma A3) multiplying the statistical sum for the graph obtained from ∂P by deleting the site i and all links from ∂P having i as their end point. Proceeding in this way we obtain finally the following bound:

$$q^{-1} \sum_{\sigma_{V(\partial P)}} \exp\left(\beta h \sum_l \delta_{\sigma_l \sigma_j}\right) \leqslant [q + \exp(\beta h)]^{|G_2|} [q^{1/d} + \exp(\beta h)]^{kd} [\exp(\beta h)]^{|L|}. \tag{2.19}$$

But

$$|G_2| = |V(\partial P)| - 1 - k = |\partial P| - |h(\partial P)| - k = |\partial P| - (2|P|/d - 1),$$
$$kd + |L| = 2|P|/d - 1,$$

so that we again get the inequality (2.16). One can easily verify that (2.16) is also true in the case when $2|P|/d - 1$ and/or $2|P|/d$ are not integers. This can be done by taking the appropriate integer parts.

From the inequalities (2.12) and (2.15) we find

$$\sum_{\sigma_\Lambda : p\in P \Rightarrow \sigma_p = 0} \exp\left(\beta h \sum_l \delta_{\sigma_l,0}\right)$$
$$\leqslant [q + \exp(\beta h)]^{|\Lambda|d - 2|P|/(d-1)} \times [q^{1/d} + \exp(\beta h)]^{2|P|/(d-1)}. \quad (2.20)$$

Finally, let us consider the sets P that are not "contractible" to a point in the sense defined above. Let us choose any Cayley tree T in ∂P. According to our assumption, there are links $l\in T$ such that the unique loop in $\{l\}\cup T$ is not a boundary of any oriented two-dimensional surface consisting of the plaquettes belonging to P. We call the set of such links Δ. We call two links $l_1, l_2\in\Delta$ equivalent if some linear combination of the loops in $\{l_i\}\cup T$ is the boundary of two-dimensional oriented surface consisting of plaquettes from P. The general solution of the equations $\sigma_p = 0$ for $p\in\partial P$ now reads

$$\sigma_l = \sigma_i - \sigma_j + \sigma_l^0. \quad (2.21)$$

Here $\sigma_l^0 \neq 0$ if and only if $l\in\Delta$; moreover, $\sigma_{l_1}^0 = \sigma_{l_2}^0$ if l_1 and l_2 are equivalent.

Corresponding, we have

$$\sum_{\sigma_{\partial P}: p\in P\Rightarrow\sigma_p=0} \exp\left(\beta h \sum_{l\in\partial P}\delta_{\sigma_l,0}\right) = \sum_{\{\sigma_l^0\}} q^{-c(\partial P)} \sum_{\sigma_{V(\partial P)}} \exp\left(\beta h \sum_{l\in\partial P}\delta_{\sigma_l^0,\sigma_j-\sigma_i}\right)$$
$$l\in\Delta:\sigma_{l_1}^0=\sigma_{l_2}^0 \text{ if } l_1,l_2 \text{ are equivalent}$$
$$\leqslant \sum_{\substack{\{\sigma_l^0\}\\ l\in\Delta}} q^{-c(\partial P)} \sum_{\sigma_{V(\partial P)}} \exp\left(\beta h \sum_{l\in\partial P}\delta_{\sigma_l^0,\sigma_j-\sigma_i}\right). \quad (2.22)$$

Now we can sum over all σ_l^0 using the fact that

$$\sum_{\sigma_l^0=0}^{q-1} \delta_{\sigma_l^0,\sigma_j-\sigma_i} = 1.$$

The right-hand side of the inequality (2.22) then becomes

$$q^{-c(\partial P)} \exp(\beta h|\Delta|) \sum_{\sigma_{V(\partial P\setminus\Delta)}} \exp\left(\beta h \sum_{l\in\partial P}\delta_{\sigma_l,\sigma_j}\right). \quad (2.23)$$

The last sum can be estimated as previously. Indeed, the graph obtained from ∂P by deleting the lines from Δ is the boundary $\partial P'$ of the subset $P'\subset P$ plus some lines that are parts of the borders of plaquettes from $P\setminus P'$ and do not form the additional loops. It can be immediately shown that the same estimate applies

$$q^{-c(\partial P)} \sum_{\sigma_{V(\partial P\setminus\Delta)}} \exp\left(\beta h \sum_{l\in\partial P}\delta_{\sigma_l,\sigma_j}\right) \leqslant [q + \exp(\beta h)]^{|\partial P\setminus\Delta| - 2|P'|/(d-1)} \times [q^{1/d} + \exp(\beta h)]^{2|P'|/(d-1)}. \quad (2.24)$$

Letting $2/(d-1)\,(|P| - |P'|) = \gamma$ and writing

$$\exp(\beta h|\Delta|) = \exp[\beta h(|\Delta| - \gamma)]\exp(\beta h\gamma) \leqslant [q + \exp(\beta h)]^{|\Delta| - [2/(d-1)](|P| - |P'|)} \times [q^{1/d} + \exp(\beta h)]^{2/(d-1)(|P| - |P'|)}.$$

We conclude from Eqs. (2.22)–(2.24) that the estimate (2.20) applies again.

Finally, by combining (2.20) with (2.10), we get

$$Z(\beta,h,q) \leqslant \sum_{|P|=0}^{|\Delta|d(d-1)/2} \binom{|\Delta|d(d-1)/2}{|P|} \times [e^\beta(q^{1/d} + e^{\beta h})^{2/(d-1)}]^{|P|} \times [(q + e^{\beta h})^{2/d-1}]^{|\Lambda|d(d-1)/2 - |P|}$$
$$= [e^\beta(q^{1/d} + e^{\beta h})^{2/(d-1)} + (q + e^{\beta h})^{2/d-1}]^{|\Lambda|d(d-1)/2}, \quad (2.25)$$

and

$$f(\beta,h,q) \geqslant -(d(d-1)/2\beta)\log[e^\beta(q^{1/d} + e^{\beta h})^{2/(d-1)} + (q + e^{\beta h})^{2/(d-1)}]. \quad (2.26)$$

Again after rescaling $\beta\to\beta\log q$ and letting $q\to\infty$ we get the needed lower bound. For $h<0$, we estimate the statistical sum by $Z(\beta,0,q)$ and use the above result. This concludes the proof.

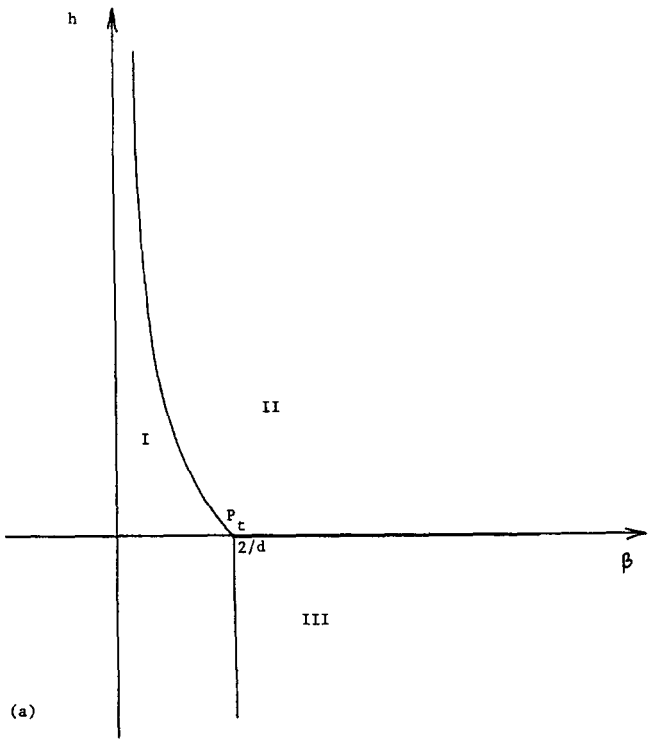
III. CONCLUSIONS AND SUMMARY

(i) As it is well known, the phase transition occurs as a result of a competition between the energy and entropy in expression for free energy $F = E - TS$. For low temperatures T, the energy is important and has to be minimalized. In the "ferromagnetic" case, this favors an order. On the other hand, for high temperatures the second term is the dominating one and one has to maximalize the entropy. This fact, in turn, favors nonordered states. In the limit $q\to\infty$, the competition between energy and entropy appears in its extreme form. For the usual (nongauge) Potts model and for the gauge Potts model in the version (a) there are two phases. In the low-temperature phase, the free energy simply equals the energy of the lowest lying state. All excited states are irrelevant and give no contribution. On the other hand, in the high-temperature phase, the energy factor is irrelevant and all states are equally probable. Consequently, the free energy simply equals T multiplied by total entropy.

The gauge Potts model in version (b) above is a bit more complicated. There are three phases. For a strong magnetic field (or for low temperatures), the free energy is dominated by the energy of the ground state (vacuum). This is the maximally ordered state with all $\sigma_l = 0$. As the magnetic field decreases (or the temperature increases), the phase transition occurs. In the new phase we meet the following situation. There are many states with the same plaquette energy as the vacuum state. These are simply the configurations gauge equivalent to the vacuum. Their energies are a bit greater than the ground state energy due to the coupling with the external field h. However, contrary to the ground state, which is unique, their number is quite large, $\sim q^{|\Lambda|}$, and, being macroscopically significant, it does contribute to the entropy per site by 1. In the phase under consideration this effect prevails the one following from the energy difference.

There exists also a third phase, the high-temperature one, in which the energy differences are again irrelevant and the free energy equals $-T$ multiplied by the total entropy.

(ii) From the consideration of Sec. II, the following interesting picture of the phase transition in case (b) emerges. For low temperatures, all the plaquette degrees of freedom are frozen to their vacuum value, $\sigma_p = 0$. They contribute an



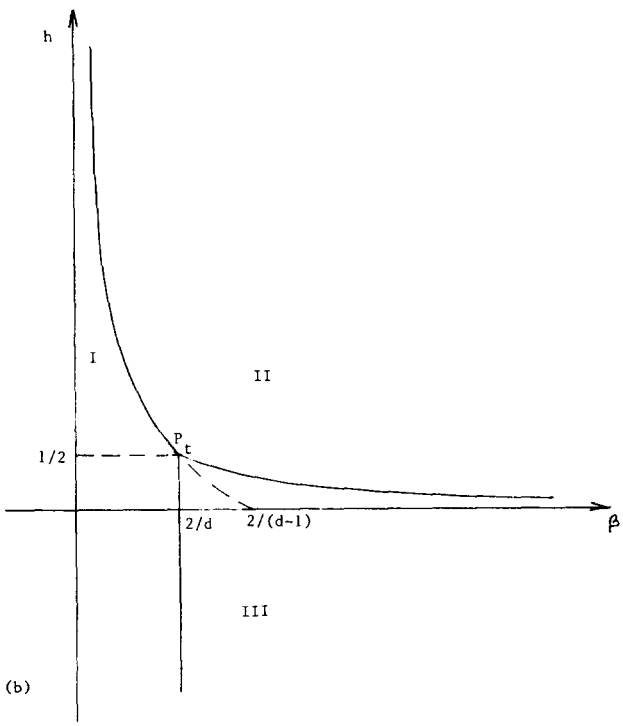


FIG. 1. (a) The phase diagram for version (a) of the gauge Potts model: $f_{\mathrm{I}} = -(d-1)/\beta$, $f_{\mathrm{II}} = -d(d-1)/2 - [h(d-1)]$, $f_{\mathrm{III}} = -d(d-1)/2$. (b) The phase diagram for version (b) of the gauge Potts model: $f_{\mathrm{I}} = -d/\beta, f_{\mathrm{II}} = -d(d-1)/2 - hd, f_{\mathrm{III}} = -d(d-1)/2 - (1/\beta)$.

amount $-d(d-1)/2$ to the free energy per site. Only the gauge degrees of freedom are unfrozen and, due to the coupling to the external field, they form a usual (nongauge) Potts model with the coupling constant equal to h and with the vanishing external field. Consequently, according to the results obtained in Ref. 7, those gauge degrees of freedom contribute an amount $-hd$ or $-1/\beta$ to the free energy in the low- or high-temperature phase, respectively. This gives the first two terms in the expression for the free energy. The third term is, as was stressed above, simply proportional to the total entropy of the system.

(iii) The phase diagrams for both cases are sketched on Fig. 1. We see that they are [especially in case (b)] fairly nontrivial. They both possess the triple point P_t; $P_t = (2/d,0)$ or $P_t = (2/d,1/2)$ in case (a) or (b), respectively. All of the phase transitions are of the first order. Let us also note that (a) shows a spontaneous magnetization for $\beta > 2/d$ and $h \to 0^+$; the second case cannot magnetize spontaneously.

(iv) Let us conclude our considerations by the remark concerning the applicability of the mean-field method. As we indicate in the Introduction, the mean-field method in its wider sense, gives the proper answer for both cases. However, if by the mean-field method we also understand the specific choice of the trial probability distribution that treats all link variables as independent, then we will fail to give the right answer for phase III of the second model. The reason for this is quite obvious. In this phase the correlations following from the gauge invariance play the dominant role. On the other hand, they are disregarded in the mean field theory in its narrow sense.

Note added in proof: Professor R. Kotecky kindly pointed out to me that there is an error in the proof starting from formula (2.22). This error can be corrected at the price of considerable lengthening of the proof. Meanwhile, I discovered that the whole proof can be simplified if we use the set $\bar{L}(P)$ constructed by Kotecky[8] and consider two cases: (i) $|\bar{L}(P)| \geqslant 2|P|/(d-1)$, (ii) $2|P|/(d-1) > |\bar{L}(P)| \geqslant 2|P|/d$. The corrected version appeared in the preprint and can be sent on request. No statement or conclusion should be changed.

ACKNOWLEDGMENT

I would like to thank Professor Kosiński for many important comments and stimulating discussions.

APPENDIX: SUBSIDIARY LEMMAS

Lemma A1: Let P be any set of plaquettes on the d-dimensional hypercubic lattice, and let ∂P be the graph spanned by the links bordering the plaquettes from P. Then the following inequality holds:

$$|h(\partial P)| \geqslant (2/d)|P|.$$

Proof: Instead of giving an independent proof we may appeal to the results obtained by Kotecky.[8] From his construction of the set $\bar{L}(P)$ in the notation used in Ref. 8, it immediately follows that $\bar{L}(P) \subset \partial P$, the graph $\partial P \setminus \bar{L}(P)$ has the same number of connected components as ∂P, and contains all vertices of ∂P. Consequently,

$$|\bar{L}(P)| = |\partial P| - |\partial P \setminus \bar{L}(P)|$$
$$\leqslant |\partial P| - |V(\partial P)| + C(\partial P) = |h(\partial P)|,$$

but Kotecky proved that $|\bar{L}(P)| \geqslant (2/d)|P|$.

Lemma A2: In the notation of Lemma A1, the following statement is valid. If

$$|h(\partial P)| = 2|P|/(d-1) - k,$$

then there exist at least k vertices such that they are the smaller (with respect to the lexicographical ordering) end points of exactly d links belonging to ∂P.

Proof: By inspecting again the construction given by Kotecky, we conclude that if the numbers of links of ∂P "emanating" from any vertex does not exceed $d-1$, then

$$|h(\partial P)| \geqslant 2|P|/(d-1).$$

Let us further note that if we delete one link belonging to ∂P, and start from the site that is the smaller end point of exactly d links, then $|h(\partial P)|$ decreases by 1 while $|P|$ decreases at most by $2(d-1)$. This fact concludes the proof.

Lemma A3: Let $\sigma_1,\ldots,\sigma_d$ be any fixed values from Z_q. Then the following inequality holds:

$$\sum_{\sigma \in Z_q} \exp\left(\beta h \sum_{i=1}^{d} \delta_{\sigma,\sigma_i}\right) \leqslant (q^{1/d} + \exp(\beta h))^d.$$

The proof is straightforward.

[1] C. J. Thompson, *Mathematical Statistical Mechanics* (Macmillan, New York, 1972), Appendix C.
[2] C. J. Thompson and H. Silver, Commun. Math. Phys. **33**, 53 (1971).
[3] P. A. Pearce and C. J. Thompson, Commun. Math. Phys. **41**, 191 (1975).
[4] A. Cant and P. A. Pearce, Commun. Math. Phys. **90**, 313 (1983).
[5] C. J. Thompson, Commun. Math. Phys. **36**, 225 (1974).
[6] P. A. Pearce and C. J. Thompson, Commun. Math. Phys. **58**, 131 (1978).
[7] P. A. Pearce and R. B. Griffiths, J. Phys. A **13**, 2143 (1980); P. A. Pearce, Physica A **125**, 247 (1984).
[8] R. Kotecky, Commun. Math. Phys. **82**, 391 (1981).

On relativistic irreducible quantum fields fulfilling CCR

Klaus Baumann[a)]
Department of Physics, Princeton University, Princeton, New Jersey 08544

(Received 12 August 1986; accepted for publication 1 October 1986)

Let ϕ be a relativistic scalar field fulfilling canonical commutation relations (CCR). Furthermore it is assumed that the time zero fields and momenta form an irreducible set. Based on estimates given by Herbst [I. W. Herbst, J. Math. Phys. **17**, 1210 (1976)], and by methods developed by Powers [R. T. Powers, Commun. Math. Phys. **4**, 145 (1967)], it is shown that ϕ has to be a free field in $n > 3$ space dimensions. For $n = 3$ (resp. $n = 2$) restrictions that look similar to the restriction in a formal $:\phi^4:_{3+1}$ (resp. $:\phi^6:_{2+1}$) theory are obtained.

I. INTRODUCTION

In 1967 Powers[1] showed that a relativistic Fermi field, which fulfills canonical anticommutation relations (CAR) and is irreducible, does not interact (it is even a free field[2]) if the number of space dimensions is greater than 1.

We want to analyze the analogous situation for Bose fields. We consider a relativistic scalar field ϕ which fulfills canonical commutation relations (CCR). Furthermore, we assume that the time zero field $\phi(f)$ together with the time zero momenta $\pi(f) = \dot{\phi}(f)$ form an irreducible set. Bose fields are represented by unbounded operators, but Fröhlich's commutator theorem[3] provides a powerful tool for dealing with this unboundedness.

In the case of fermions CAR imply the bound $\|\psi(t,f)\| \leqslant \|f\|_2$ for the field operator. Powers[1] uses locality and these operator bounds to show that

$$[\psi(f)\{\psi(g),\dot{\psi}(h)\}] = 0 \text{ for } n > 1 \text{ space dimensions.}$$

In the Bose case we also use locality but instead of $\|\psi(t,f)\| \leqslant \|f\|_2$ we use bounds for n-point functions given by Herbst[4] to estimate $\|[\pi(f)[\pi(g),\dot{\pi}(h)]]\Omega\|$.

Based upon "$\partial_j\,\phi$-bounds" we show in Sec. II that $[\pi(g_N)[\cdots[\pi(g_1),\dot{\pi}(h)]\cdots]\Omega = 0$ for $n > (N+3)/(N-1)$ space dimensions. From irreducibility we conclude in Sec. III that ϕ has to be a free field in $n > 3$ space dimensions. In $n = 3$ (resp. $n = 2$) space dimensions we get restrictions on multiple commutators $[\pi(g_N)[\cdots[\pi(g_1),\dot{\pi}(h)]\cdots]$ which are similar to the restrictions one would expect in a formal $:\phi^4:_{3+1}$ (resp. $:\phi^6:_{2+1}$) theory.

These results have to be compared with the known models: $P(\phi)_{1+1}$ fulfills CCR and so does the sine-Gordon model. In two-space dimensions $:\phi^4:_{2+1}$ is a canonical theory. The models with exponential interaction $:\exp\alpha\phi:$ are very interesting (see Ref. 5). In two or more space dimensions the regularized Schwinger functions converge to the Schwinger functions of a free field. In one-space dimension the same phenomenon occurs for large $|\alpha|$. For $\alpha^2 < 4\pi$ we have nontrivial Schwinger functions. In all cases the Wightman functions fulfill CCR.

In the references we have also listed the pioneering work by Araki[6] because it is fundamental for the estimates given by Herbst.[4] Furthermore, we want to mention the unpublished thesis by Sinha,[7] in which he tried to adapt the methods developed by Powers[1] to the Bose case. But at this time the very strong estimates given by Herbst[4] were not yet available and Sinha had to replace these estimates by assumptions.

Throughout this paper we make the following assumptions.

(i) Relativistic quantum field theory (QFT).

(a) $\phi(t,x)$ fulfills the Wightman axioms for a scalar, neutral field in $n+1$ space-time dimensions. The self-adjoint Hamiltonian $H \geqslant 0$ is the generator of time translations. We assume $(\Omega,\phi(t,x)\Omega) \equiv 0$.

(b) Existence of sharp time fields. For $f \in \mathscr{S}(\mathbb{R}^n)$ and $\psi \in D$ we assume

$$\phi(t,f)\psi = \lim_{\epsilon\to 0} \phi(\delta_t^\epsilon, f)\psi, \tag{1.1}$$

and

$$\pi(t,f)\psi = i[H,\phi(t,f)]\psi = \lim_{\epsilon\to 0} \phi(-\dot{\delta}_t^\epsilon, f)\psi \tag{1.2}$$

exist, where $\delta_t^\epsilon \in \mathscr{S}(\mathbb{R})$ is a δ-sequence and D is an invariant domain, i.e., $\phi(t,f)D \subset D$ and $\pi(t,f)D \subset D$. By $\phi(f)$ [resp. $\pi(f)$] we denote the time zero fields $\phi(0,f)$ [resp. $\pi(0,f)$].

(c) We assume that

$$\mathscr{D}_0 = \text{linear span } \langle e^{i\phi(f)}\Omega; f \in \mathscr{S}_{\text{real}}(\mathbb{R}^n)\rangle$$

is dense in the Hilbert space $\mathscr{H}$ and furthermore is a core for the Hamiltonian H.

In the following we assume always real test functions!

Remark 1:1: Assumption (c) is necessary because many estimates are based on Araki's formula[6]

$$(e^{i\phi(f)}\Omega, He^{i\phi(g)}\Omega) = \tfrac{1}{2}(f,g)(e^{i\phi(f)}\Omega, e^{i\phi(g)}\Omega), \tag{1.3}$$

and we want H to be uniquely determined by (1.3).

(ii) Canonical commutation relations (CCR). For the symmetric operators $\phi(t,f)$ and $\pi(t,f)$ we have (a) form bounds,

$$\begin{aligned} &\pm\phi(t,f) \leqslant |f|_1(H+1), \\ &\pm\pi(t,f) \leqslant |f|_2(H+1) \text{ as forms on } Q(H), \end{aligned} \tag{1.4}$$

[a)] Address after September 1986: Institut für Theoretische Physik, Bunsenstrasse 9, D-34 Göttingen, Federal Republic of Germany.

 0022-2488/87/030697-08$02.50

where $|\cdot|_1$ and $|\cdot|_2$ are norms on $\mathscr{S}(\mathbb{R}^n)$; and

$$\text{(b)}\quad [\phi(t,f),\phi(t,g)]=0=[\pi(t,f),\pi(t,g)], \tag{1.5}$$
$$[\phi(t,f),\pi(t,g)]=i\int_{\mathbb{R}^n} d^n x f(x)g(x)$$

weakly on $D(H)\times D(H)$.

Remark 1.2: By Fröhlich's commutator theorem[3] the CCR for the Weyl operators $e^{i\phi(t,f)}$ and $e^{i\pi(t,g)}$ follow from (a) and (b).

(iii) Irreducibility.

A bounded operator B which commutes with $e^{i\phi(f)}$ and $e^{i\pi(g)}$ for all $f,g\in\mathscr{S}(\mathbb{R}^n)$ (remember we take only real test functions!) is a c-number, i.e., $B=(\Omega,B\Omega)$. Again by Fröhlich's commutator theorem we can reformulate irreducibility as follows.

Proposition 1.3: Assume the symmetric operator C fulfills

$$\text{(a) the form bound } \pm C\leqslant c_0(H+1) \text{ on } Q(H), \tag{1.6}$$

and

$$\text{(b)}\quad [\phi(f),C]=0=[\pi(f),C] \text{ on } D(H)\times D(H) \text{ for all } f\in\mathscr{S}(\mathbb{R}^n), \tag{1.7}$$

then $C=(\Omega,C\Omega)$.

Proof: See Ref. 3 for details. Because of the form bounds (1.3) and (1.6) it follows from (1.7) that

$$e^{i\phi(f)}e^{i\lambda C}e^{-i\phi(f)}=e^{i\lambda C}, \tag{1.8}$$
$$e^{i\pi(f)}e^{i\lambda C}e^{-i\pi(f)}=e^{i\lambda C} \quad \text{for } \lambda\in\mathbb{R}. \tag{1.9}$$

From irreducibility we conclude

$$e^{i\lambda C}=(\Omega,e^{i\lambda C}\Omega), \tag{1.10}$$

and from this $C=(\Omega,C\Omega)$ follows. This formulation of irreducibility is much more convenient for the applications even if we have to assume a form bound (1.6).

(iv) Existence of $\dot{\pi}(t,f)$.

(a) For $f\in\mathscr{S}(\mathbb{R}^n)$ and $\psi\in D$,

$$\dot{\pi}(t,f)\psi=i[H,\pi(t,f)]\psi=\lim_{\epsilon\to 0}\phi(\ddot{\delta}_t^{\epsilon},f)\psi \tag{1.11}$$

exists and D is an invariant domain.

(b) We have the form bound

$$\pm\dot{\pi}(t,f)\leqslant|f|_3(H+1) \quad \text{on } Q(H), \tag{1.12}$$

where $|\cdot|_3$ is a norm on $\mathscr{S}(\mathbb{R}^n)$.

Remark 1.4: The assumed existence of $\phi(f)\Omega,\pi(f)\Omega$, and $\dot{\pi}(f)\Omega$ can be expressed by the Källen–Lehmann weight function $\rho(m^2)$ as follows:

$$\|\phi(f)\Omega\|^2=\int_0^\infty dm^2\rho(m^2)\int_{\mathbb{R}^n} d^n p\,\frac{|\tilde{f}(p)|^2}{2\sqrt{m^2+p^2}}<\infty, \tag{1.13}$$

$$\|\pi(f)\Omega\|^2=\int_0^\infty dm^2\rho(m^2)\int_{\mathbb{R}^n} d^n p\,\frac{\sqrt{m^2+p^2}}{2}|\tilde{f}(p)|^2<\infty, \tag{1.14}$$

$$\|\dot{\pi}(f)\Omega\|^2=\int_0^\infty dm^2\rho(m^2)\int_{\mathbb{R}^n} d^n p\,\frac{[m^2+p^2]^{3/2}}{2}|\tilde{f}(p)|^2<\infty. \tag{1.15}$$

Therefore the moments $\int_0^\infty dm^2\rho(m^2)m^k$ exist for $k=0,1,2,3$.

For the vacuum expectation value of the commutator we have the representation

$$(\Omega,[\phi(t,x),\phi(s,y)]\Omega)=-i\int_0^\infty dm^2\rho(m^2)D(t-s,x-y;m^2), \tag{1.16}$$

where

$$(\Box+m^2)D(t,x;m^2)=0, \tag{1.17}$$

with the initial conditions

$$D(0,x;m^2)=0,(\partial_t D)(0,x;m^2)=\delta(x). \tag{1.18}$$

The canonical commutation relations imply

$$\int_0^\infty dm^2\rho(m^2)=1. \tag{1.19}$$

II. ESTIMATES FOR MULTIPLE COMMUTATORS

In this section we want to get estimates for

$$[(\pi(g_N)[\pi(g_{N-1})[\cdots[\pi(g_1),\dot{\pi}(h)]\cdots]\Omega.$$

Our main tool is the combination of methods developed by Powers[1] for analyzing CAR with estimates given by Herbst[4] for a Wightman field fulfilling CCR.

From the paper by Herbst[4] one can easily extract the following estimates.

Proposition 2.1: Let $\zeta_k=\xi_k+i\eta_k\in\mathbb{R}+i\mathbb{R}_+$, $k=1,\ldots,N$, and let $f_k\in\mathscr{S}(\mathbb{R}^n)$, $k=1,\ldots,N$. Then the analytic vector-valued function

$$F_N(\underline{\zeta};(\mathbf{a}\cdot\nabla)f):=e^{i\zeta_1 H}\phi(\mathbf{a}\nabla f_1)e^{i\zeta_2 H}\cdots e^{i\zeta_N H}\phi(\mathbf{a}\nabla f_N)\Omega, \tag{2.1}$$

where $\mathbf{a}\in\mathbb{R}^n$ is a unit vector, is bounded by

$$\|F_N(\underline{\zeta};(\mathbf{a}\cdot\nabla)f)\|\leqslant A^N(N!)^{1/2}\left\{1+\max_{k=2,\ldots,N}\left(\frac{|\xi_k|}{\eta_k}\right)^{(N-1)/2}\right\}\times\prod_{k=1}^N\left(\lambda\|(\mathbf{a}\cdot\nabla)f_k\|_2+\frac{1}{\lambda}\|f_k\|_2\right), \tag{2.2}$$

for all $\lambda>0$.

Sketch of the proof: The starting point is the "$\nabla\phi$" bounds (see Theorem 2.6 in Ref. 4)

$$\pm\phi(\mathbf{a}\cdot\nabla f)\leqslant H+\tfrac{1}{2}\|f\|_2^2. \tag{2.3}$$

From this the estimates for n-point functions [see Theorem 2.5 (B) in Ref. 4],

$$(\Omega,|\phi(\mathbf{a}\nabla f)|^N\Omega)\leqslant B^N(N!)^{1/2}(4\|\mathbf{a}\cdot\nabla f\|_2\|f\|_2)^{N/2}, \tag{2.4}$$

can be easily derived. But of course

$$2(\|\mathbf{a}\nabla f\|_2\|f\|_2)^{1/2}\leqslant\lambda\|\mathbf{a}\nabla f\|_2+(1/\lambda)\|f\|_2 \tag{2.5}$$

for all $\lambda > 0$ and therefore we have

$$|(\Omega,\phi(\mathbf{a}\nabla f)^N\Omega)| \leqslant B^N(N!)^{1/2}[\lambda\|\mathbf{a}\nabla f\|_2 + (1/\lambda)\|f\|_2]^N. \tag{2.6}$$

As explained by Herbst[4] after the proof of Proposition 3.7 from the bound (2.6) the estimates (2.2) follow. This is so, because e^{-tH} is a positivity preserving contraction semigroup and therefore the Schwinger functions exist. Furthermore, one can do the analytic continuation in a similar way as Herbst[4] does it in the proof of Proposition 3.7.

The main result of this section is the following lemma.

Lemma 2.2: For $g_0,g_1,\ldots,g_N \in \mathscr{D}(\mathbb{R}^n)$ we have

$$[\pi(g_N)[\pi(g_{N-1})\cdots[\pi(g_1),\dot\pi(g_0)]\cdots]\Omega \equiv 0 \text{ if } n > (N+3)/(N-1). \tag{2.7}$$

Proof: (a) As we shall see later we only need the cases $N = 2, 3, 4$, and 6. The corresponding values of n, the number of space dimensions for which (2.7) is true, are $n \geqslant 6, 4, 3$, and 2.

(b) Remember that $\pi(g_k) = \dot\phi(0,g_k)$ and $\dot\pi(g_0) = \ddot\phi(0,g_0)$. By using δ-sequences $f_0^\epsilon, f_1^\epsilon, \ldots, f_N^\epsilon \in \mathscr{D}(\mathbb{R})$ we write

$$\begin{aligned}[\pi(g_N)[\cdots[\pi(g_1),\dot\pi(g_0)]\cdots]\Omega &= \lim_{\epsilon\to 0}(-1)^N[\phi(\dot f_N^\epsilon,g_N) \\ &\times[\cdots[\phi(\dot f_1^\epsilon,g_1),\phi(\ddot f_0^\epsilon,g_0)]\cdots]\Omega.\end{aligned} \tag{2.8}$$

Let us choose $f_j \in \mathscr{D}(\mathbb{R}), j = 0,\ldots,N$, such that

$$\text{(i) } \operatorname{supp} f_j \subseteq \left[-\frac{1}{5N},\frac{1}{5N}\right], \quad \text{(ii) } \int_{\mathbb{R}} f_j(t)dt = 1, \tag{2.9}$$

and for $\epsilon > 0$ we define

$$\text{(iii) } f_j^\epsilon(t) := (1/\epsilon) f_j(t/\epsilon). \tag{2.10}$$

(c) For $0 < \epsilon \leqslant 1$ let $E_{\underline{k}}^\epsilon$, $\underline{k} \in \mathbb{Z}^n$, be a smooth partition of the unity as defined in Appendix A. For $\underline{k} = (k_1,\ldots,k_n) \in \mathbb{Z}^n$ we have

$$\begin{aligned}\operatorname{supp} E_{\underline{k}}^\epsilon \subseteq [(k_1 - \tfrac{3}{4})\epsilon,(k_1+\tfrac{3}{4})\epsilon] \\ \times\cdots\times[(k_n-\tfrac{3}{4})\epsilon,(k_n+\tfrac{3}{4})\epsilon].\end{aligned} \tag{2.11}$$

By linearity we get

$$\begin{aligned}[\phi(\dot f_N^\epsilon,g_N)[\cdots[\phi(\dot f_1^\epsilon,g_1),\phi(\ddot f_0^\epsilon,g_0)]\cdots]\Omega \\ = \sum_{\underline{k}(N),\ldots,\underline{k}(0)\in\mathbb{Z}^n}[\phi(\dot f_N^\epsilon,E_{\underline{k}(N)}^\epsilon g_N)[\cdots[\phi(\dot f_1^\epsilon, \\ E_{\underline{k}(1)}^\epsilon g_1),\phi(\ddot f_0^\epsilon,E_{\underline{k}(0)}^\epsilon g_0)]\cdots]\Omega.\end{aligned} \tag{2.12}$$

This is a finite sum because $g_0,\ldots,g_N \in \mathscr{D}(\mathbb{R}^N)$.

(d) As Powers[1] has done in the case of CAR we now use locality to reduce the number of terms. The support of $f_j^\epsilon \times E_{\underline{k}}^\epsilon g_j$ is by construction contained in

$$\begin{aligned}O_{\underline{k}} := \left[-\frac{\epsilon}{5N},\frac{\epsilon}{5N}\right]\times\left[\left(k_1-\frac{3}{4}\right)\epsilon,\left(k_1+\frac{3}{4}\right)\epsilon\right] \\ \times\cdots\times\left[\left(k_n-\frac{3}{4}\right)\epsilon,\left(k_n+\frac{3}{4}\right)\epsilon\right].\end{aligned} \tag{2.13}$$

This $O_{\underline{k}}$ is spacelike separated from $O_{\underline{l}}$ if $|k_i - l_i| \geqslant 2$ for some $i \in \{1,\ldots,n\}$, because if we assume $k_i \geqslant l_i + 2$, then for $x \in O_{\underline{k}}$ and $y \in O_{\underline{l}}$ we have

$$\begin{aligned}(x-y)^2 &\leqslant ((2/5N)\epsilon)^2 - [(k_i - \tfrac{3}{4}) - (l_i + \tfrac{3}{4}))]^2 \\ &\leqslant \epsilon^2[(2/5N)^2 - (\tfrac{1}{2})^2] < 0.\end{aligned} \tag{2.14}$$

Therefore we get by locality

$$\begin{aligned}[\phi(\dot f_N^\epsilon,g_N)[\cdots[\phi(\dot f_1^\epsilon,g_1),\phi(\ddot f_0^\epsilon,g_0)]\cdots]\Omega \\ = \sum_{\underline{k}(N),\ldots,\underline{k}(0)\in\mathbb{Z}^n}[\phi(\dot f_N^\epsilon,E_{\underline{k}(N)}^\epsilon g_N) \\ \times[\cdots[\phi(\dot f_1^\epsilon,E_{\underline{k}(1)}^\epsilon g_1),\phi(\ddot f_0^\epsilon,E_{\underline{k}(0)}^\epsilon g_0)]\cdots]\Omega,\end{aligned} \tag{2.15}$$

and $|(\underline{k}(j) - \underline{k}(0))_i| \leqslant 1$ for all $i = 1,\ldots,n$ and $j = 1,\ldots,N$. Let L denote the sidelength of a cube in $\mathbb{R}^n$ containing $\operatorname{supp} g_0$. Then the number of N-fold commutators appearing on the rhs of Eq. (2.15) is at most $3^{nN}((L+2)/\epsilon)^n$.

(e) Proposition 2.3, which we shall prove later, provides a bound for each of these N-fold commutators. Therefore we have the estimate

$$\begin{aligned}\|[\phi(\dot f_N^\epsilon,g_N)[\cdots[\phi(\dot f_1^\epsilon,g_1),\phi(\ddot f_0^\epsilon,g_0)]\cdots]\Omega\| \\ \leqslant 3^{nN}\left(\frac{L+2}{\epsilon}\right)^n D(n,N)\left(\prod_{j=0}^N \max|g_j|\right) \\ \times\max\left|\frac{d^{K+2}f_0}{dt^{K+2}}\right| \\ \times\left(\prod_{j=1}^N \max\left|\frac{d^{K+1}f_j}{dt^{K+1}}\right|\right)\epsilon^{(1/2)[n(N+1)-N-3]}\end{aligned} \tag{2.16}$$

$$\begin{aligned}= C(n,N)(L+2)^n\epsilon^{[(N-1)/2](n-(N+3)/(N-1))} \\ \times\prod_{j=0}^N \max|g_j|\max|f_0^{(K+2)}|\prod_{j=1}^N \max|f_j^{(K+1)}|,\end{aligned} \tag{2.17}$$

where $C(n,N)$ is a constant and $K = [N/2]$ ($\hat{=}$ smallest natural number greater than $N/2$). As ϵ goes to zero Lemma 2.2 follows from (2.17).

Proposition 2.3: For $g_0,\ldots,g_N \in \mathscr{D}(\mathbb{R}^n)$ and $f_0,\ldots,f_N \in \mathscr{D}([-1/5N,1/5N])$ we have

$$\begin{aligned}\|[\phi(\dot f_N^\epsilon,E_{\underline{k}(N)}^\epsilon g_N)[\cdots[\phi(\dot f_1^\epsilon,E_{\underline{k}(1)}^\epsilon g_1), \\ \phi(\ddot f_0^\epsilon,E_{\underline{k}(0)}^\epsilon g_0)]\cdots]\Omega\| \\ \leqslant D(n,N)\prod_{j=0}^N \max|g_j|\max|f_0^{(K+2)}| \\ \times\prod_{j=1}^N \max|f_j^{(K+1)}|\epsilon^{(1/2)[n(N+1)-N-3]}\end{aligned} \tag{2.18}$$

provided $|(\underline{k}(j) - \underline{k}(0))_1| \leqslant 1$ for all $j = 1,2,\ldots,N$, where $D(n,N)$ is a constant and $K = [N/2]$ ($\hat{=}$ smallest natural number greater than $N/2$).

Proof: (a) In a first step we replace the test functions $E_{\underline{k}(j)}^\epsilon g_j$ by test functions $\partial_1 h_{j,\underline{k}(j)}^\epsilon$ and we show that because of locality this does not affect the N-fold commutator. The idea is the following. Consider the commutator $[\phi(s,f),\phi(t,g)]$, where $f,g \in \mathscr{D}(\mathbb{R}^n)$. Define $f_a(x) := f(x-a)$ (translation by $a \in \mathbb{R}^n$). For $a \in \mathbb{R}^n$ large enough we have

$$[\phi(s,f),\phi(t,g)] = [\phi(s,f-f_a),\phi(t,g-g_{-a})]$$

by locality. If we take $a \in \mathbb{R}^n$ parallel to the x_1 direction and

define

$$F(x) := \int_{-\infty}^{x_1} [f(y,x_2,\ldots,x_n) - f(y-a,x_2,\ldots,x_n)]\,dy,$$

then $F \in \mathscr{D}(\mathbb{R}^n)$ and $f - f_a = \partial_1 F$. With a similar definition for G we finally get

$$[\phi(s,f),\phi(t,g)] = [\phi(s,\partial_1 F),\phi(t,\partial_1 G)].$$

(b) To apply the above idea to our problem we define

$$h^{\epsilon}_{0,\underline{k}(0)}(x) := \int_{-\infty}^{x_1} dy\{(E^{\epsilon}_{\underline{k}(0)} g_0)(y,x_2,\ldots,x_n) - (E^{\epsilon}_{\underline{k}(0)} g_0)(y+3\epsilon,x_2,\ldots,x_n)\} \quad (2.19)$$

and for $j = 1,\ldots,N$

$$h^{\epsilon}_{j,\underline{k}(j)}(x) := \int_{-\infty}^{x_1} dy\{(E^{\epsilon}_{\underline{k}(j)} g_j)(y,x_2,\ldots,x_n) - (E^{\epsilon}_{\underline{k}(j)} g_j)(y-3\epsilon,x_2,\ldots,x_n)\}. \quad (2.20)$$

From this definition we get with $a := 3\epsilon e_1$

$$\partial_1 h^{\epsilon}_{0,\underline{k}(0)} = E^{\epsilon}_{\underline{k}(0)} g_0 - (E^{\epsilon}_{\underline{k}(0)} g_0)_{-a}, \quad (2.21)$$

and for $j = 1,\ldots,N$

$$\partial_1 h^{\epsilon}_{j,\underline{k}(j)} = E^{\epsilon}_{\underline{k}(j)} g_j - (E^{\epsilon}_{\underline{k}(j)} g_j)_a. \quad (2.22)$$

Because of the constraint $|(\underline{k}(j) - \underline{k}(0))_1| \leqslant 1$, $j = 1,\ldots,N$, our choice of $a = 3\epsilon e_1$, the support properties of $f^{\epsilon}_j \times E^{\epsilon}_{\underline{k}(j)}$, $f^{\epsilon}_j \times (E^{\epsilon}_{\underline{k}(j)})_a$, and $f^{\epsilon}_0 \times (E^{\epsilon}_{\underline{k}(0)})_{-a}$ [see (2.13)], and locality we get

$$\begin{aligned}&[\phi(\dot f^{\epsilon}_N, E^{\epsilon}_{\underline{k}(N)} g_N)[\cdots[\phi(\dot f^{\epsilon}_1, E^{\epsilon}_{\underline{k}(1)} g_1),\\ &\quad \phi(\ddot f^{\epsilon}_0, E^{\epsilon}_{\underline{k}(0)} g_0)]\cdots]\Omega\\ &= [\phi(\dot f^{\epsilon}_N, \partial_1 h^{\epsilon}_{N,\underline{k}(N)})[\cdots[\phi(\dot f^{\epsilon}_1, \partial_1 h^{\epsilon}_{1,\underline{k}(1)}),\\ &\quad \phi(\ddot f^{\epsilon}_0, \partial_1 h^{\epsilon}_{0,\underline{k}(0)})]\cdots]\Omega. \end{aligned} \quad (2.23)$$

(c) For later use let us estimate the L_2-norms $\|\partial_1 h^{\epsilon}_{j,\underline{k}(j)}\|_2$ and $\|h^{\epsilon}_{j,\underline{k}(j)}\|_2$,

$$\begin{aligned}&\|\partial_1 h^{\epsilon}_{j,\underline{k}(j)}\|_2\\ &= \|E^{\epsilon}_{\underline{k}(j)} g_j - (E^{\epsilon}_{\underline{k}(j)} g_j)_a\|_2 \leqslant \sqrt{2}\|E^{\epsilon}_{\underline{k}(j)} g_j\|_2\\ &\leqslant \sqrt{2}\max|g_j|\,\|E^{\epsilon}_{\underline{k}(j)}\|_2 \leqslant \sqrt{2}\max|g_j|\epsilon^{n/2}, \end{aligned} \quad (2.24)$$

$$\begin{aligned}|h^{\epsilon}_{j,\underline{k}(j)}(x)| &\leqslant \int_{-\infty}^{+\infty} dy|E^{\epsilon}_{\underline{k}(j)} g_j(y,x_2,\ldots,x_n)|\\ &\leqslant \max|g_j|\int_{-\infty}^{+\infty} dy\left|E\left(\frac{y}{\epsilon}\right)\right| \leqslant \max|g_j|\epsilon. \end{aligned} \quad (2.25)$$

And because the support of $h^{\epsilon}_{j,\underline{k}(j)}$ with respect to x_1 is contained in $[(k_1(j) - \frac{3}{4})\epsilon, (k_1(j) + 3 + \frac{3}{4})\epsilon]$ we have

$$\|h^{\epsilon}_{j,\underline{k}(j)}\|_2 \leqslant (\tfrac{9}{2})^{1/2}\max|g_j|\epsilon^{(n/2)+1}. \quad (2.26)$$

These estimates are true for $j = 1,\ldots,N$ and also for $j = 0$ if we replace $a = 3\epsilon e_1$ by $-3\epsilon e_1$.

(d) The second step consists in using Proposition 2.1 to estimate the rhs of Eq. (2.23). The N-fold commutator is built up by 2^N terms, each of which is the boundary value of an analytic vector-valued function. Consider for example the first term

$$\begin{aligned}&\phi(\dot f^{\epsilon}_N, \partial_1 h^{\epsilon}_{N,\underline{k}(N)})\cdots\phi(\dot f^{\epsilon}_1, \partial_1 h^{\epsilon}_{1,\underline{k}(1)})\phi(\ddot f^{\epsilon}_0, \partial_1 h^{\epsilon}_{0,\underline{k}(0)})\Omega\\ &= \int_{\mathbb{R}^{N+1}} d^{N+1}t\,\dot f^{\epsilon}_N(t_N)\cdots\ddot f^{\epsilon}_0(t_0), \end{aligned} \quad (2.27)$$

$$\begin{aligned}&e^{itH}\phi(\partial_1 h^{\epsilon}_{N,\underline{k}(N)})e^{i(t_{N-1}-t_N)H}\cdots e^{i(t_0-t_1)H}\phi(\partial_1 h^{\epsilon}_{0,\underline{k}(0)})\Omega\\ &= \int_{\mathbb{R}^{N+1}} d^{N+1}\xi f^{\epsilon}(\underline{\xi})\\ &\quad \times F_{N+1}(\underline{\xi} + i\underline{0};\partial_1 h^{\epsilon}_{N,\underline{k}(N)},\ldots,\partial_1 h^{\epsilon}_{0,\underline{k}(0)}), \end{aligned} \quad (2.28)$$

where we have introduced new coordinates

$$\underline{\xi} = (\xi_1,\ldots,\xi_{N+1}) = (t_N, t_{N-1} - t_N,\ldots,t_0 - t_1), \quad (2.29)$$

and defined the function $f^{\epsilon}(\underline{\xi}) \in (\mathbb{R}^{N+1})$ as

$$f^{\epsilon}(\underline{\xi}) = \dot f^{\epsilon}_N(\xi_1)\dot f^{\epsilon}_{N-1}(\xi_1+\xi_2)\cdots\ddot f^{\epsilon}_0(\xi_1+\cdots\xi_{N+1}). \quad (2.30)$$

From (2.10) we derive the following scaling behavior:

$$f^{\epsilon}(\underline{\xi}) = \epsilon^{-(N+2)}\epsilon^{-(N+1)} f((1/\epsilon)\underline{\xi}), \quad (2.31)$$

with

$$f(\underline{\xi}) := \dot f_N(\xi_1)\dot f_{N-1}(\xi_1+\xi_2)\cdots\ddot f_0(\xi_1+\cdots+\xi_{N+1}). \quad (2.32)$$

(e) Now we are able to use Proposition 2.1. We claim that

$$\begin{aligned}&\left|\left|\int_{\mathbb{R}^{N+1}} d^{N+1}\xi f^{\epsilon}(\underline{\xi}) F_{N+1}(\underline{\xi} + i\underline{0};\partial_1 h^{\epsilon}_{N,\underline{k}(N)},\ldots,\partial_1 h^{\epsilon}_{0,\underline{k}(0)})\right|\right|\\ &\leqslant D\prod_{j=0}^{N}\max|g_j|\max|f_0^{(K+2)}|\\ &\quad \times\prod_{j=1}^{N}\max|f_j^{(K+1)}|\epsilon^{(1/2)(N+1)(n+1)-(N+2)}, \end{aligned} \quad (2.33)$$

where D is a constant (depending on n and N) and $K = [N/2]$. From estimate (2.2) we know that for any $\lambda > 0$

$$\begin{aligned}&\|F_{N+1}(\underline{\xi} + i\underline{\eta};\partial_1 h^{\epsilon}_{N,\underline{k}(N)},\ldots,\partial_1 h^{\epsilon}_{0,\underline{k}(0)})\|\\ &\leqslant A^{N+1}[(N+1)!]^{1/2}\\ &\quad \times\prod_{j=0}^{N}\left(\lambda\|\partial_1 h^{\epsilon}_{j,\underline{k}(j)}\|_2 + \frac{1}{\lambda}\|h^{\epsilon}_{j,\underline{k}(j)}\|_2\right)\\ &\quad \times\left\{1 + \max_{j=2,\ldots,N+1}\left(\frac{|\xi_j|}{\eta_j}\right)^{N/2}\right\}. \end{aligned} \quad (2.34)$$

From the estimate (2.24) and (2.26) we get for the special choice $\lambda = \epsilon^{1/2}$

$$\begin{aligned}&\lambda\|\partial_1 h^{\epsilon}_{j,\underline{k}(j)}\|_2 + \frac{1}{\lambda}\|h^{\epsilon}_{j,\underline{k}(j)}\|_2\\ &\leqslant \max|g_j|\epsilon^{(n+1)/2}[2^{1/2} + (\tfrac{9}{2})^{1/2}]\\ &\leqslant (\tfrac{25}{2})^{1/2}\max|g_j|\epsilon^{(n+1)/2}, \end{aligned} \quad (2.35)$$

and this explains the factor $\Pi_{j=0}^{N}\max|g_j|\epsilon^{(1/2)(N+1)(n+1)}$ in (2.33). Finally we have to estimate

$$\int_{\mathbb{R}^{N+1}} d^{N+1}\xi f^{\epsilon}(\underline{\xi}) F_{N+1}(\underline{\xi} + i\underline{0}),$$

where

$$f^{\epsilon} \in \mathscr{D}(\mathbb{R}^{N+1}),$$

$$\|F_{N+1}(\underline{\xi} + i\underline{\eta})\| \leqslant C\left\{1 + \sum_{j=1}^{N+1}\left(\frac{|\xi_j|}{\eta_j}\right)^{N/2}\right\}. \quad (2.36)$$

In Appendix B we derive under the above assumption the

bound

$$\left\| \int_{\mathbb{R}^{N+1}} d^{N+1}\xi f^{\epsilon}(\xi) F_{N+1}(\xi + i\underline{0}) \right\|$$
$$\leqslant C \sum_{l=0}^{K} \int_{\mathbb{R}^{N+1}} d^{N+1}\xi \, |(\hat{\eta}\cdot\underline{\partial}_{\xi})^{l} f^{\epsilon}(\xi)|$$
$$\times \left\{ 1 + \sum_{j=1}^{N+1} \left(\frac{|\xi_j|}{\hat{\eta}_j} \right)^{N/2} \right\}, \tag{2.37}$$

where $\hat{\eta} \in \mathbb{R}_{+}^{N+1}$ can be any point and $K = [N/2]$ denotes the smallest natural number greater than $N/2$. If we take $\hat{\eta} := \epsilon(1,...,1) = \epsilon \cdot \underline{1}$ and remember that

$$f^{\epsilon}(\xi) = \epsilon^{-(N+2)} \epsilon^{-(N+1)} f((1/\epsilon)\xi), \tag{2.31}$$

we get

$$\left\| \int_{\mathbb{R}^{N+1}} d^{N+1}\xi f^{\epsilon}(\xi) F_{N+1}(\xi + i\underline{0}) \right\|$$
$$\leqslant C\epsilon^{-(N+2)} \sum_{l=0}^{K} \int_{\mathbb{R}^{N+1}} d^{N+1}\xi \, |(\underline{1}\cdot\underline{\partial}_{\xi})^{l} f(\xi)|$$
$$\times \left\{ 1 + \sum_{j=1}^{N+1} |\xi_j|^{N/2} \right\}. \tag{2.38}$$

Because

$$f(\xi) = \dot{f}_N(\xi_1)\dot{f}_{N-1}(\xi_1+\xi_2)\cdots\dot{f}_0(\xi_1+\cdots+\xi_{N+1})$$

has compact support the rhs of (2.38) is bounded by

$$\leqslant \hat{C}\cdot\epsilon^{-(N+2)} \max|f_0^{(K+2)}| \prod_{j=1}^{N} \max|f_j^{(K+1)}|. \tag{2.39}$$

This proves the bound (2.33) and for a suitable $\hat{C}$ the estimate (2.39) applies for each of the 2^N terms building up the N-fold commutator on the rhs of Eq. (2.23). This proves Proposition 2.3.

Remark 2.4: Instead of Proposition 2.1 we could have used the following estimate (see Proposition 3.7 in Ref. 4):

$$F_N(\underline{\xi};\underline{f}) \leqslant A^N N! \prod_{k=1}^{N} \|f_k\|_2 \left\{ 1 + \max_{k=2,...,N} \left(\frac{|\xi_k|}{\eta_k} \right)^{N-1} \right\}, \tag{2.40}$$

which is valid under the additional assumption of a mass gap but for all $f_k \in \mathscr{S}(\mathbb{R}^n)$ and not only for f_k's which are derivatives of test functions, i.e., $f_k = (\mathbf{a}\nabla)g_k, g_k \in \mathscr{S}(\mathbb{R}^n)$. But (2.40) is weaker than (2.2) and we would only get

$$[\pi(g_N)[\cdots[\pi(g_1),\dot{\pi}(g_0)]\cdots]\Omega$$
$$\equiv 0 \text{ if } n > 2(N+2)/(N-1), \tag{2.41}$$

instead of the stronger result (2.7).

III. IRREDUCIBLE FIELDS FULFILLING CCR ARE FREE IN $n \geqslant 4$ SPACE DIMENSIONS

Now we shall show how we can use Lemma 2.2 to prove that irreducible fields fulfilling CCR are free fields in $n \geqslant 4$ space dimensions. First let us consider multiple commutators which also contain a time zero field $\phi(f)$.

Lemma 3.1: For $f, g_0, ..., g_N \in \mathscr{S}(\mathbb{R}^n)$, $N \in \mathbb{N}$, and all $\psi \in D$ we have

$$[\phi(f), \dot{\pi}(g_0)]\psi = 0, \tag{3.1a}$$

and

$$[\phi(f)[\pi(g_N)[\cdots[\pi(g_1),\dot{\pi}(g_0)]\cdots]\psi = 0. \tag{3.1b}$$

Proof: From CCR and the existence of $\dot{\pi}(g_0)$ it follows that

$$\left(\frac{d}{dt}[\phi(t,f),\pi(t,g_0)]\right)\psi$$
$$= 0, \quad \text{for } \psi \in D$$
$$= \underbrace{[\pi(t,f),\pi(t,g_0)]}_{\equiv 0}\psi + [\phi(t,f),\dot{\pi}(t,g_0)]\psi. \tag{3.2}$$

This proves (3.1a). If we use the Jacobi identity

$$[\phi(f)[\pi(g),A]]\psi = [\pi(g)[\phi(f),A]]\psi$$
$$- \underbrace{[A[\phi(f),\pi(g)]]}_{\equiv 0}\psi, \tag{3.3}$$

and take A successively as $\dot{\pi}(g_0), [\pi(g_1), \dot{\pi}(g_0)], ..., [\pi(g_{N-1})[\cdots[\pi(g_1),\dot{\pi}(g_0)]\cdots]$ we get (3.1b).

Let us define the following form bounds for N-fold commutators.

For $g_0, ..., g_N \in \mathscr{D}(\mathbb{R}^n)$ there is a constant $C_{g_N\cdots g_0}$ such that

$$(O_N): \quad \pm i^N[\pi(g_N)[\cdots[\pi(g_1),\dot{\pi}(g_0)]\cdots]$$
$$\leqslant C_{g_N\cdots g_0}(H+1), \tag{3.4}$$

as a quadratic form on $Q(H)$, where H is the Hamiltonian.

With these form bounds (O_N) we can exploit irreducibility via Fröhlich's commutator theorem[3] as follows.

Lemma 3.2: Assume for all $f, g_0, ..., g_N \in \mathscr{D}(\mathbb{R}^n)$

$$[\pi(f)[\pi(g_N)[\cdots[\pi(g_1),\dot{\pi}(g_0)]\cdots]\Omega = 0, \tag{3.5}$$

then the form bound (O_N), CCR, and irreducibility imply

$$[\pi(g_N)[\cdots[\pi(g_1),\dot{\pi}(g_0)]\cdots]$$
$$= (\Omega,[\pi(g_N)[\cdots[\pi(g_1),\dot{\pi}(g_0)]\cdots]\Omega). \tag{3.6}$$

Proof: For fixed test functions the $(N+1)$-fold commutator $[\pi(f)[\pi(g_N)[\cdots[\pi(g_1),\dot{\pi}(g_0)]\cdots]$ is an element of the algebra $P(O)$, $O \subset \mathbb{R}^{n+1}$ compact, and by the Reeh–Schlieder theorem (see Theorem 4.3 in Streater–Wightman[8]) we get from Eq. (3.5)

$$[\pi(f)[\pi(g_N),...,\dot{\pi}(g_0)]\cdots] = 0. \tag{3.7}$$

From Lemma 3.1 we know that

$$[\phi(f)[\pi(g_N),...,\dot{\pi}(g_0)]\cdots] = 0. \tag{3.8}$$

This remains true even for $f \in \mathscr{S}(\mathbb{R}^n)$ because as long as $g_0 \in \mathscr{D}(\mathbb{R}^n)$ locality acts as a cutoff for supp f. Because of the form bound (O_N) and the corresponding form bounds for $\phi(f)$ and $\pi(f)$ we conclude from irreducibility that (3.7) and (3.8) imply (3.6).

If we combine Lemma 2.2 with Lemma 3.2 for the case $N = 1$ we get that in more than five-space dimensions the commutator $[\pi(g),\dot{\pi}(h)]$ is a c-number, i.e., $[\pi(g),\dot{\pi}(h)] = (\Omega,[(g),\dot{\pi}(h)]\Omega)$. As shown in the following lemma irreducibility and the Källen–Lehmann representation imply that ϕ is a free field.

Lemma 3.3: Assume that for all $g, h \in \mathscr{D}(\mathbb{R}^n)$

$$[\pi(g),\dot{\pi}(h)] = (\Omega,[\pi(g),\dot{\pi}(h)]\Omega), \tag{3.9}$$

then irreducibility implies

$$\dot\pi(h) - \phi(\Delta h) + M^2\phi(h) = 0, \tag{3.10}$$

with

$$M^2 = \int_0^\infty dm^2\,\rho(m^2)m^2. \tag{3.11}$$

Proof: (a) From the Källen–Lehmann representation for two-point functions we know that

$$i(\Omega,[\pi(g),\dot\pi(h)]\Omega)$$

$$= \int_0^\infty dm^2\,\rho(m^2)\int_{\mathbb{R}^n} d^nx\,g(x)\{(\Delta h)(x) - m^2h(x)\} \tag{3.12}$$

$$= \int_{\mathbb{R}^n} d^nx\,g(x)(\Delta h)(x) - M^2\int_{\mathbb{R}^n} d^nx\,g(x)h(x), \tag{3.13}$$

because CCR implies $\int_0^\infty dm^2\,\rho(m^2) = 1$ and the existence of $\dot\pi(h)\Omega$ implies $\int_0^\infty dm^2\,\rho(m^2)m^2 = M^2 < \infty$. By the canonical commutation relations we can write the rhs of Eq. (3.13) as $i[\pi(g),\phi(\Delta h) - M^2\phi(h)]$ and therefore using (3.9) we end up with

$$[\pi(g),\dot\pi(h) - \phi(\Delta h) + M^2\phi(h)] = 0. \tag{3.14}$$

(b) Lemma 3.1 and CCR imply that also

$$[\phi(g),\dot\pi(h) - \phi(\Delta h) + M^2\phi(h)] = 0. \tag{3.15}$$

From the bounds for $\phi(f)$, $\pi(f)$, and $\dot\pi(f)$ and from irreducibility we conclude

$$\dot\pi(h) - \phi(\Delta h) + M^2\phi(h) = 0, \tag{3.10}$$

because $(\Omega,\phi(t,x)\Omega)\equiv 0$ by assumption. Furthermore, (3.10) shows that $\rho(m^2)$ has to be $\delta(m^2 - M^2)$. By continuity Eq. (3.10) can be extended to all $h\in\mathscr{S}(\mathbb{R}^n)$.

Remark 3.4: The above proof also shows that among all generalized free fields ϕ, which fulfill CCR and for which $\|\ddot\Phi(t,h)\Omega\| < \infty$, only fields of a definite mass M are irreducible.

If we combine Lemma 2.2, Lemma 3.2, and Lemma 3.3 we get the following as a first result.

Theorem 3.5: Under our general assumptions (i)–(iv), and the additional assumption of a form bound (O_1) for $i[\pi(g_1),\dot\pi(g_0)]$ we have that in $n > 5$ space dimensions ϕ is a free field.

To get a result only in more than five space dimensions is not quite satisfactory. How can we improve this? From Lemma 2.2 it follows that

$$[\pi(g_3)[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]]]\Omega = 0,$$

in $n > 3$ space dimensions. If we also assume a form bound (O_2) then Lemma 3.2 tells us that

$$[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]] = (\Omega,[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]]\Omega).$$

But as we shall show in the next lemma the positivity of the Hamiltonian H implies under the above circumstances that $[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]]$ vanishes.

Lemma 3.6: Assume that for $g_0, g_1, g_2\in\mathscr{D}(\mathbb{R}^n)$

$$[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]] = (\Omega,[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]]\Omega), \tag{3.16}$$

and assume the form bound (O_1) then

$$[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]]\equiv 0. \tag{3.17}$$

Proof: By the positivity of H we have for all $f\in\mathscr{D}(\mathbb{R}^n)$ and all $\lambda\in\mathbb{R}$

$$0\leqslant e^{i\lambda\pi(f)}He^{-i\lambda\pi(f)} \tag{3.18}$$

$$= H - \lambda\dot\pi(f) - (\lambda^2/2!)i[\pi(f),\dot\pi(f)]$$
$$- (\lambda^3/3!)i^2[\pi(f)[\pi(f),\dot\pi(f)]]. \tag{3.19}$$

All higher commutators vanish by (3.16) and because of the form bounds for $\dot\pi(f)$ and $i[\pi(f),\dot\pi(f)]$ we can apply Fröhlich's commutator theorem[3] to write (3.18) as a series. By taking the vacuum expectation value of the inequality (3.18) we get

$$0\leqslant -(\lambda^2/2!)(\Omega,i[\pi(f),\dot\pi(f)]\Omega)$$
$$- (\lambda^3/3!)(\Omega,i^2[\pi(f)[\pi(f),\dot\pi(f)]]\Omega), \tag{3.20}$$

because $H\Omega = 0$ and $(\Omega,\dot\pi(f)\Omega) = 0$. For this to be nonnegative for all $\lambda\in\mathbb{R}$ the coefficient of λ^3 has to vanish! If we now take

$$f = \mu_0g_0 + \mu_1g_1 + \mu_2g_2, \quad \mu_i\in\mathbb{R}, \tag{3.21}$$

and use the Jacobi identity we get

$$(\Omega,[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]]\Omega)\equiv 0, \tag{3.22}$$

and by (3.16) we have shown Lemma 3.6. This improves Theorem 3.5 as follows.

Theorem 3.7: Under the assumptions of Theorem 3.5 and assuming the additional form bound (O_2) for double commutators $i^2[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]]$ we have that ϕ is a free field in $n > 3$ space dimensions.

Proof: Follows from Lemma 2.2 for $N = 3$, Lemma 3.2, Lemma 3.6, again, Lemma 3.2, and finally Lemma 3.3.

What can be said about the remaining cases $n\leqslant 3$? From (2.7) it is obvious that always $n > 1$ no matter how large N is chosen. The following lemma shows how the interaction is restricted in two- and three-space dimensions.

Lemma 3.8: Under our general assumptions (i)–(iv) and assuming in addition the form bound (a) (O_3), we have for $n = 3$

$$[\pi(g_3)[\cdots[\pi(g_1),\dot\pi(g_0)]]]$$
$$= (\Omega,[\pi(g_3)[\cdots,\dot\pi(g_0)]]\Omega) \tag{3.23}$$
("ϕ^4 interaction");

or (b) (O_5), we have for $n = 2$

$$[\pi(g_5)[\cdots[\pi(g_1),\dot\pi(g_0)]\cdots]$$
$$= (\Omega,[\pi(g_5)[\cdots,\dot\pi(g_0)]\cdots]\Omega) \tag{3.24}$$
("ϕ^6 interaction").

Proof: Combine Lemma 2.2 with Lemma 3.2. With a little bit more effort we can conclude more, e.g., assuming also a form bound (O_2), we have for $n = 3$

$$i^2[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]]$$
$$= \lambda\phi(g_2g_1g_0) + (\Omega,i^2[\pi(g_2)[\pi(g_1),\dot\pi(g_0)]]\Omega)$$
$$= \lambda\phi(g_2g_1g_0) + \mu i[\pi(g_2),\phi(g_1g_0)], \tag{3.25}$$

and we can even derive bounds on $\lambda\geqslant 0$ if we explicity estimate all constants involved in the proof of Lemma 2.2. For going beyond (3.25) it is necessary to define operator prod-

ucts $:\phi^2:$ and $:\phi^3:$ but we did not succeed in solving this problem.

IV. DISCUSSION

First of all we think that the estimates of Proposition 2.1—extracted from Herbst's paper[4]—are very strong and therefore the final results are quite optimal. For example even if one assumes the bounds

$$\|\phi(t_N,f_N)\cdots\phi(t_1,f_1)\Omega\|\leqslant C_N\prod_{k=1}^{N}\|\phi(t_k,f_k)\Omega\|, \tag{4.1}$$

which are modeled after a free field theory and which look terribly restringent, one does not get better results than the ones listed in Sec. III.

Second to get rid of certain assumptions which were necessary because we dealt with unbounded field operators one should modify the method such that only Weyl operators $e^{i\phi(f)}$, $e^{i\pi(f)}$, and e^{iHt} enter in the proof. Or even better one should give a proof within Euclidean field theory, because Herbst[4] got his estimates by first constructing Euclidean fields.

Finally let us summarize our results as follows. In more than four space-time dimensions only free fields can fulfill CCR and irreducibility. In four space-time dimensions the same is true with the possible exception of a formal $:\phi^4:_{3+1}$ theory. In three space-time dimensions $:\phi^4:_{2+1}$ fulfills CCR and we ruled out any other but the $:\phi^6:_{2+1}$ interaction as possible candidates for CCR. In two space-time dimensions our analysis does not impose any restriction on the interaction. This fits nicely with the constructed models, because $P(\phi)_{1+1}$, sine–Gordon model, and exponential interaction all fulfill canonical commutation relations.

ACKNOWLEDGMENTS

It is a pleasure to thank Arthur Wightman for suggesting this problem and for his constant encouragement. Furthermore I want to thank Chris King and John Klauder for many helpful discussions and also Ira Herbst for explaining to me certain details of his paper.[4]

The work for this paper was supported by a grant from the Max Kade Foundation, New York.

APPENDIX A: A SMOOTH PARTITION OF THE UNITY

Take $\rho\in\mathscr{D}([-\frac{1}{4},\frac{1}{4}])$ such that

(i) $\rho(x)\geqslant 0$,

(ii) $\rho(-x)=\rho(x)$,

and

$$\text{(iii)}\ \int_{\mathbb{R}}\rho(x)dx=1. \tag{A1}$$

For $0<\epsilon\leqslant 1$ and $k\in\mathbb{Z}$ we define

$$E_k^\epsilon(x):=\int_{-\infty}^{x}\left\{\rho\left(\frac{y}{\epsilon}-k-\frac{1}{2}\right)-\rho\left(\frac{y}{\epsilon}-k+\frac{1}{2}\right)\right\}\frac{dy}{\epsilon}. \tag{A2}$$

This E_k^ϵ has the following properties:

$$\text{(i)}\ E_k^\epsilon\in\mathscr{D}(\mathbb{R}),\quad \operatorname{supp}E_k^\epsilon\subseteq[(k-\tfrac{3}{4})\epsilon,(k+\tfrac{3}{4})\epsilon], \tag{A3}$$

$$\text{(ii)}\ 0\leqslant E_k^\epsilon(x)\leqslant 1,\quad E_k^\epsilon(x)=1,\quad \text{for } x\in[(k-\tfrac{1}{4})\epsilon,(k+\tfrac{1}{4}\epsilon)], \tag{A4}$$

$$\text{(iii)}\ \sum_{k=K}^{L}E_k^\epsilon(x)=\begin{cases}1, & \text{for } x\in[(K-\frac{1}{4})\epsilon,(L+\frac{1}{4})\epsilon],\\ 0, & \text{for } x<(K-\frac{3}{4})\epsilon \text{ or } x>(L+\frac{3}{4})\epsilon,\end{cases} \tag{A5}$$

$$\text{(iv)}\ \sum_{k\in\mathbb{Z}}E_k^\epsilon(x)\equiv 1, \tag{A6}$$

$$\text{(v)}\ \int_{\mathbb{R}}|E_k^\epsilon(x)|^2dx\leqslant\int_{\mathbb{R}}E_k^\epsilon(x)dx=\epsilon. \tag{A7}$$

In $n>1$ dimensions we define for $0\leqslant\epsilon<1$ and $\underline{k}=(k_1,\ldots,k_n)\in\mathbb{Z}^n$,

$$E_{\underline{k}}^\epsilon(x):=E_{k_1}^\epsilon(x_1)\cdots E_{k_n}^\epsilon(x_n). \tag{A8}$$

We call $E_{\underline{k}}^\epsilon$ a smooth partition of the unity of width ϵ. We have the estimate

$$\|E_k^\epsilon\|_2\leqslant\epsilon^{n/2}. \tag{A9}$$

APPENDIX B: AN ESTIMATE FOR BOUNDARY VALUES

(a) Let $F(\underline{\zeta})$ be a function analytic in $\underline{\zeta}\in[\mathbb{R}+i(0,\infty)]^{\otimes n}$ which fulfills the estimate

$$|F(\underline{\xi}+i\underline{\eta})|\leqslant C\left\{1+\sum_{k=1}^{n}\left(\frac{|\xi_k|}{\eta_k}\right)^\alpha\right\},\quad 0<\alpha<\infty, \tag{B1}$$

therefore $F(\underline{\xi}+i\underline{0})$ is a distribution over $\mathscr{S}(\mathbb{R}^n)$.

(b) Define $K:=[\alpha]$ ($\hat{=}$ smallest natural number greater than α). We expand $F(\underline{\xi}+i\underline{\eta})$ in a Taylor series around $\underline{\xi}+i\underline{\hat{\eta}}$,

$$F(\underline{\xi}+i\underline{\eta})=\sum_{l=0}^{K-1}\frac{1}{l!}(i(\underline{\eta}-\underline{\eta})\underline{\partial}_{\underline{\xi}})^lF(\underline{\xi}+i\underline{\hat{\eta}})+\frac{1}{(K-1)!}\int_0^1d\lambda(1-\lambda)^{K-1}(i(\underline{\eta}-\underline{\eta})\underline{\partial}_{\underline{\xi}})^K\times F(\underline{\xi}+i\underline{\eta}+\lambda i(\underline{\eta}-\underline{\eta})). \tag{B2}$$

If we integrate by parts we get for $f\in\mathscr{S}(\mathbb{R}^n)$

$$\int_{\mathbb{R}^n}f(\underline{\xi})F(\underline{\xi}+i\underline{\eta})d^n\xi=\sum_{l=0}^{K-1}\frac{1}{l!}\int_{\mathbb{R}^n}F(\underline{\xi}+i\underline{\eta})(-i(\underline{\eta}-\underline{\eta})\underline{\partial}_{\underline{\xi}})^lf(\underline{\xi})d^n\xi+\frac{1}{(K-1)!}\int_0^1d\lambda(1-\lambda)^{K-1}\int_{\mathbb{R}^n}F(\underline{\xi}+i\underline{\eta}+\lambda i(\underline{\eta}-\underline{\eta}))(-i(\underline{\eta}-\underline{\eta})\underline{\partial}_{\underline{\xi}})^Kf(\underline{\xi})d^n\xi. \tag{B3}$$

(c) From estimate (B1) we get

$$\left| \int_{\mathbf{R}^n} f(\underline{\xi}) F(\underline{\xi} + i\underline{\eta}) d^n\xi \right|$$
$$\leqslant \sum_{l=0}^{K-1} \frac{c}{l!} \int_{\mathbf{R}^n} \left\{ 1 + \sum_{k=1}^{N} \left(\frac{|\xi|_k}{\hat{\eta}_k} \right)^{\alpha} \right\}$$
$$\times |((\underline{\eta} - \underline{\eta})\partial_{\underline{\xi}})^l f(\underline{\xi})| d^n\xi + \frac{c}{(K-1)!}$$
$$\times \int_{\mathbf{R}^n} |((\underline{\eta} - \underline{\eta})\partial_{\underline{\xi}})^K f(\underline{\xi})| \int_0^1 d\lambda (1-\lambda)^{K-1}$$
$$\times \left\{ 1 + \sum_k \left(\frac{|\xi_k|}{\hat{\eta}_k + \lambda(\eta_k - \hat{\eta}_k)} \right)^{\alpha} \right\} d^n\xi. \tag{B4}$$

If we take $0 < \eta_k \leqslant \hat{\eta}_k$, $k = 1,...,n$, then we have the estimate

$$\int_0^1 d\lambda (1-\lambda)^{K-1} \left(\frac{|\xi|}{\hat{\eta} + \lambda(\eta - \hat{\eta})} \right)^{\alpha}$$
$$= \left(\frac{|\xi|}{\hat{\eta}} \right)^{\alpha} \int_0^1 d\lambda \frac{(1-\lambda)^{K-1}}{[1 - \lambda(1 - (\eta/\hat{\eta}))]^{\alpha}}$$
$$\leqslant \left(\frac{|\xi|}{\hat{\eta}} \right)^{\alpha} \int_0^1 \frac{d\lambda}{(1-\lambda)^{\alpha - K + 1}} \leqslant \frac{1}{[\alpha] - \alpha} \left(\frac{|\xi|}{\hat{\eta}} \right)^{\alpha}. \tag{B5}$$

Therefore the last term in Eq. (B4) is bounded by

$$\frac{C}{K!} \int_{\mathbf{R}^n} |((\underline{\eta} - \hat{\eta})\partial_{\underline{\xi}})^K f(\underline{\xi})|$$
$$\times \left\{ 1 + \frac{K}{K - \alpha} \sum_{k=1}^{n} \left(\frac{|\xi_k|}{\hat{\eta}_k} \right)^{\alpha} \right\} d^n\xi. \tag{B6}$$

Now we can put $\underline{\eta} \equiv 0$ and for any $\underline{\eta} \in \mathbf{R}^n_+$ we get the crude estimate

$$\left| \int_{\mathbf{R}^n} f(\underline{\xi}) F(\underline{\xi} + i\underline{0}) d^n\xi \right|$$
$$\leqslant \frac{c}{[\alpha] - \alpha} \sum_{l=0}^{[\alpha]} \int_{\mathbf{R}^n} |((\underline{\eta}\partial_{\underline{\xi}}))^l f(\underline{\xi})|$$
$$\times \left\{ 1 + \sum_{k=1}^{n} \left(\frac{|\xi_k|}{\hat{\eta}_k} \right)^{\alpha} \right\} d^n\xi. \tag{B7}$$

Only the derivatives of f up to order $[\alpha]$ enter in this estimate.

[1] R. T. Powers, "Absence of interaction as a consequence of good ultraviolet behavior in the case of a local Fermi field," Commun. Math. Phys. **4**, 145 (1967).

[2] K. Baumann, "Three remarks on Powers' theorem about irreducible fields fulfilling CAR," J. Math. Phys. **27**, 2373 (1986).

[3] J. Fröhlich, "Application of commutator theorems to the integration of representations of Lie algebras and commutation relations," Commun. Math. Phys. **54**, 135 (1977).

[4] I. W. Herbst, "On canonical quantum field theories," J. Math. Phys. **17**, 1210 (1976).

[5] S. Albeverio, G. Gallavotti, and R. Hoegh-Krohn, "Some results for the exponential interaction in two or more dimensions," Commun. Math. Phys. **70**, 187 (1979).

[6] H. Araki, "Hamiltonian formalism and the canonical commutation relations in quantum field theory," J. Math. Phys. **1**, 492 (1960).

[7] K. B. Sinha, Ph.D. thesis, University of Rochester, 1969 (unpublished).

[8] R. F. Streater and A. S. Wightman, PCT, *Spin and Statistics, and All That* (Benjamin, New York, 1964).

An action principle combining electromagnetism and general relativity

H. Gardner Moyer
Dutch Digital Systems, Post Office Box 476, Holbrook, New York 11741

(Received 18 August 1986; accepted for publication 5 November 1986)

The stationary action problem for a single, classical, point particle in external gravitational and electromagnetic fields is written in optimal control format. The relativistic interval is the independent variable and time, space, and action are the five dependent variables. A general metric is used for the space-time manifold so that the equations are manifestly covariant. The form of the system equations guarantees that the particle moves with unit speed with respect to interval. The Lagrangian is a function of the metric tensor and the electromagnetic four-potential, but not of particle parameters such as electric charge q and mass m. The Hamiltonian is not identically zero, unlike those derived in many earlier analyses. A constant of the motion is found that is identified with q/mc^2. An explanation is presented for the classical inequality $m \geqslant 0$. The trajectories can reduce to geodesics and even further to those governed by Fermat's principle of stationary time.

I. INTRODUCTION

The trajectory of a single, classical, point particle subject to gravity and electromagnetism will be derived using optimal control.[1] The five state variables will be time, space, and action. The independent variable will be the relativistic interval or proper time τ. All scalar products will be taken with a general metric so that the equations will be manifestly covariant. The simplicity of this problem will permit us to concentrate our attention on the new viewpoint and methods.

The four differential equations to be used here for the space-time variables x formulate the geodesic problem of general relativity. The form of these equations guarantees that the particle moves with unit speed with respect to proper time. The usual differential equation for the action s is easily put in a manifestly covariant form representing the electromagnetic interaction only. The system of five first-order equations differs from earlier formulations in that it does not contain particle parameters such as mass m and electric charge q. Each extremal represents a trade-off between maximized proper time and either maximized or minimized action. In the extreme case with the Lagrange multiplier $\lambda_s = 0$, τ is maximized with s open. The extremal is then a geodesic. We show that λ_s is the constant that experimental physicists call q and that the Hamiltonian $\mathscr{H}$ is the constant they call the self-energy mc^2. Our equations require mass to be non-negative.

The first-order differential equations for the state variables and Lagrange multipliers are used to derive the second-order accelerations. The writing of the latter equations is simplified by introducing the Christoffel symbols as abreviations. The usual postulating of these symbols as affine connections is unnecessary.

When $q = m = 0$, the trajectories reduce to lightlike paths governed by Fermat's principle of stationary time. The eikonals of geometric optics can be superposed. These surfaces bound Maxwell's four-dimensional (4-D) waves. Here we will show that when the values allowed for q,m are generalized, EM waves should be extended into a six-dimensional space. The wave equation itself is beyond the scope of the present paper, although the Hamilton–Jacobi equation for the wave fronts $\tau(x,s) = \text{const}$ is written.

Although the correct velocity and acceleration equations have been derived previously from manifestly covariant equations,[2–4] the Hamiltonians have been unsatisfactory in that they have been identically zero. This has made the Hamilton–Jacobi and wave equations difficult to derive. A vanishing Hamiltonian occurs whenever the system equations are homogeneous in the control variables to any power other than zero. A specific formulation that has been used previously will be discussed in Sec. VI after notation and methods have been established. Here the Hamiltonian will be homogeneous of degree 1 in the velocity with respect to proper time, but of degree zero in the control variables. Here $\mathscr{H}$ will equal the self-energy and will be equated to the usual function of the energy-momentum and potential four-vectors.

II. A SUMMARY OF OPTIMAL CONTROL

Optimal control will be used rather than the calculus of variations. The former is the more systematic, general, and powerful. It is the more tutorial because it provides a geometric picture of the optimization process. The problem is expressed as a system of n first-order differential equations with an initial manifold. The n-tuple of *state variables* x is composed of all the derivated quantities including action. The nonderivated variables u make up the *control variable m-tuple*. For each value of the independent variable (usually designated t), a boundary of the reachable set or *wave front* is defined in state variable space. By definition, no point outside the wave front can be reached by a trajectory that satisfies the inequality and differential constraints. The optimal trajectories or extremals terminate on the wave front.

The n-tuple λ is the *outward pointing normal* to the wave front. (Some authors prefer an inward pointing normal.) The elements of λ are known in the calculus of variations as the Lagrange multipliers canonically conjugate to the state variables. The magnitude of the normal n-tuple is irrelevant.

 0022-2488/87/030705-06$02.50

Thus our equations will be homogeneous in the elements of λ and only their ratios will affect the extremals.

The scalar product of dx/dt and the normal λ is called the *control Hamiltonian* $H(t,x,\lambda,u)$. If an extremal is to remain on the wave front as t increases, its control variables must be chosen to make the control Hamiltonian a maximum. This *maximum principle* is what physicists call Huygens' principle of wavelets and wave fronts. The optimal control approach is particularly appropriate for the least action problem because it turns the later introduction of quantum waves into a natural development. The optimal u is a function of t,x,λ and when this expression is substituted into the control Hamiltonian the *true Hamiltonian* $\mathscr{H}(t,x,\lambda)$ of the calculus of variations is obtained.

The calculation of extremals does not require the definition of either final boundary conditions or a performance functional to be optimized. A state variable is maximized, minimized, or open according to whether the corresponding element of the final (outward pointing) λ is positive, negative, or zero. When there are two state variables, an extremal that optimizes x^1 subject to a fixed final value of x^2 also solves the reciprocal problem of optimizing x^2 subject to x^1 fixed. Similar statements apply when there are more than two state variables.

The initial manifold is sometimes simply a point that can lie either inside or outside the wave fronts. In the former case there is no upper bound on the values of t defined by trajectories that terminate at a given final point. In the latter case extremals that provide the maximum value of t exist and determine the trailing edge of the wave front. The velocity and outward normal n-tuples form an acute angle on the leading edge ($\mathscr{H}>0$) and an obtuse angle on the trailing edge ($\mathscr{H}<0$). The latter will be much more important for the present problem so that the obtuse angle and negative $\mathscr{H}$ would be inconvenient, λ will therefore be defined to point inward. This will be accomplished by using the *minimum principle* rather than the maximum principle. These concepts are illustrated schematically in Fig. 1.

When an $(n-1)$-dimensional initial manifold is smooth, *transversality conditions* require the $\lambda(0)$ to point along its normals. If an outward pointing $\lambda(0)$ is at a corner of a manifold, it is qualified if it makes obtuse angles with the sides. Should the manifold degenerate to a point, $\lambda(0)$ would be unrestricted. When the minimum principle is being used, an inward pointing $\lambda(0)$ satisfies transversality if it points diametrically opposite to a qualified outward normal of the maximum principle.

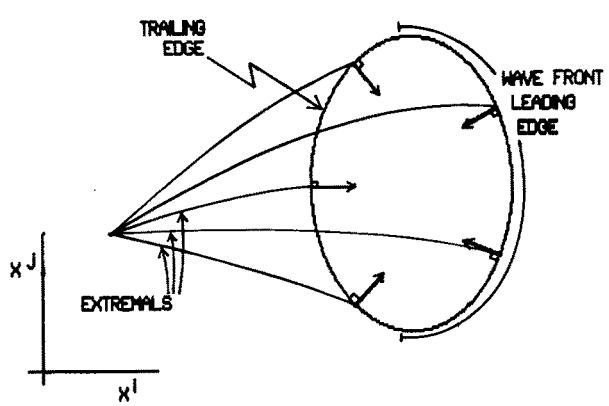


FIG. 1. Wave front with leading and trailing edges. Extremals and inward pointing normals are shown.

The optimal control equations reduce to those of the calculus of variations when λ is scaled so that its (constant) element associated with action is -1. Optimal control is more general because it allows the action to be either minimized or maximized. It can be used for those problems that have no multipliers that are constants of the motion. It is more powerful because it shows how to treat nondifferential equality and inequality constraints.

The phrase "minimum or saddle point" will be abreviated to "minimum." A minimizing (or maximizing) extremal becomes a saddle point after passing a Jacobi conjugate point. Its final point then lies in the interior of the wave front.

III. TRAJECTORIES SUBJECT TO ELECTROMAGNETISM ONLY

The least action problem for a particle with rest mass m and electric charge q will be formulated at first using the usual position-action space $(\mathbf{x},x^4)$ with $\mathscr{R}^4$ metric. (Three-tuples that transform under spatial rotation as vectors are written in boldface.) The independent variable is $x^0 \equiv ct$ — the speed of light multiplying time. The control variables are the three-tuple of dimensionless velocity $\mathbf{v}$. The Lagrangian L is a function of the electromagnetic scalar potential $A_0(x^0,\mathbf{x})$ and the vector potential $\mathbf{A}(x^0,\mathbf{x})$. Radiation reaction and gravity are neglected. The four state variable equations with the initial point when $x^0=0$ are then[5,6]

$$\frac{d\mathbf{x}}{dx^0}=\mathbf{v}, \quad \mathbf{x}(0)=\bar{\mathbf{x}}, \tag{3.1a}$$

$$\frac{dx^4}{dx^0}\equiv\frac{ds}{dx^0}\equiv L=-mc(1-\mathbf{v}^2)^{1/2} - \frac{q}{c}(A_0-\mathbf{A}\cdot\mathbf{v}), \quad s(0)=0. \tag{3.1b}$$

The scalar product of the normal four-tuple $\boldsymbol{\lambda},\lambda_s$ with $(\mathbf{v},L)$ is called the control Hamiltonian $H(x^0,\mathbf{x},\boldsymbol{\lambda},\lambda_s,\mathbf{v})$:

$$H=-mc\lambda_s(1-\mathbf{v}^2)^{1/2}-q\lambda_s A_0/c+(\boldsymbol{\lambda}+q\lambda_s\mathbf{A}/c)\cdot\mathbf{v}, \tag{3.2}$$

The control variables $\mathbf{v}$ appear nonlinearly and inhomogeneously. There are four state variables and three initial Lagrange multiplier ratios. Therefore a three-parameter family of extremals issues from the initial point. The natural choice for the parameters is $\mathbf{v}(0)$.

IV. TRAJECTORIES SUBJECT TO ELECTROMAGNETISM AND GRAVITY

In order to obtain space-time symmetric equations that take account of gravity, a new independent variable must be found. The relativistic interval or proper time τ will be introduced. It is defined using the symmetric, indefinite, covariant, metric tensor $\mathbf{G}$ which reduces to the Minkowski metric $\boldsymbol{\eta}$ for flat space. An arrow will indicate that the general metric is being specialized to that of Minkowski. An "M" will be written after the number of such equations. They may be

checked with the literature, but it should be noted that space-time is curved when the potential A is non-null,

$$g_{\mu\nu}(x^\rho)\equiv \mathbf{G}(x),\quad g^{\mu\nu}(x^\rho)\equiv \mathbf{G}^{-1}(x),$$

$$x^\rho = x,\quad \mu,\nu,\rho = 0,\ldots,3 \tag{4.1}$$

$$\rightarrow \boldsymbol{\eta} = \boldsymbol{\eta}^{-1}\equiv \mathrm{diag}[1,-1,-1,-1]. \tag{4.1M}$$

Since the metric tensor has been written in two forms, there will be two notations for the relativistic interval

$$\delta\tau = +(g_{\mu\nu}\delta x^\mu \delta x^\nu)^{1/2}\equiv +(\delta x, \mathbf{G}\delta x)^{1/2} \tag{4.2}$$

$$\rightarrow +[(\delta x^0)^2 - \delta \mathbf{x}^2]^{1/2}. \tag{4.2M}$$

Terms with repeated Greek indices are summed from 0 to 3. When scalar products are written as $(\cdot,\cdot)$, the first vector is contravariant and the second is covariant.

The system (3.1) can be generalized to the following five equations when x^0 is a dependent variable. All scalar products, including that in the Lagrangian, will be taken in terms of the metric tensor $\mathbf{G}$. The equations will be simplified by using the following abreviation for the magnitude of the control contravariant four-vector u,

$$|u|\equiv +(u, \mathbf{G}u)^{1/2}. \tag{4.3}$$

The first four state variable equations are

$$\frac{dx}{d\tau}\equiv \dot{x} = \frac{u}{|u|},\quad x(0)=\bar{x}. \tag{4.4}$$

The form of these equations guarantees that the nonholonomic constraint

$$(\dot{x}, \mathbf{G}\dot{x}) = 1 \tag{4.5}$$

implied by (4.2), will be satisfied for any choice of u including one that makes $|u|$ zero. They also require u to be timelike ($|u|$real). The nonlinearity in the control vector u will become significant when we look for the minimum of the control Hamiltonian. When $\mathbf{G}$ is positive definite, a well-posed *geodesic* problem is completely formulated by (4.4). A Lagrangian is not required. When $\mathbf{G}$ is indefinite, a control variable inequality constraint must be adjoined or the problem would be trivial. Therefore the restriction

$$u^0 > 0 \tag{4.4'}$$

will be imposed since classical trajectories always move forward in time. Although the usual relativistic metrics do not permit $\dot{x}^0$ to pass through zero continuously, it must be forbidden to jump from a positive to a negative value. The four dimensions do not appear symmetrically in this constraint —but time has already been given special treatment when a relativistic metric was chosen [see (4.1M)]. The problem of the previous section does not have a corresponding explicit constraint because the independent variable x^0 is tacitly assumed to be monotonically increasing. The necessity for this constraint will be clarified in the discussion of transversality below.

Equations (4.4) and (4.4') require u to point into the *future half-cone*. The latter is defined in 4-D space-time with its vertex at the particle's location $x(\tau)$ and with sides composed of null geodesics with $x^0 \geqslant x^0(\tau)$. It follows that (4.4') is preserved by Lorentz transformations.

To derive a manifestly covariant Lagrangian, we first multiply (3.1b) by $dx^0/d\tau$:

$$\dot{s} = \frac{-mc[(dx^0)^2 - d\mathbf{x}]^{1/2}}{d\tau} - \frac{q}{c}(A_0\dot{x}^0 - \mathbf{A}\cdot\dot{\mathbf{x}}). \tag{4.6M}$$

The four-potential will be defined in the covariant form $A=[A_0, -\mathbf{A}]$. After substituting from (4.2M) and (4.4), and generalizing the metric, this equation becomes

$$\dot{s}\equiv L = -mc - (q/c)(u, A)/|u|. \tag{4.6}$$

The symmetry and unity of idea of this Lagrangian are marred by its first term. Since mc is a constant, Ockham's razor can be applied. The relation can then be further simplified by absorbing the multiplicative constant q/c into the action s,

$$\dot{x}^4\equiv \dot{s}\equiv L = -(u, A)/|u|,\quad s(0)=0. \tag{4.4''}$$

The five-tuple of state variables is composed of the contravariant four-vector x and the scalar $x^4\equiv s$. The dimension of s is now charge.

The system (4.4) is remarkable in that it does not contain constants such as mass, electric charge, or magnetic charge. Later a constant of the motion will be found that will be recognized as being the q/mc^2 of experimental physics. It will vary from trajectory to trajectory.

When the initial manifold is a point, transversality conditions permit all directions for $\lambda(0), \lambda_s$. However, contrary to (4.4), it is a mistake to consider the present initial manifold as a point or even a surface. Both $\dot{x}$ and $\dot{s}$ are unbounded when u points in a null direction. Any point that can be reached via a combination of such lightlike subarcs should be considered to be part of the initial manifold. This indicates that those $\lambda(0), \lambda_s$ that generate lightlike extremals are at the limit of the family of multipliers that satisfy transversality. All points of state variable space could be reached in this way had we not imposed inequality (4.4').

As τ increases, (4.5) requires the points of the initial manifold to move. Equations (4.4) require the motion to be toward higher values of x^0 along timelike and lightlike world lines. The smallest interval τ required to reach an attainable final point x, s is zero as discussed in the previous paragraph. The boundary of the reachable set is a trailing edge defined by the family of maximum τ trajectories that originate at $(\bar{x},0)$ with $\lambda(0)$'s that satisfy the transversality condition.

According to Huygens' principle, extremals will be generated when the control vector u is chosen to minimize the control Hamiltonian H:

$$H = (\dot{x}, \lambda) + \lambda_s \dot{s} = (u, \lambda - \lambda_s A)/|u|. \tag{4.7}$$

(The usual situation with $\lambda - \lambda_s A$ non-null will be assumed at first. The exceptional case will be discussed in Sec. VI.) The wave front normal five-tuple is composed of the vector λ and the scalar λ_s. The vector λ will be shown to be equal to a scalar times a gradient and is therefore covariant. If there were no constraints on u, the Hamiltonian could always be made to approach minus infinity. The numerator could be made negative by choosing a u that makes an obtuse angle with $\lambda - \lambda_s A$. The denominator $|u|$ could be made arbitrarily small by approaching a null direction.

However, (4.4) and (4.4') require the vector u to point inside the future half-cone. Section VI will show that transversality requires $\lambda(0) - \lambda_s A(0)$ to be future pointing also. This restriction can be stated mathematically as

$$\lambda_0 - \lambda_s A_0 \geqslant 0, \quad (\mathbf{G}^{-1}(\lambda - \lambda_s A), \lambda - \lambda_s A) \geqslant 0 \quad (\tau = 0) \tag{4.8}$$

$$\rightarrow \lambda_0 - \lambda_s A_0 \geqslant + [(\boldsymbol{\lambda} - \lambda_s \mathbf{A})^2]^{1/2}. \tag{4.8M}$$

Now H cannot be made negative by any admissible u. Since u enters the Hamiltonian to zero degree, if H is stationary at a point $\bar{u}$, then it will have the same value at any point of the half-line $u = k\bar{u}$ defined for $k > 0$. In the future half-cone the stationary points are improper minimums that satisfy

$$\frac{\partial H}{\partial u} = |u|^{-1}(\lambda - \lambda_s A) - (u, \lambda - \lambda_s A)|u|^{-3}\mathbf{G}u = 0. \tag{4.9}$$

(This equation is used in optimal control to eliminate u just as $P = \partial L / \partial \dot{x}$ is used in the calculus of variations to eliminate $\dot{x}$.) After multiplying by $|u|$ it is easy to see that

$$\lambda - \lambda_s A = (u, \lambda - \lambda_s A)|u|^{-2}\mathbf{G}u = H\,\mathbf{G}u/|u| = H\,\mathbf{G}\dot{\mathbf{x}}. \tag{4.10}$$

Although these four simultaneous equations are quadratic in the unknowns $u/|u|$, the latter can be eliminated just by taking the scalar product of the vector equation with itself. Then H becomes the true Hamiltonian $\mathscr{H}(x,\lambda,\lambda_s)$:

$$\mathscr{H}^2 = (\mathbf{G}^{-1}(\lambda - \lambda_s A), \lambda - \lambda_s A). \tag{4.11}$$

Since the Hamiltonian has been shown to be non-negative, it must be set to the positive value of the square root,

$$\mathscr{H} = + (\mathbf{G}^{-1}(\lambda - \lambda_s A), \lambda - \lambda_s A)^{1/2} \tag{4.12}$$

$$\rightarrow + [(\lambda_0 - \lambda_s A_0)^2 - (\boldsymbol{\lambda} - \lambda_s \mathbf{A})^2]^{1/2}. \tag{4.12M}$$

Equations (4.4) and (4.4″) can be rewritten using (4.10) with $H = \mathscr{H}$:

$$\dot{x} = \frac{\mathbf{G}^{-1}(\lambda - \lambda_s A)}{\mathscr{H}} = \frac{\partial H}{\partial \lambda}, \tag{4.13a}$$

$$\dot{s} = \frac{-(\mathbf{G}^{-1}(\lambda - \lambda_s A), A)}{\mathscr{H}} = \frac{\partial \mathscr{H}}{\partial \lambda_s}. \tag{4.13b}$$

The velocity $\dot{x}$ has unit magnitude for any choice of λ, λ_s. The latter obey Hamilton's equations also. The derivation will be given for those readers who are unfamiliar with the optimal control approach. Since λ, λ_s are normal to the wave front, they satisfy

$$(\delta x, \lambda) + \lambda_s \delta s = 0 \quad (\tau > 0). \tag{4.14}$$

The derivative of this equation will be found using $d(\delta x)/d\tau = \delta \dot{x}$. Equations (4.4) will be more convenient than (4.13):

$$\begin{aligned} \frac{d}{d\tau}[(\delta x, \lambda) + \lambda_s \delta s] &= (\delta x, \dot{\lambda}) + \dot{\lambda}_s \delta s - \frac{(u, \lambda)\partial_\nu g_{\mu\rho} u^\mu u^\rho \delta x^\nu}{2|u|^3} \\ &+ \lambda_s \left[-\partial_\nu A_\rho + \frac{(u, A)\partial_\nu g_{\mu\rho} u^\mu}{2|u|^2} \right] \frac{u^\rho \delta x^\nu}{|u|} \\ &+ \frac{\partial [(\dot{x}, \lambda) + \lambda_s \dot{s}]}{\partial u^\mu} \delta u^\mu = 0 \end{aligned} \tag{4.15}$$

($\partial_\nu \cdot \equiv \partial \cdot / \partial x^\nu$). The coefficients of the δu^μ vanish by virtue of (4.9). Now $\dot{\lambda}, \dot{\lambda}_s$ will be defined so that the coefficients of the $\delta x^\nu, \delta s$ vanish:

$$\begin{aligned} \dot{\lambda}_\nu &= (H\,\partial_\nu g_{\mu\rho}\, u^\mu/2|u| + \lambda_s\, \partial_\nu A_\rho) u^\rho/|u| = -\partial_\nu H \\ &= (H \partial_\nu g_{\mu\rho}\, \dot{x}^\mu/2 + \lambda_s \partial_\nu A_\rho)\dot{x}^\rho = -\partial_\nu \mathscr{H}, \end{aligned} \tag{4.16a}$$

$$\dot{\lambda}_s = 0 = \partial_s \mathscr{H}. \tag{4.16b}$$

These equations may be checked using (4.12) with $\partial_\nu g^{\sigma\kappa} = -g^{\sigma\mu} g^{\kappa\rho}\, \partial_\nu g_{\mu\rho}$. Here λ_s is a constant of the motion. This property is easily proved for $\mathscr{H}$ also, using (4.13) and (4.16),

$$\dot{\mathscr{H}} = \partial_\mu \mathscr{H}\, \dot{x}^\mu + \frac{\partial \mathscr{H}}{\partial \lambda_\mu} \dot{\lambda}_\mu + \frac{\partial \mathscr{H}}{\partial \lambda_s} \dot{\lambda}_s = 0. \tag{4.17}$$

The equation for the acceleration $\ddot{x}$ will use the following notation: $F_{\nu\rho}$ is the electromagnetic field tensor, $\Gamma_{\rho\mu,\nu}, \Gamma^\sigma_{\mu\rho}$ are Christoffel three index symbols of the first and second kind,

$$F_{\nu\rho} \equiv -\partial_\rho A_\nu + \partial_\nu A_\rho, \quad F^\sigma_\rho \equiv g^{\sigma\nu} F_{\nu\rho}, \tag{4.18a}$$

$$\Gamma_{\mu\rho,\nu} \equiv (\partial_\rho g_{\nu\mu} - \partial_\nu g_{\mu\rho} + \partial_\mu g_{\rho\nu})/2, \quad \Gamma^\sigma_{\mu\rho} \equiv g^{\sigma\nu} \Gamma_{\mu\rho,\nu}, \tag{4.18b}$$

$$\Gamma_{\mu\rho,\nu} \dot{x}^\mu \dot{x}^\rho = (\partial_\rho g_{\mu\nu} - \partial_\nu g_{\mu\rho}/2)\dot{x}^\mu \dot{x}^\rho, \quad \dot{A}_\nu = \partial_\rho A_\nu \dot{x}^\rho. \tag{4.19}$$

Multiplying (4.13a) by $\mathbf{G}$, differentiating, and substituting from (4.16) and the second of (4.19) yield

$$\begin{aligned} &\partial_\rho g_{\mu\nu} \dot{x}^\rho \dot{x}^\mu + g_{\mu\nu} \ddot{x}^\mu \\ &\quad = \frac{\dot{\lambda}_\nu - \lambda_s \dot{A}_\nu}{\mathscr{H}} \\ &\quad = \frac{\partial_\nu g_{\mu\rho} \dot{x}^\mu \dot{x}^\rho}{2} + \left(\frac{\lambda_s}{\mathscr{H}}\right)(\partial_\nu A_\rho - \partial_\rho A_\nu)\dot{x}^\rho. \end{aligned} \tag{4.20}$$

After solving for $\ddot{x}$ and introducing the definitions of (4.18), this equation can be written as

$$\ddot{x}^\sigma = (\lambda_s/\mathscr{H}) F^\sigma_\rho \dot{x}^\rho - \Gamma^\sigma_{\mu\rho} \dot{x}^\mu \dot{x}^\rho. \tag{4.21}$$

The two terms on the right are similar except for the number of indices. This also applies to the definitions of $F_{\nu\rho}$ and $\Gamma_{\mu\rho,\nu}$ in (4.18).

When the last equation is written with the Minkowski metric, it will agree with Refs. 2,3,7, and 8 if the constant $\lambda_s/\mathscr{H}$ is identified with q/mc^2. Since the extremals reduce to geodesics that are not influenced by the electromagnetic force when $\lambda_s = 0$, they correspond to experiment if

$$\lambda_s \equiv q, \quad \mathscr{H} \equiv mc^2. \tag{4.22}$$

Putting the m with $\mathscr{H}$ rather than λ_s permits the Hamiltonian to have different values on different trajectories. Putting the c^2 with $\mathscr{H}$ gives the Hamiltonian its traditional dimension "energy." Our equations allow charge to have either sign, but require mass to be non-negative. We may now speak of the principle of least action only when the charge is positive. Action is maximized when q is negative. Proper time is maximized when the minimized $\mathscr{H}$ is positive and "open" when $\mathscr{H}$ is zero. The vector λ will be written as P when it accompanies the notation of (4.22). The optimal control terminology in (4.21) may now be replaced with that of experimental physics[9]

$$\ddot{x}^\sigma = (q/mc^2)F^\sigma_\rho \dot{x}^\rho - \Gamma^\sigma_{\mu\rho}\dot{x}^\mu\dot{x}^\rho \tag{4.23a}$$

$$= [qF^\sigma_\rho - \Gamma^\sigma_{\mu\rho}g^{\mu\nu}(P_\nu - qA_\nu)]g^{\rho\kappa}[P_\kappa - qA_\kappa]/m^2c^4 \tag{4.23b}$$

$$\rightarrow (q/m^2c^4)\eta^{\sigma\mu}(\partial_\mu A_\rho - \partial_\rho A_\mu)\eta^{\rho\kappa}(P_\kappa - qA_\kappa). \tag{4.23M}$$

The trajectory family with the given q/m and initial conditions can be calculated from (4.23a) by varying three of the $\dot{x}^\mu(0)$. The fourth can be found from $(\dot{\mathbf{x}}, \mathbf{G}\dot{\mathbf{x}}) = 1$ using the quadratic formula. (A problem may occur because $\mathbf{G}$ is indefinite.) Alternatively, (4.13) and (4.16) can be used with three elements of P varied and the fourth determined by (4.12) and (4.22).

V. THE HAMILTON–JACOBI AND EIKONAL EQUATIONS

A general state variable increment

$$\Delta x = \delta x + \dot{x}\Delta\tau \tag{5.1}$$

is composed of a component δx in the wave front tangent plane and another $\dot{x}\Delta\tau$ that is out of the plane. Equation (4.14) generalizes to

$$(\Delta x, \lambda) + \Delta s\lambda_s = [(\dot{x}, \lambda) + \dot{s}\lambda_s]\Delta\tau = \mathscr{H}\Delta\tau. \tag{5.2}$$

This equation shows that when $\mathscr{H} \neq 0$,

$$\partial_\mu\tau = \lambda_\mu/\mathscr{H}, \quad \partial_s\tau = \lambda_s/\mathscr{H}. \tag{5.3}$$

(These relations confirm that the five-tuple λ,λ_s is normal to the wave front. We might mention that $\partial_s\tau = q/mc^2$.) Substituting these relations into (4.11) yields the Hamilton–Jacobi partial differential equation that governs the wave fronts $\tau(x,s) = \text{const}$,

$$g^{\mu\nu}(\partial_\mu\tau - A_\mu\,\partial_s\tau)(\partial_\nu\tau - A_\nu\,\partial_s\tau) = 1 \quad (\mathscr{H}\neq 0) \tag{5.4}$$

$$\rightarrow (\partial_0\tau - A_0\,\partial_s\tau)^2 - (\nabla\tau - \mathbf{A}\,\partial_s\tau)^2 = 1. \tag{5.4M}$$

It is natural to speculate that there are waves bounded by these surfaces. The amplitude ψ would have the functional dependence $\psi = \psi(x,s,\tau)$. The 6–D wave equation would presumably separate into components that include the 4-D Klein–Gordon or Dirac equation.

When $\lambda_s = \mathscr{H} = 0$, (4.11) and (5.2) – (5.3) become

$$(\mathbf{G}^{-1}\lambda, \lambda) = 0, \tag{5.5}$$

$$(\Delta x, \lambda) = \lambda_0\Delta x^0 + \boldsymbol{\lambda}\cdot\Delta\mathbf{x} = (\lambda_0\nabla x^0 + \boldsymbol{\lambda})\cdot\Delta\mathbf{x} = 0, \tag{5.6}$$

$$\nabla x^0 = -\boldsymbol{\lambda}/\lambda_0. \tag{5.7}$$

Equation (5.5) states that λ is tangent to the surface of the future half-cone which means that λ_0 is positive.Substituting (5.7) into (5.5) yields the equation for the eikonal surfaces $x^0(\mathbf{x}) = \text{const}$ of geometric optics[10]

$$g^{00} - 2g^{0i}\partial_i x^0 + g^{ij}\partial_i x^0\partial_i x^0 = 0 \quad (\lambda_s = \mathscr{H} = 0) \tag{5.8}$$

$$\rightarrow (\nabla x^0)^2 = 1. \tag{5.8M}$$

(Repeated Latin indices are summed from 1 to 3.) As x^0 increases from $\bar{x}^0$ to infinity, these surfaces sweep out the future half-cone associated with the point $\bar{x}$. The eikonals bound Maxwell's electromagnetic 4-D waves $\psi = \psi(x)$. Thus the 6-D waves bounded by (5.4) should reduce to EM waves as mass and charge approach zero.

Physicists have found that quantum mechanical equations can be integrated formally when the initial probability distribution is Gaussian. Since this distribution is nonzero over all space-time, there are no wave fronts initially or later. This obscures the relationship between quantum mechanics and both analytical dynamics and EM waves. It would be interesting to see the figures of Ref. 11 for a disturbance that is intially concentrated at a point.

VI. REMARKS

In the notation of experimental physics, (4.12) and (4.13a) become

$$mc^2 = +(G^{-1}(P - qA), P - qA)^{1/2} \tag{6.1}$$

$$\rightarrow + [(P_0 - qA_0)^2 - (\mathbf{P} - q\mathbf{A})^2]^{1/2}, \tag{6.1M}$$

$$\dot{x} = \mathbf{G}^{-1}(P - qA)/mc^2. \tag{6.2}$$

These equations show that when $P(0) - qA(0)$ is non–null but in a null direction, $m = 0$ and $\dot{x} = \infty$. (Recall that this means that transversality is satisfied marginally.) Since m and q are unrelated in the present equations, charge is permitted to have any value on the resulting lightlike trajectory. (Presumably $q \neq 0$ will be forbidden by quantum theory.) When q also vanishes, P_0 must be positive so that $\dot{x}^0 > 0$. The transversality conditions for inward pointing normals then state that time is minimized with action and relativistic interval open. This is Fermat's principle.

Assume that the potential A is due to a charged mass point. Then repulsive trajectories also minimize time. Maximum time extremals are possible in the attractive case. [See the first of (4.8).]

Equation (6.2) can be written as

$$P = mc^2\mathbf{G}\dot{x} + qA = mc^2\mathbf{G}\,dx/(dx,\mathbf{G}\,dx)^{1/2} + qA \tag{6.3}$$

$$\rightarrow \frac{mc^2\boldsymbol{\eta}}{(1-\mathbf{v}^2)^{1/2}}\frac{dx}{dx^0} + qA. \tag{6.3M}$$

When $\mathbf{G}, A$ are time independent, P_0 is a constant of the motion. Let us assume that the metric is in diagonal form:

$$\mathbf{G} = \text{diag}[g_{00},g_{11},g_{22},g_{33}], \quad v \equiv \frac{dx}{dx^0} = [1,\mathbf{v}], \tag{6.4}$$

$$P_0 = mc^2 g_{00}/(v, \mathbf{G}v)^{1/2} + qA_0. \tag{6.5}$$

After substituting into $\mathbf{G}$ a line element such as that of Reissner–Nordstrom[12] for a charged mass point, it is easy to see from the weak field and low speed approximations that the first term on the right contains the self-energy, kinetic energy, and gravitational potential energy. The second term of course adds the EM potential energy qA_0. Thus P_0 is the energy of the particle. (The energy of the fields produced by the particle is neglected.) Here $\mathbf{P}/c$ is its covariant canonical momentum.

We now consider the exceptional case for which the four equations $P - qA = 0$ are satisfied. The surface H vs u, defined by (4.7), becomes the horizontal plane $H = 0$. The minimum principle cannot select the optimal $u/|u|$. Those controls that keep $P - qA$ null as τ increases produce *singular extremals.*[13] Singular controls satisfy

$$\frac{d(P_\nu - qA_\nu)}{d\tau} = \frac{q(\partial_\nu A_\rho - \partial_\rho A_\nu)u^\rho}{|u|} = \frac{qF_{\nu\rho}u^\rho}{|u|} = 0, \tag{6.6}$$

and therefore exist only when $\det[F_{\nu\rho}] = 0$. Although this

condition is often satisfied, singular trajectories do not appear to be important physically, probably do not survive quantization, and will not be discussed further here.

Suppose that A is null. Then (4.4″) implies the holonomic constraint $s = 0$. The extremals that originate at a point can not fill a five-dimensional volume. The direction of the wave front normal is then not completely determined. There will be a one-parameter family of normal five-tuples associated with each extremal. An optimal trajectory whose $\lambda_s = q \neq 0$ will be identical to one with $q = 0$ and can be properly interpreted as maximizing τ with s open.

A formulation in the literature[2–4] for the present problem can be obtained from (4.6M) with a general metric together with (4.4) and (4.4′). The control four-vector is $y \equiv u/|u|$,

$$\dot{x} = y, \quad \dot{s} = -mc(y, \mathbf{G}y)^{1/2} - (q/c)(y, A), \quad y^0 > 0. \tag{6.7}$$

The control Hamiltonian for this system is homogeneous of degree 1 in y. Euler's relation $nH = y^\mu \, \partial H / \partial y^\mu$ with n equal to the degree requires the Hamiltonian to vanish at its minimum. References 2–4 show that the extremals are identical to those of this paper provided that the final step sets $(y, \mathbf{G}y) = 1$. The final transversality conditions indicate that s is optimized with τ open. But when $q = 0, s = -mc\tau$ so that if s is optimized so is τ. The resolution of this paradox will be left to the interested reader. Although there are four initial Lagrange multiplier ratios, the extremal family has only three degrees of freedom due to the restriction on $\mathscr{H}$. This is the number found in Sec. II for the formulation with four state variables. When the quantities q,m are allowed to change, the order of the extremal family increases from 3 to only 4. This is due to the invariance of the extremals when λ_s is modified with q,m varied inversely. Thus a four-parameter family issues from the initial point just as in Sec. IV.

The system (4.4) can be generalized to the following $4 + j$ equations to allow for additional forces:

$$\dot{x} = u/|u|, \quad \dot{s}_i = L_i(x,u), \quad u^0 > 0 \quad (i = 1,...,j). \tag{6.8}$$

The Lagrange multipliers λ_{4+i} are constants that can be identified with the corresponding charges.

VII. CONCLUSIONS

The principle of stationary action has been formulated as a manifestly covariant optimal control problem whose independent variable is proper time. There are Lagrange multipliers for all five state variables including action. The dimension of action has been changed from erg seconds to charge.

Classical particles would always traverse null geodesics if they could move in both time directions. Nontrivial trajectories can be described by an action principle only if motion to earlier times is prohibited. Particles that disobey must be quantum mechanical.

When a family of extremals issues from a point, the wave fronts for most variational problems have either leading edges only or else both leading and trailing edges. However, when the metric tensor of a geodesic problem is indefinite and there is a control variable inequality constraint, the wave fronts have trailing edges only. This explains why $\mathscr{H} = mc^2$ is always non-negative in classical mechanics.

A Lagrangian formed from the scalar product of the velocity and potential four-vectors, but with no constants, leads to an EM field, an electric charge, and a non-negative mass. It would be interesting to apply the methods of this article to the theory given in Ref. 14. Both electric and magnetic charges with corresponding four-potentials are present in this classical treatment.

The normal five-tuple and Hamiltonian employed by the optimal control formulation are familar to experimental physicists. They call λ_0 the "particle's energy," $\boldsymbol{\lambda}/c$ the "canonical momentum," λ_s the "charge," and $\mathscr{H}$ the "self-energy."

A four-parameter family of extremals issues the from an initial point. Three may be regarded as usual as the components of the velocity with respect to time. The new parameter is q/mc^2.

When q is zero, the particle moves along a geodesic that maximizes relativistic interval with action open. If, in addition, the mass is zero, the interval is also open and the trajectory can be regarded as minimizing time (Fermat's principle). In this case the speed of the particle must be that of light.

The principle of stationary action-interval time developed in this paper unifies the trajectories of lightlike (e.g., photons) and subluminal particles. The eikonals $t(\mathbf{x}) = \text{const}$ of geometric optics are generalized to the wave fronts $\tau(x,s)$ Similarly, EM waves should be extended to 6-D waves with amplitudes $\psi = \psi(x,s,\tau)$. In accordance with the correspondence principle of wave mechanics, ψ should vanish at the wave front, be large at a short distance in from the wave front, and be small elsewhere in the interior.

The simplicity, symmetry, and beauty of manifestly covariant equations have been purchased at a high price. We have given up the specific q/m of a particular particle. The present treatment of classical particles has returned some interesting diagnostic relations. Whether this have been a step toward truth will not be known until 6-D waves and their transition to 4-D quantum mechanics have been explored.

[1] M. R. Hestenes, *The Calculus of Variations and Optimal Control Theory* (Wiley, New York, 1966).
[2] A. O. Barut, *Electrodynamics and the Classical Theory of Fields and Particles* (MacMillan, New York, 1964), p. 62.
[3] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), 2nd ed., p. 577.
[4] O. D. Johns, Am. J. Phys. **53**, 982 (1985).
[5] H. Goldstein, *Classical Mechanics* (Addison–Wesley, Reading, MA, 1980), 2nd ed., p. 322.
[6] See Ref. 3, p. 574.
[7] See Ref. 5, p. 330.
[8] J. L. Anderson, *Principles of Relativity Physics* (Academic, New York, 1967), p. 217.
[9] See Ref. 8, p. 356.
[10] See Ref. 5, p. 489.
[11] R. L. W. Chen and M. B. Rhodes, Am. J. Phys. **52**, 988 (1984).
[12] See Ref. 8, p. 398.
[13] H. G. Moyer, SIAM Control Optim. **11**, 620 (1973).
[14] W. Hauser, *Introduction to the Principles of Electromagnetism* (Addison–Wesley, Reading, MA, 1971), p. 242 f.

Stochastic quantization of a Fermi field: Fermions as solitons

P. Bandyopadhyay and K. Hajra
Indian Statistical Institute, Calcutta-700035, India

(Received 5 February 1986; accepted for publication 1 October 1986)

It is shown that the stochastic quantization of a fermion introducing an anisotropy in the internal space so that this gives rise to two internal helicities corresponding to particle and antiparticle leads us to describe a fermion as a Skyrme soliton. The Skyrme term appears here as a consequence of this internal anisotropy and can be treated as a quantum effect. Some topological properties of this fermionization are then discussed.

I. INTRODUCTION

In a series of papers Skyrme[1] argued that nucleons and other baryonic resonances can be treated as solitons that arise as solutions of nonlinear Lagrangians where the pseudoscalar mesons are described by 2×2 unitary matrices commonly known as the nonlinear sigma model. Skyrme and Williams[2] gave general arguments to show that if a suitable quantization scheme is adopted, the spin of these particles will emerge correctly as a half-odd integer. Recent developments regarding strong interaction dynamics have revived the old idea of Skyrme as, at low energy, the theory reduced to a nonlinear sigma model of spontaneously broken chiral symmetry. Indeed Pak and Tze[3] studied in detail the current algebraic and topological aspects of the chiral model. Gipson and Tze[4] also argued that the usual weak interaction model as well predicts Skyrme solitons with exotic properties. Later on Witten[5] as well as Balchandran *et al.*[6] studied various properties of such solitonic solutions and interpreted them as physical baryonic states.

Finkelstein and Rubinstein[7] showed in a very general framework how the quantization of a soliton may lead to a fermion based on a homotopy classification of soliton solutions. Mickelsson[8] has proved the spin and statistics connection in classical point-particle mechanics in the spirit of the field-theoretic homotopy proof by Finkelstein and Rubinstein when the geometry of the configuration space is determined by a parallelization describing a soliton field. All these features indicate that fermions may appear as solitons when a consistent quantization procedure is taken into account.

In a recent paper[9] it has been shown that Nelson's stochastic quantization procedure[10] can be generalized to the relativistic case and a fermion can be quantized when we consider universal Brownian motion in the external space as well as in the internal space of the particle. An anisotropy is introduced in the internal space so that this gives rise to two internal helicities depicting particle and antiparticle states. Thus the internal helicities may be taken to represent a geometrical interpretation of the fermion number. This is also the case for the hydrodynamical quantization method where a fermion is generated when a vortex line is introduced corresponding to the anisotropy of the internal domain.[11] In this paper we shall show that these features relating to the quantization procedure of a fermion and the geometrical interpretation of fermion number effectively leads to the fact that fermions in general correspond to Skyrme solitons.

In Sec. II, we shall recapitulate for completeness the stochastic quantization procedure of a fermion and in Sec. III we shall show that fermions in this picture effectively appear as Skyrme solitons. In Sec. IV we shall discuss the geometrical and topological properties of fermionization.

II. STOCHASTIC QUANTIZATION OF A FERMION

Nelson's stochastic quantization procedure[10] is based on the assumption that the configuration variable $q(t)$ is promoted to a Markov process, which will be denoted here by the same symbol $q(t)$. The process $q(t)$ is determined by two conditions, the first is the hypothesis of universal Brownian motion, the second is the validity of the Euler–Lagrange equations. In the present framework we take that apart from a Brownian motion process in the external space, there is a Brownian motion process in the internal space also. In view of this we denote the configuration variable as $Q(t,\xi_0)$, where ξ_0 is the fourth component (real) of the internal four-vector ξ_μ. We assume that $Q(t,\xi_0)$ is a separable function and can be denoted as

$$Q(t,\xi_0)=q(t)q(\xi_0). \tag{1}$$

The process $Q(t,\xi_0)$ is assumed to satisfy the stochastic differential equations

$$dQ_i(t,\xi_0)=b_i(Q(t,\xi_0),t,\xi_0)dt+d\omega_i(t), \tag{2}$$

$$dQ_i(t,\xi_0)=b_i'(Q(t,\xi_0),t,\xi_0)d\xi_0+d\omega_i(\xi_0), \tag{3}$$

which depict the Brownian motion processes in the external and internal space, respectively. Here $b_i(Q(t,\xi_0),t,\xi_0)$ and $b_i'(Q(t,\xi_0),t,\xi_0)$ correspond to certain velocity fields in the external and internal space and $d\omega_i$ are independent Brownian motions. It is assumed that $d\omega_i(t)(d\omega_i(\xi_0))$ does not depend on $Q(S,S')$ for $S\leqslant t(S'\leqslant\xi_0)$ and the expectations have the following values:

$$\begin{aligned}
&\langle d\omega_i(t)\rangle=0,\\
&\langle d\omega_i(t)d\omega_j(t')\rangle=(\hbar/m)\delta_{ij}\delta(t-t')dt\,dt',\\
&\langle d\omega_i(\xi_0)\rangle=0,\\
&\langle d\omega_i(\xi_0)d\omega_j(\xi_0')\rangle=(\hbar/\pi_0)\delta_{ij}\delta(\xi_0-\xi_0')d\xi_0\,d\xi_0',
\end{aligned} \tag{4}$$

where $\hbar$ is Planck's constant divided by 2π and m and π_0 are suitable constants. The description is asymmetrical in both "external" and "internal" time, but we can also write

$$dQ_i(t,\xi_0)=b_i^*(Q(t,\xi_0),t,\xi_0)dt+d\omega_i^*(t), \tag{5}$$

$$dQ_i(t,\xi_0)=b_i'^*(Q(t,\xi_0),t,\xi_0)d\xi_0+d\omega_i^*(\xi_0), \tag{6}$$

 0022-2488/87/030711-06$02.50

where now ω^* has the same properties as ω except that $d\omega_i^*(t)(d\omega_i^*(\xi_0))$ are independent of $Q(S,S')$ for $S \geqslant t$ $(S' \geqslant \xi_0)$. Now we introduce the mean forward derivatives $D_t Q_i(t,\xi_0)$, $D_{\xi_0} Q_i(t,\xi_0)$ and the mean backward derivatives $D_t^* Q_i(t,\xi_0)$, $D_{\xi_0}^* Q_i(t,\xi_0)$ through the following definitions in analogy with those proposed by Nelson:

$$
\begin{aligned}
D_t Q_i(t,\xi_0) &= \lim_{h\to 0^+} E_i \frac{Q_i(t+h,\xi_0) - Q_i(t,\xi_0)}{h}, \\
D_{\xi_0} Q_i(t,\xi_0) &= \lim_{h\to 0^+} E_i \frac{Q_i(t,\xi_0+h) - Q_i(t,\xi_0)}{h}, \\
D_t^* Q_i(t,\xi_0) &= \lim_{h\to {}^+0} E_i \frac{Q_i(t,\xi_0) - Q_i(t-h,\xi_0)}{h}, \\
D_{\xi_0}^* Q_i(t,\xi_0) &= \lim_{h\to {}^+0} E_i \frac{Q_i(t,\xi_0) - Q_i(t,\xi_0-h)}{h},
\end{aligned}
\tag{7}
$$

where E_i is the conditional expectation with respect to the σ algebra Σ, generated by the random variables $Q_i(t,\xi_0)_{i=1,2,\ldots,n}$.

Since, by definition,

$$
\begin{aligned}
&E_i(d\omega_i(t)) = E(d\omega_i(t)) = 0, \\
&E_i(d\omega_i(\xi_0)) = E(d\omega_i(\xi_0)) = 0, \\
&E_i(d\omega_i^*(t)) = E(d\omega_i^*(t)) = 0, \\
&E_i(d\omega_i^*(\xi_0)) = E(d\omega_i^*(\xi_0)) = 0,
\end{aligned}
\tag{8}
$$

we have

$$
\begin{aligned}
&D_t(Q_i(t,\xi_0)) = b_i(Q(t,\xi_0),t,\xi_0), \\
&D_{\xi_0}(Q_i(t,\xi_0)) = b_i'(Q(t,\xi_0),t,\xi_0), \\
&D_t^*(Q_i(t,\xi_0)) = b_i^*(Q(t,\xi_0),t,\xi_0), \\
&D_{\xi_0}^*(Q_i(t,\xi_0)) = b_i'^*(Q(t,\xi_0),t,\xi_0).
\end{aligned}
\tag{9}
$$

In general, for the sufficiently regular function $F(Q(t,\xi_0),t,\xi_0)$ we have

$$
\begin{aligned}
&D_t F(Q(t,\xi_0),t,\xi_0) \\
&\quad = \left(\frac{\partial}{\partial t} + \sum_{i=1}^n b_i \frac{\partial}{\partial Q_i} + \frac{\hbar}{2m}\Delta\right) F(Q(t,\xi_0),t,\xi_0), \\
&D_{\xi_0} F(Q(t,\xi_0),t,\xi_0) \\
&\quad = \left(\frac{\partial}{\partial \xi_0} + \sum_{i=1}^n b_i' \frac{\partial}{\partial Q_i} + \frac{\hbar}{2\pi_0}\Delta\right) F(Q(t,\xi_0),t,\xi_0), \\
&D_t^* F(Q(t,\xi_0),t,\xi_0) \\
&\quad = \left(\frac{\partial}{\partial t} + \sum_{i=1}^n b_i^* \frac{\partial}{\partial Q_i} - \frac{\hbar}{2m}\Delta\right) F(Q(t,\xi_0),t,\xi_0), \\
&D_{\xi_0}^* F(Q(t,\xi_0),t,\xi_0) \\
&\quad = \left(\frac{\partial}{\partial \xi_0} + \sum_{i=1}^n b_i'^* \frac{\partial}{\partial Q_i} - \frac{\hbar}{2\pi_0}\Delta\right) F(Q(t,\xi_0),t,\xi_0),
\end{aligned}
\tag{10}
$$

where

$$
\Delta = \sum_{i=1}^n \frac{\partial^2}{\partial Q_i^2}.
$$

Now we can derive the following moments[12]:

$$
\begin{aligned}
&\langle q_i(t)\rangle = 0, \\
&\langle q_i(t) q_j(t')\rangle = (\hbar/2m\omega)\delta_{ij} e^{-\omega(t-t')} \quad (t > t'), \\
&\langle q_i(\xi_0)\rangle = 0, \\
&\langle q_i(\xi_0) q_j(\xi_0')\rangle = (\hbar/2\pi_0\omega')\delta_{ij} e^{-\omega'(\xi_0 - \xi_0')} \quad (\xi_0 > \xi_0').
\end{aligned}
\tag{11}
$$

From these expressions the moments of the product variables can be derived and are given by

$$
\begin{aligned}
&\langle Q_i(t,\xi_0)\rangle = 0, \\
&\langle Q_i(t,\xi_0) Q_j(t',\xi_0')\rangle \\
&\quad = (\hbar/2m\omega)\,(\hbar/2\pi_0\omega')\delta_{ij} e^{-\omega(t-t')} e^{-\omega'(\xi_0 - \xi_0')} \\
&\qquad (t > t', \quad \xi_0 > \xi_0').
\end{aligned}
\tag{12}
$$

Let $\{e_i(\mathbf{x})\}$ denote the complete orthonormal set of eigenfunctions of the three-dimensional Laplacian Δ,

$$
\Delta e_i(\mathbf{x}) = -k_i^2 e_i(\mathbf{x}). \tag{13}
$$

Also we denote $\{e_j(\boldsymbol{\xi})\}$ as the set of complete orthonormal set of eigenfunctions of the three-dimensional Laplacian Δ' in terms of the variables

$$
\boldsymbol{\xi}\left(\Delta' = \frac{\partial^2}{\partial \xi_1^2} + \frac{\partial^2}{\partial \xi_2^2} + \frac{\partial^2}{\partial \xi_3^2}\right)
$$

so that

$$
\Delta' e_j(\boldsymbol{\xi}) = -\pi_j^2 e_j(\boldsymbol{\xi}). \tag{14}
$$

Now we can construct a stochastic nonlocal field φ, which can be expressed as an orthonormal expansion in terms of $q_i(t)$, $e_i(\mathbf{x})$, $q_i(\xi_0)$, and $e_j(\boldsymbol{\xi})$, and write

$$
\varphi(x,t,\xi) = \sum_{i,j} q_i(t) e_i(\mathbf{x}) q_j(\xi_0) e_j(\boldsymbol{\xi}). \tag{15}
$$

Now from the moments of $Q_i(t,\xi_0)$ we can determine the moments of $\varphi(x,t,\xi)$,

$$
\begin{aligned}
&\langle \varphi(x,t,\xi)\rangle = 0, \\
&\langle \varphi(x,t,\xi)\varphi(x',t',\xi')\rangle \\
&\quad = \frac{1}{(2\pi)^4}\int \frac{d^4k\, e^{i(k,(x-x'))}}{(k,k)+m^2} \frac{1}{(2\pi)^4}\int \frac{d^4\pi\, e^{i(\pi,(\xi-\xi'))}}{(\pi,\pi)+\pi_0^2},
\end{aligned}
\tag{16}
$$

where (a,b) denotes the Euclidean product.

It is noted that in the limit, $\xi_0 = \xi_0' = 0$, and integrating over the internal space variable $\boldsymbol{\xi}$, the correlation function (16) reduces to

$$
\langle \varphi(x,t)\varphi(x',t')\rangle = \frac{1}{(2\pi)^4}\int \frac{d^4k\, e^{i(k,(x-x'))}}{(k,k)+m^2}, \tag{17}
$$

which is the correlation function of a scalar field. This is the Euclidean Markov field result which has been obtained starting from Nelson's real time formalism of Brownian motion and in this sense gives rise to the equivalence of these two formalisms as advocated by Guerra and Ruggiero.[13]

Now we want to show that when the anisotropic feature of the internal space-time corresponding to the variable ξ_μ is

taken into account implicitly we can obtain the fermionic propagator in Euclidean space-time. To this end, we introduce the anistropy by having two opposite orientations of the internal variable ξ_μ (and hence of $\pi_\mu = i\partial/\partial\xi_\mu$) and take that each orientation denotes a separate field and the two opposite orientations depict two separate fields having two opposite internal helicities corresponding to particle and antiparticle configurations. That is, we can take for these two configurations, the internal space-time variable $(+\xi_1, +\xi_2, +\xi_3, +\xi_0)$ and $(-\xi_1, -\xi_2, -\xi_3, -\xi_0)$ and this indicates that $i\pi_\mu$ and $-i\pi_\mu$ ($\pi_\mu = i\partial/\partial\xi_\mu$) will correspond to two different internal helicities depicting two different configurations giving rise to particle and antiparticle states.

Now from Eq. (16), we see that it is effectively a correlation function in eight-dimensional space-time, four-dimensional in the external space-time variable and four-dimensional in the internal space-time variable. To make it an effective four-dimensional expression in the external space-time variable so that the role of the anisotropic feature of the internal space is exhibited properly, we introduce a mapping of the external and internal space as follows:

$$k^2 = (k',\pi), \quad x^2 = (x',\xi), \quad m^2 = m'\pi_0, \tag{18}$$

where $(k',\pi)[(x',\xi)]$ denotes an Euclidean product and each component of $k(x)$ is given by $k_i = \sqrt{k_i'\pi_i}$ $[x_i = \sqrt{x_i'\xi_i}]$. By introducing these new variables, we can write the correlation function of the new field variables from the expression (17) as follows:

$$\langle\bar{\varphi}(x,t,\xi)\bar{\varphi}(x',t',\xi')\rangle = \frac{1}{(2\pi)^4}\int \frac{e^{i(\sqrt{k'\pi},(x(\xi)-x'(\xi')))}\, d^4\sqrt{k'\pi}}{(k,'\pi)+m'\pi_0} = \frac{1}{(2\pi)^4} \times \int \frac{e^{i(\sqrt{k'\pi},(x(\xi)-x'(\xi')))}\, d^4\sqrt{k'\pi}}{(i\sqrt{(k',\pi)}+\sqrt{m'\pi_0})(-i\sqrt{(k',\pi)}+\sqrt{m'\pi_0})}. \tag{19}$$

This mapping in effect means that the behavior of the particle in the external space is a manifestation of the behavior of the internal constituents and the motion of the particle is governed by the motion of the constituents in the internal space as a whole. Now as we have assumed that the internal space is anisotropic in nature so that a direction vector is fixed in the internal space, which gives rise to the internal helicity, and the two opposite helicities correspond to particle and antiparticle states, we can take that $i\sqrt{\pi}$ and $-i\sqrt{\pi}$ correspond to the different internal helicity states and denote two separate fields depicting particle and antiparticle states. So for a single helicity state depicting a particle (or antiparticle) state we should take $-i\sqrt{\pi}$ (or $i\sqrt{\pi}$) as a vanishing term. Taking $-i\sqrt{\pi} = 0$, we see that the expression (19) just reduces to the form

$$\langle\bar{\varphi}(x,t,\xi)\bar{\varphi}(x',t',\xi')\rangle = \frac{1}{(2\pi)^4}\int \frac{e^{i(k,(x-x'))}\, d^4k}{i\sqrt{k^2}+m}, \tag{20}$$

where we have choosen the unit $m = \pi_0 = 1$.

Now we can choose a matrix $(\gamma_\mu k_\mu + m) = (\not{k} + m)$ with two degenerate eigenvalues $\pm i\sqrt{k^2}+m$, which can be diagonalized by a unitary matrix U:

$$(\not{k}+m) = U^+ \begin{pmatrix} i\sqrt{k^2}+m & 0 & 0 & 0 \\ 0 & i\sqrt{k^2}+m & 0 & 0 \\ 0 & 0 & -i\sqrt{k^2}+m & 0 \\ 0 & 0 & 0 & -i\sqrt{k^2}+m \end{pmatrix} U. \tag{21}$$

Thus we just get the fermionic propagator in the Euclidean space-time

$$\langle\bar{\varphi}(x,t,\xi)\bar{\varphi}(x',t',\xi')\rangle = \frac{1}{(2\pi)^4}\int \frac{d^4k\, e^{i(k,(x-x'))}}{(\not{k}+m)}, \tag{22}$$

and the new field $\bar{\varphi}(x,t,\xi)$, where the anisotropic feature of the internal space is manifested by the internal handedness, depicts a fermionic field.

This shows that when a direction vector giving rise to an internal helicity in an anisotropic microlocal space-time is taken into account, we can have quantized fermionic field from Brownian motion processes.

In an earlier paper[11] it has been argued that the hydrodynamical quantization procedure is equivalent to stochastic quantization and as in the stochastic quantization procedure, the anisotropy is introduced in the internal space to have a fermion field, a similar situation arises in hydrodynamical quantization, too, where the quasi-irrotational (vortex) fluid motion is found to be the classical analog of a relativistic quantum mechanical system which gives rise to a fermionic field. In this picture, the classical analog of the relativistic fermionic field is the hydrodynamical motion with vorticity and only then, to define the motion around the vortex line, is the concept of circulation necessary. Precisely, the condition of circulation quantization is the condition of relativistic quantization, which gives rise to a fermion field, and a fermion can be thought of as a classically circular vortex.

III. INTERNAL HELICITY, NONLINEAR SIGMA MODEL, AND SKYRME SOLITON

From the above picture of realizing a fermion by introducing an anisotropy in the internal space so that a particular direction (say Z axis) is fixed or a vortex line in the liquid drop in Madelung fluid is introduced, it can be shown that fermions appear as solitons. This can be explicitly demonstrated by exploring the link between this geometrical for-

malism and the nonlinear sigma model. Since in this formalism a space-time point x_μ is depicted by the pair (x_μ, ξ_μ), where ξ_μ is an attached vector denoting the internal space-time variable, we can reformulate this picture by taking into account a complexified Minkowski space-time for which the coordinate is given by $Z_\mu = x_\mu + i\xi_\mu$. Now to introduce the effect of anisotropy in the internal space so that two internal helicities give rise to particle and antiparticle states, we use the formulation of twistor geometry where a vector x^a is written as a 2×2 complex matrix having the SL(2,c) group structure. Thus we write for the position vector Z^μ in complex space-time as

$$Z^\mu \to Z^{AA'} = X^{AA'} + i\xi^{AA'}. \tag{23}$$

Now $\xi^{AA'}$ is decomposed into two two-component spinorial variables by the relation $\xi^{AA'} = \overline{\theta}^A \theta^{A'}$. It can be shown that $\overline{\theta}^A(\theta^{A'})$ corresponds to two internal helicity states $+\frac{1}{2}(-\frac{1}{2})$ corresponding to particle and antiparticle states and is thus related to the fermion number.[14] Thus the position operator of a fermion becomes a non-Hermitian one as envisaged in Dirac equation and can be written as $Z^\mu \to Z^{AA'} = X^{AA'} + i\overline{\theta}^A \theta^{A'}$, where $X^{AA'}$ corresponds to the conventional position in space-time and $\overline{\theta}^A(\theta^{A'})$ corresponds to the internal helicity realized from the anisotropic nature of the internal structure and is related to the fermion number. With such a coordinate a particle can be viewed as moving in a superspace (x,θ), where supersymmetry is achieved in the case when θ is a Majorana spinor so that $\theta = \overline{\theta}$.

In this complexified space-time exhibiting the internal helicity states, we can write the metric tensor

$$g_{\mu\nu}(x,\theta) = g_{\mu\nu}(x)\overline{\theta}^A \theta^{A'}. \tag{24}$$

Now writing $g_{\mu\nu}(x) = e^i_\mu(x) e^j_\nu(x)\eta_{ij}$, where η_{ij} is the Minkowski metric, we write

$$\begin{aligned} g_{\mu\nu}(x,\theta) &= e^i_\mu(x)\overline{\theta}^A e^j_\nu(x)\theta^{A'}\eta_{ij} \\ &= \eta^{iA}_\mu(x)\eta^{jA'}_\nu(x)\eta_{ij}, \end{aligned} \tag{25}$$

where

$$\begin{aligned} \eta^{iA}_\mu(x) &= e^i_\mu(x)\overline{\theta}^A, \\ \eta^{jA'}_\nu(x) &= e^j_\nu(x)\theta^{A'}, \end{aligned} \tag{26}$$

and $\eta^{iA}_\mu(\eta^{jA'}_\nu)$ is a mixed quantity behaving as a spinor regarding the index A and tensor regarding indices i, μ.

The vierbein field e^i_μ is transformed as

$$e^i_\mu \to e'^i_\mu = [\Lambda(x)]^i_k e^k_\mu, \tag{27}$$

where $\Lambda(x)$ is an x-dependent Lorentz matrix. The spinor is transformed as

$$\overline{\theta}^A \to \overline{\theta}'^A = S(\lambda)^A_B \overline{\theta}^B, \tag{28}$$

where $S(\lambda)^A_B$ represents a SL(2,c) group operator. Using Eqs. (27) and (28), the transformation of η^{iA}_μ is given by

$$\begin{aligned} \eta^{iA}_\mu \to \eta'^{iA}_\mu &= [\Lambda(x)]^i_k S(\lambda)^A_B e^k_\mu \overline{\theta}^B \\ &= [A(x)]^{iA}_{kB}\eta^{kB}_\mu, \end{aligned} \tag{29}$$

where

$$[A(x)]^{iA}_{kB} = [\Lambda(x)]^i_k S(\lambda)^A_B, \tag{30}$$

and

$$\eta^{kB}_\mu = e^k_\mu \overline{\theta}^B.$$

In analogy with Yang–Mills gauge theory, we take

$$S(\lambda)^A_B = \delta^A_B, \tag{31}$$

and write

$$[A(x)]^{iA}_{kB} = [A(x)]^i_k. \tag{32}$$

Thus from (29) we have

$$\eta^{iA}_\mu \to \eta'^{iA}_\mu = [A(x)]^i_k \eta^{kA}_\mu. \tag{33}$$

Now if we assume that spinors η^{iA}_μ are scalars under coordinate transformation in superspace, i.e.,

$$\eta'^{iA}_\mu = \eta^{iA}_\mu, \tag{34}$$

we get

$$\eta^{iA}_\mu = [A(x)]^i_k \eta^{kA}_\mu. \tag{35}$$

In terms of the antisymmetric fundamental tensor $\epsilon_{i\mu}$, with the property $\epsilon_{i\mu}\epsilon^{i\mu} = 1$, we can write

$$\begin{aligned} \eta^{iA}_\mu &= \epsilon_{i\mu}\epsilon^{i\mu}[A(x)]^i \eta^A_\mu = \epsilon_{i\mu}[A(x)]^i \epsilon^{i\mu}\eta^A_\mu \\ &= [A(x)]_\mu \eta^{iA}. \end{aligned} \tag{36}$$

This equation (36) defines the four-vector A_μ whose components are 2×2 matrices and can be treated as a gauge potential. It is to be noted that η^{iA} represents a spin-$\frac{3}{2}$ particle.

We define a two-component spinor ξ^A_a $(a = 0,1)$ at each point of space-time. The components ξ^A_0 and ξ^A_1 represent $\overline{\theta}^A$ and $\theta^{A'}$, respectively, representing the internal helicity states. Here η^{iA} may be written in terms of ξ^A_a as

$$\eta^{iA} = \epsilon^{ij}\eta^A_{;j} = \epsilon^{ij}\xi^A_{a;j}, \tag{37}$$

where $\eta^A = \xi^A_a$ represents a two-component spinor and the symbol (;) denotes the covariant derivative.

In terms of this we can write

$$\eta^{iA}_{;\mu} = [A(x)]_\mu \epsilon^{ij}\xi^A_{a;j}, \tag{38}$$

which gives

$$\xi^A_{a;\mu} = [A(x)]_\mu \xi^A_a. \tag{39}$$

For convenience, the above relation can be rewritten as

$$\nabla_\mu \xi = A_\mu \xi, \tag{40}$$

where ξ is the (2×2) matrix whose elements are ξ^A_a. Now applying the commotator of covariant derivatives $[\nabla_\nu, \nabla_\mu]$ on ξ, we write

$$\begin{aligned} (\nabla_\nu\nabla_\mu - \nabla_\mu\nabla_\nu)\xi &= \{\partial_\nu A_\mu - \partial_\mu A_\nu + [A_\mu, A_\nu]\}\xi \\ &= F_{\mu\nu}\xi. \end{aligned} \tag{41}$$

Thus we have the field strength tensor $F_{\mu\nu}$ given in terms of the gauge fields A_μ and for zero curvature, we have the relation

$$F_{\mu\nu} = \partial_\nu A_\mu - \partial_\mu A_\nu + [A_\mu, A_\nu] = 0. \tag{42}$$

This zero curvature condition then implies that we can write the non-Abelian gauge field as

$$A_\mu = U^+ \partial_\mu U, \text{ where } U \in \mathrm{SL}(2,c). \tag{43}$$

Now substituting A_μ with $U^+\partial_\mu U$, we can write the Lagrangian corresponding to the equation of motion given by the relation (42),

$$L = M^2\,\mathrm{Tr}(\partial_\mu U^+\partial_\mu U) + \mathrm{Tr}[\partial_\mu UU^+,\partial_\nu UU^+]^2, \tag{44}$$

where M is a suitable constant having the dimension of mass.

Thus we find that the quantization of a Fermi field considering an anisotropy in the internal space leading to an internal helicity corresponds to the realization of a nonlinear sigma model where the Skyrme term in the Lagrangian ($\mathscr{L}_{\text{Skyrme}} = \mathrm{Tr}[\partial_\mu UU^+,\partial_\nu UU^+]^2$) automatically arises stabilizing the soliton. Indeed, this is no surprise as the anisotropic feature of the internal space prevents it from shrinking to zero size.

IV. TOPOLOGICAL PROPERTIES OF FERMIONIZATION

We have shown above that the anisotropic feature of the internal space leading to two internal helicity states which correspond to particle and antiparticle states gives rise to a quantized fermion. Again this feature helps us to depict a fermion as a Skyrme soliton. Now to investigate the topological properties of such a soliton, we may take into account that when the internal variable ξ_μ is attached to the space-time point x_μ, the space-time manifold is given by a de Sitter space $E(4,1)$. So we can study an O(5) nonlinear sigma model, which is characterized by a real unit vector $n(x)$ with the properties $n(x)\in\mathbb{R}^5$ and $n^+(x)n(x) = 1$. This O(5) nonlinear sigma model has been investigated by Felsager and Leinaas[15] in detail. We shall use their results here to show that this O(5) model can be decomposed into two three-space-dimensional solitons (Skyrme solitons) so that we can get two topological invariants corresponding to the fermion numbers $+1$ and -1.

The action of the O(5) nonlinear model is given by

$$S = \frac{1}{2}\int \|\mathbf{F}\|^2\,d^4x = \frac{1}{2}\int \|\mathbf{d}n\wedge\mathbf{d}n^+\|^2\,d^4x, \tag{45}$$

where $\mathbf{F}(x)$ is the SO(4) gauge field. Here $\|\mathbf{F}\|^2$ is defined as

$$\|\mathbf{F}\|^2 = -\tfrac{1}{2}\mathrm{Tr}[F_{\mu\nu}F^{\mu\nu}], \tag{46}$$

which can also be written as

$$S = -\frac{1}{2}\int \mathrm{Tr}(\mathbf{F}\wedge{}^*\mathbf{F}), \tag{47}$$

where * denotes Hodge's duality operation

$${}^*F_{\mu\nu} = \tfrac{1}{2}\epsilon_{\mu\nu\rho\sigma}F^{\rho\sigma}. \tag{48}$$

Here $n(x)$ is a point on a four-sphere S^4. Since $\mathbf{F}(x)$ is a SO(4) gauge field, and SO(4) can be decomposed locally as the product of two SU(2) groups,

$$\mathrm{SO}(4)\simeq\mathrm{SU}(2)\times\mathrm{SU}(2), \tag{49}$$

it is possible to decompose the SO(4) field $\mathbf{F}(x)$ into two SU(2) fields denoted by $\mathbf{F}_+(x)$ and $\mathbf{F}_-(x)$:

$$\mathbf{F}(x) = \mathbf{F}_+(x) + \mathbf{F}_-(x). \tag{50}$$

Using this definition, we find that

$$\|\mathbf{F}_\pm\|^2 = \tfrac{1}{2}\|\mathbf{F}\|^2, \tag{51}$$

and the action (45) can be written as

$$S = \int \|\mathbf{F}_\pm\|^2\,d^4x. \tag{52}$$

Now noting that in the four-sphere S^4, apart from the three-space dimension, the fourth-space dimension can be split into two segments $X_4 = \theta_+ + \theta_-$ so that θ_+ (θ_-) is defined in the positive (negative) segment of the axis, the anisotropic feature of the internal space as mentioned earlier can be realized and θ_+ (θ_-) corresponds to the positive (negative) ξ axis, where ξ is the internal space variable. This in effect gives rise to two internal helicity states corresponding to particle and antiparticle. Now noting that in this formalism $\mathbf{F}_+$ ($\mathbf{F}_-$) is defined only in the θ_+ (θ_-) region of space, we can decompose the action (52) as

$$S = S_1 + S_2 = \tfrac{1}{2}\left[\int \|\mathbf{F}_+\|^2\,d^3x\,d\theta_+ + \int \|\mathbf{F}_-\|^2\,d^3x\,d\theta_-\right], \tag{53}$$

where $\mathbf{F}_+$ ($\mathbf{F}_-$) is SU(2) gauge field defined in S^3. This splitting can also be realized from the fact that since $\mathbf{F}(x)$ is defined in S^4 we can decompose S^4 into $S^4\to S^3\times S^1$. Since S^1 is a doubly connected space we can take S^3 into two disconnected regions so that we can write $S^4 = S^3_+ + S^3_-$. Now since $\mathbf{F}_+$ ($\mathbf{F}_-$) is defined in S^3 we get two three-space-dimensional solitons from this and this disconnectedness is ensured by the Skyrme term which stabilizes the solitons so that these cannot shrink to zero size.

Now following Felsager and Leinaas[15] we can relate the Pontryagin density $\lambda(x)$ and the Euler density $\chi(x)$ of the SO(4) gauge field with the Pontryagin densities $\lambda_+(x)$ and $\lambda_-(x)$ of the two SU(2) gauge fields. The Pontryagin density decomposes in the following way

$$\begin{aligned}\lambda(x) &= -(1/16\pi^2)\mathrm{Tr}(\mathbf{F}\wedge\mathbf{F})\\ &= (1/16\pi^2)(\mathbf{F}|\mathbf{F})\\ &= (1/16\pi^2)(\mathbf{F}_+|\mathbf{F}_+) + (1/16\pi^2)(\mathbf{F}_-|\mathbf{F}_-)\\ &= \lambda_+(x) + \lambda_-(x).\end{aligned} \tag{54}$$

Similarly the Euler density is given by

$$\begin{aligned}\chi(x) &= -(1/16\pi^2)\mathrm{Tr}({}^*\mathbf{F}\wedge\mathbf{F})\\ &= (1/16\pi^2)({}^*\mathbf{F}|\mathbf{F})\\ &= (1/16\pi^2)(\mathbf{F}_+|\mathbf{F}_+) - (1/16\pi^2)(\mathbf{F}_-|\mathbf{F}_-)\\ &= \lambda_+(x) - \lambda_-(x).\end{aligned} \tag{55}$$

Since the Pontryagin density of the SO(4) gauge field vanishes identically since $\mathbf{F}\wedge\mathbf{F} = 0$, we find

$$\lambda_+(x) = -\lambda_-(x) = \tfrac{1}{2}\chi(x). \tag{56}$$

Thus the Pontryagin densities of the two SU(2) gauge fields are of opposite sign and in magnitude equal to half the Euler density of the SO(4) gauge field.

The winding number associated with the mapping $\mathbb{R}^4\to S^4$ is given by

$$m = \frac{1}{(8\pi)^2}\int_{\mathbb{R}^4} \epsilon_{ijklm}\,n^i\,\mathbf{d}n^j\wedge\mathbf{d}n^k\wedge\mathbf{d}n^l\wedge\mathbf{d}n^m. \tag{57}$$

The Euler number χ is given by[15]

$$\chi = -\frac{1}{16\pi^2}\int_{\mathbf{R}^4} \mathrm{Tr}(^*\mathbf{F}\wedge\mathbf{F}) = 2m. \tag{58}$$

From this we find that the Pontryagin number of the two SU(2) gauge fields is given by

$$\lambda_+(x) = m, \quad \lambda_-(x) = -m. \tag{59}$$

This suggests that the Pontryagin number of the two SU(2) gauge fields actually corresponds to the fermion numbers $+1$ and -1. Thus the topological properties of fermionization give rise to two Pontryagin invariants corresponding to the fermion and the antifermion.

V. DISCUSSION

We have shown above that the stochastic quantization of a fermion introducing an anisotropy in the internal space so that this gives rise to two internal helicities corresponding to particle and antiparticle states leads us to describe a fermion as a Skyrme soliton. The quartic Skyrme term appears here as a consequence of this internal anisotropy. In this sense, the Skyrme term may be taken as a quantum effect. Indeed Pak and Tze[3] have also pointed out that the Skyrme term may be viewed to have a quantum origin and have shown that by giving rise to an effective interaction the radiative corrections in the $SU(2)\times SU(2)$ invariant model not only make solitons with topological fermion number but also fermionic spin states.

In recent times a lot of work is being done to interpret a nucleon as a Skyrme soliton on the basis of the chiral model. From the present analysis it appears that the chiral model is a specific case and all massive fermions can be considered as Skyrme solitons when the Skyrme term arises in the quantization procedure. Indeed the extension to the soliton sectors of chiral theories invariant under $SU(N)\times SU(N)$, $N>2$, may be achieved by embedding $SU(2)\times SU(2)$ in larger groups when the special role of the topological properties of the $SU(2)\times SU(2)$ subgroup incorporating the correlation between space-time and internal symmetries become transparent. Specifically this may help us to have a geometrical origin of the internal symmetry and a dynamical understanding of the symmetry breaking of a $SU(3)\times SU(3)$ theory.

It may be pointed out that since the fermionization procedure of the stochastic quantization method incorporates universal Brownian motion in the internal space also apart from the external space, the splitting of the soliton field into a point singularity plus cloud should be abandoned and the extension of the soliton plays the dominant role. However a latticization may be possible in the manner of Polyakov's work on the two-dimensional Heisenberg ferromagnet.[16] In fact when the four-sphere S^4 is split into $S^2\times S^2$ instead of $S^3\times S^1$ as in the present case, the O(5) nonlinear sigma model gives a lattice picture when finite energy static solutions in two space dimensions such as vortices with finite energy per unit length appears. This can be related with Liouville field theory and a Polyakov string through a correspondence between the sigma model and the Liouville model.[17]

Finally, we may point out that in $(1+1)$ dimensions, Coleman[18] has shown explicitly the equivalence of the sine–Gordon soliton with the massive Thirring fermion. The fermionization procedure discussed here may be considered here as a generalization of Coleman's significant result in $(3+1)$ dimensions so that the stochastic quantization of a fermion leads to view them as Skyrme solitons in three space dimensions.

[1] T. H. R. Skyrme, Proc. R. Soc. London Ser. A **260**, 127 (1961); Nucl. Phys. **31**, 556 (1962); J. Math. Phys. **12**, 1735 (1971).
[2] J. G. Williams, J. Math. Phys. **11**, 2611 (1970).
[3] N. K. Pak and Ch. H. Tze, Ann. Phys. (NY) **117**, 164 (1979).
[4] J. M. Gipson and Ch. H. Tze, Nucl. Phys. B **183**, 524 (1981).
[5] E. Witten, Nucl. Phys. B **223**, 422, 433 (1983).
[6] A. P. Balchandran, V. P. Nair, S. G. Rajeev, and A. Stern, Phys. Rev. Lett. **49**, 1124 (1982); Phys. Rev. D **27**, 1153 (1983).
[7] D. Finkelstein and J. Rubinstein, J. Math. Phys. **9**, 1762 (1968).
[8] J. Mickelsson, Phys. Rev. D **30**, 1843 (1984).
[9] P. Bandyopadhyay, preprint.
[10] E. Nelson, *Dynamical Theories of Brownian Motion* (Princeton U.P., Princeton, NJ, 1967).
[11] K. Hajra, preprint.
[12] S. M. Moore, J. Math. Phys. **21**, 2102 (1980).
[13] F. Guerra and P. Ruggiero, Phys. Rev. Lett. **31**, 1022 (1973).
[14] P. Bandyopadhyay, preprint.
[15] B. Felsager and J. M. Leinaas, Ann. Phys (NY) **130**, 461 (1980).
[16] P. Bandyopadhyay, and K. Hajra, preprint.
[17] H. Bohr, B. O. Hon, and S. Saito, Nuovo Cimento A **84**, 237 (1984).
[18] S. Coleman, Phys. Rev. D **11**, 2088 (1975).

Nonperturbative confinement in quantum chromodynamics. IV. Improved treatment of Schoenmaker's equation

D. Atkinson and M. Koopmans
Institute for Theoretical Physics, University of Groningen, P. O. Box 800, 9700 AV Groningen, The Netherlands

(Received 17 June 1986; accepted for publication 22 October 1986)

An improved ansatz for the three-gluon vertex function is treated; and it is shown that the gluon propagator has a double pole at the origin of the p^2 plane, as well as a tachyon on the spacelike real axis, at least in this approximation.

I. INTRODUCTION

The present paper constitutes the conclusion of a study of the infrared behavior of the gluon propagator in the Landau gauge. In previous papers,[1] to which we shall refer as I, II, and III, respectively, we investigated an approximation scheme for the gluon propagator Dyson–Schwinger equation that was initiated by Mandelstam.[2]

The basic idea of the approximations is to truncate the Dyson–Schwinger equation by introducing an ansatz for the three-gluon vertex that (a) involves only the propagator itself, and (b) is inspired by (Mandelstam's ansatz I and II), or is strictly consistent with (Schoenmaker's ansatz, III) the Slavnov–Taylor identity. It has been argued that the longitudinal part of the vertex function (i.e., the part that contributes to the Slavnov–Taylor identity) need not be relevant to the Dyson–Schwinger equation. Indeed Gardner[3] constructed a model in which the vertex function consists of two parts, one of which contributes only to the Slavnov–Taylor identity, while the other contributes only to the Dyson–Schwinger equation. However, it can be shown that the two parts of Gardner's ansatz do not have the same scaling properties under renormalization: one part has the correct number of factors Z_3, while the other does not. Accordingly, we may reject the Gardner ansatz as a valid criticism of the general method. On the other hand, a recent paper of Zhang[4] is much more convincing. He shows that the contribution of the longitudinal part of the vertex function reduces essentially to a term that is indistinguishable from a tadpole, so that the nontrivial parts of the approximate Dyson–Schwinger equation arise from transverse contributions to the vertex function. Both Gardner's and Zhang's treatments apply specifically to an axial gauge propagator, which has special properties (in particular, orthogonality to the gauge vector); and hence they are not relevant to a study in the Landau gauge.

In III we studied an improved version of the Mandelstam ansatz, but with a simplification that allowed us to reduce the equation to a fourth-order nonlinear differential equation. In this paper, we complete this analysis by removing the simplification, which involves us in the treatment of a sixth-order nonlinear differential equation. This equation is subjected to numerical analysis as it stands: we confirm the p^{-4} behavior as $p \to 0$, p being the gluon momentum; and we also find a tachyon state, as in III. There are probably no first-sheet complex branch points—a deficiency of the model of I and II—although there is a neighborhood of the origin that is inaccessible to the computer, because of large cancellations, so one cannot be completely sure. In order to remove these cancellations, the sixth-order differential equation is transformed into an integral equation by two successive implementations of the method of variation of parameters; and in this form the equation is suited to a rigorous demonstration of the existence of a solution. We show that the solution is analytic in a (cut) neighborhood of the origin, so that an accumulation of first-sheet complex branch points is excluded.

II. NUMERICAL ANALYSIS OF DIFFERENTIAL EQUATION

The form factor multiplying the bare gluon propagator (see III) can be written

$$F(x) = A/x + \gamma x + \phi(x)x^3, \tag{2.1}$$

where A is an unknown constant (which, however, can be scaled away), where

$$\gamma = \tfrac{240}{389} \approx 0.617, \tag{2.2}$$

and where $\phi(x)$ is a function that satisfies the nonlinear integral equation

$$\begin{aligned} x^6 G(x) = & -\frac{1}{4}\int_0^x dy (x-y)^3 (x^2 + 10xy + y^2)\phi(y) \\ & + \frac{1}{8}\int_0^{x/4} dy\, x^{3/2}(x-4y)^{3/2} \\ & \times (x^2 + 20xy + 12y^2)\phi(y), \end{aligned} \tag{2.3}$$

where

$$G(x) = [\gamma + x^2\phi(x)]/[1 + \gamma x^2 + x^4\phi(x)]. \tag{2.4}$$

The details can be found in III.

The second integral in (2.3), involving the surd, is rather awkward; but fortunately it has been shown numerically,[5] by means of cubic splines, that the truncated equation

$$x^6 G(x) = -\frac{1}{4}\int_0^x dy (x-y)^3 (x^2 + 10xy + y^2)\phi(y) \tag{2.5}$$

has a solution that resembles that of the full equation (2.3), the difference being not qualitative, but merely quantitative and relatively minor. In order to study the qualitative properties of the solution, it suffices to look at (2.5). We shall

 0022-2488/87/030717-03$02.50

accordingly subject this equation to numerical analysis and to a rigorous proof of existence.

The integral equation can be converted into a differential equation, namely

$$\left[\frac{d}{dx}\right]^6 [x^6 G(x)] + 18x^2\phi''(x) + 144x\phi'(x) + 138\phi(x) = 0. \tag{2.6}$$

This equation has been treated numerically (in double precision) by the Runge–Kutta method. As usual, an asymptotic series expansion must be made in a neighborhood of the infrared point $x = 0$.

The results of the numerical work[6] can be summarized as follows.

(a) There are complex branch points on secondary Riemann sheets that are connected to the principal Riemann sheet through the timelike cut along $-\infty < x < 0$. There seem to be no complex branch points on the principal Riemann sheet, although this result is not completely conclusive, since a small region around the infrared point is inaccessible, due to large cancellations.

(b) There is a (ghost) pole on the spacelike axis at

$$x = x_p \equiv 2.831, \tag{2.7}$$

much as in case III. This may be a signal that our vacuum is unstable, or it may be merely an artifact of our approximations.

III. REFORMULATION OF THE EQUATION

In this section, we reformulate Eq. (2.5) in such a way that there are no cancellations in the infrared region. Using such a form, we shall outline the proof that a solution exists (by means of the Banach Theorem). Moreover, we could also set up a computer analysis that is not plagued by cancellations (however, we have not done this).

The necessary manipulations are tedious, and we shall merely sketch the method here. Further details can be found in Ref. 6. Equation (2.6) is a nonlinear sixth-order differential equation. We transform it in two steps. First, observe that the linear, homogeneous equation

$$18x^2\psi''(x) + 144x\psi'(x) + 138\psi(x) = 0 \tag{3.1}$$

has the two solutions

$$\psi_\pm(x) = x^{\pm\beta - 7/2}, \tag{3.2}$$

where $\beta = (165)^{1/2}/6$. We resolve Eq. (2.6) by the method of variation of parameters, treating the first term, involving a sixth-order derivative of $x^6 G$, as an inhomogeneity. The result is an integral over this derivative, multiplied by a kernel. The six derivatives can be removed by six partial integrations, the result being

$$x^4 G''''(x) + 20x^3 G'''(x) + \frac{355}{3}x^2 G''(x) + \frac{680}{3}xG'(x) + \frac{865}{9}G(x) = -18\phi(x) - \frac{455}{54\beta}x^{-7/2} \times \int_0^x dy\, y^{5/2} \left(\left(\frac{y}{x}\right)^\beta - \left(\frac{x}{y}\right)^\beta G(y)\right). \tag{3.3}$$

The second step consists in resolving this fourth-order nonlinear integrodifferential equation by applying the method of variation of parameters again. In order to do this expeditiously, we add terms proportional to x^2G'', xG', and G to both sides of (3.3), as well as $18G/x^2$, the latter being the most singular part of $18\phi(x)$. The corresponding homogeneous equation is

$$x^6 H'''' + 20x^5 H''' + \frac{975}{8}x^4 H'' + 225x^3 H' + \left[\frac{36465}{256}x^2 + 18\right]H = 0, \tag{3.4}$$

with the four solutions

$$H_i(x) = x^{-11/4}\exp[\beta_i x^{-1/2}], \tag{3.5}$$

$$\beta_1 = (72)^{1/4}(1+i), \quad \beta_2 = (72)^{1/4}(1-i), \quad \beta_3 = (72)^{1/4}(-1+i), \quad \beta_4(72)^{1/2}(-1-i). \tag{3.6}$$

These homogeneous solutions are used to resolve Eq. (3.3), the result being

$$G(x) = \sum_{i=1}^4 G_i(x), \tag{3.7}$$

where

$$G_i(x) = \tfrac{1}{2}72^{-1}\beta_i x^{-11/4}\exp(\beta_i x^{-1/2}) \times \int_0^x dy\, y^{5/4}\exp(-\beta_i y^{-1/2})\Sigma(y), \tag{3.8}$$

with

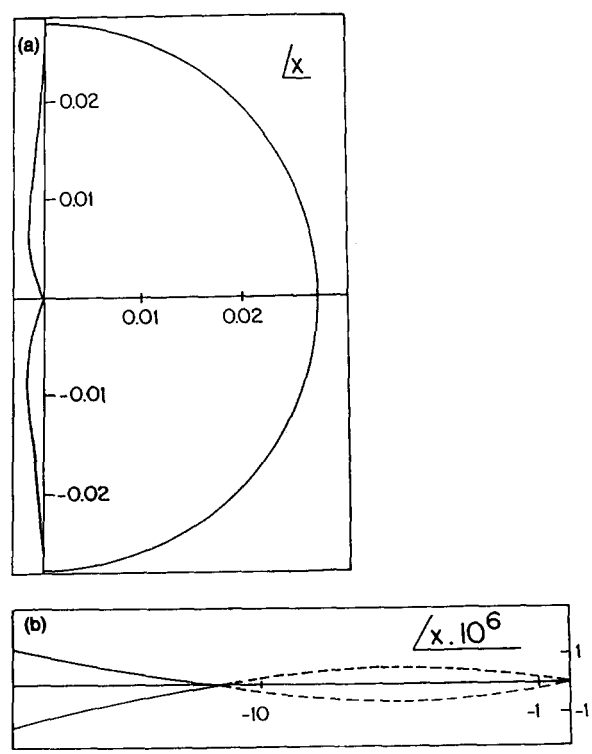


FIG. 1. Sketch of the domain of applicability of the Banach Theorem. (a) The full domain. (b) The region near the negative imaginary axis on a larger scale. Note that the boundary of the region of proved analyticity intercepts the negative real axis and penetrates the second Riemann sheet (dotted lines). Thus the origin cannot be an accumulation point of first-sheet singularities.

$$\Sigma(x) = 18\gamma + \frac{106745}{2304}x^4 G''(x) + \frac{85}{3}x^3 G'(x) + \frac{85}{24}x^2 G(x) + \frac{18x^2 G^2(x)}{x^2 G(x) - 1} - \frac{455}{54\beta}x^{-3/2} \times \int_0^x dy\, y^{5/2} \left(\left(\frac{y}{x}\right)^\beta - \left(\frac{x}{y}\right)^\beta G(y) \right). \quad (3.9)$$

The first- and second-order derivatives in Eq. (3.9) can be removed by partially integrating Eq. (3.8); but there is a subtlety. One has to deform the contour $(0,x)$ in such a way that the boundary term at $y = 0$ vanishes; it turns out to be sufficient to ensure that the contour for G_2 and G_3 approaches the origin along the positive imaginary axis, while that for G_1 and G_4 approaches it along the negative imaginary axis. Such deformations are allowed if the G_i are analytic in the cut plane; and we can show that this is the case.

After this adjustment, Eqs. (3.7)–(3.9) constitute a nonlinear integral equation without derivatives. It is eminently suited to an existence proof by means of the contraction mapping (Banach) principle. As usual, the integration interval $(0,x)$ is transformed to $(0,\infty)$; and one finally proves that a unique solution $G(x)$ exists, that is analytic in a half-circle in the right half-plane and in a curved region in the left half-plane (see Fig. 1). It is most significant that this curved region crosses the cut $(-\infty,0)$ and penetrates the second Riemann sheet. Thus the infrared singularity is certainly *not* the accumulation point of first-sheet complex branch points (as it was in I and II). Further details can be found in Ref. 6.

[1] D. Atkinson, J. K. Drohm, P. W. Johnson, and K. Stam, J. Math. Phys. **22**, 2704 (1981); D. Atkinson, P. W. Johnson, and K. Stam, *ibid.* **23**, 1917 (1982); D. Atkinson, H. Boelens, S. J. Hiemstra, P. W. Johnson, W. J. Schoenmaker, and K. Stam, *ibid.* **25**, 2095 (1984).

[2] S. Mandelstam, Phys. Rev. D **20**, 3223 (1979).

[3] E. J. Gardner, J. Phys. G: Nucl. Phys. **9**, 139 (1983). See also, B. K. Jennings and R. M. Woloshyn, *ibid.* **9**, 997 (1983).

[4] R. B. Zhang, Phys. Rev. D **31**, 1512 (1985).

[5] A. de Winkel, "Numerical evaluation of the improved Mandelstam equation for the gluon propagator by an iterative method," Groningen internal report No. 195, 1984.

[6] M. Koopmans, "Confinement in non-perturbative QCD. Implications of the improved Mandelstam equation," Groningen internal report No. 213, 1985.

Self-energy operator for an electron in an external Coulomb potential

Levere C. Hostler
Physics Department, Wilkes College, Wilkes Barre, Pennsylvania 18766

(Received 24 June 1986; accepted for publication 5 November 1986)

The self-energy operator for an electron in an external Coulomb potential is investigated analytically using a mass eigenfunction expansion concept reported earlier. Contour integration techniques in the complex m^2 plane are used to combine bound state and continuum contributions into a single integral. The result is a relatively simple integral representation for the mass operator. Only terms ignoring the "shift correction" are considered in this preliminary study. A transformation to a basis of relativistic Coulomb Sturmian functions exhibits the $Z\alpha$ dependence of the integrand in a strikingly simple way. The entire investigation is set in the framework of the "scalar formalism" for quantum electrodynamics investigated earlier by a number of authors and based on the "second-order" Dirac equation, $\{\Pi\cdot(1+i\sigma)\cdot\Pi+m^2\}\Phi=0$, where Φ is a 2×1 Pauli spinor.

I. INTRODUCTION

A mass eigenfunction expansion concept reported earlier[1] is applied in an analytical investigation of the mass operator for an electron in an external Coulomb potential. The investigation is geared toward a future calculation of the Lamb shift, obtained from the mass operator by taking the expectation value in an unperturbed state. It is this application that gives the results presented here their physical interest. In addition, the application illustrated here of the mass eigenfunction expansion and the relativistic Coulomb Sturmian basis may provide a guide for other relativistic perturbation calculations, for example for positronium calculations.

All calculations are set in the framework of a "scalar formalism" for quantum electrodynamics (QED) investigated earlier by a number of authors.[2-13] The scalar formalism for QED is based on the "second-order" Dirac equation[14]

$$\{\Pi\cdot(1+i\sigma)\cdot\Pi+m^2\}\Phi=0,\qquad \Pi\equiv-i\partial_\mu-eA_\mu, \tag{1.1}$$

in which Φ is a 2×1 Pauli spinor, and σ is the Lorentz spin tensor

$$\sigma_{\mu\nu}\equiv\left[\begin{array}{ccc|c} 0 & \sigma_3 & -\sigma_2 & \sigma_1 \\ -\sigma_3 & 0 & \sigma_1 & \sigma_2 \\ \sigma_2 & -\sigma_1 & 0 & \sigma_3 \\ \hline -\sigma_1 & -\sigma_2 & -\sigma_3 & 0 \end{array}\right], \tag{1.2}$$

where $\sigma_{1,2,3}$ denote the ordinary 2×2 Pauli spin matrices. The use of Eq. (1.1) instead of the usual linear Dirac equation to describe the electron brings out a close similarity between the quantum theory of a spin-$\frac{1}{2}$ particle and the quantum theory of a simple scalar particle. Together with the small dimension of the matrices involved this suggests that calculations with the scalar formalism could lead to some simplification compared to the use of the linear Dirac equation. For example a positronium wave function in the scalar formalism is a 2×2 matrix as opposed to a 4×4 matrix in conventional quantum electrodynamics.

The form of the second-order Dirac equation, Eq. (1.1), suggests the possibility of viewing that equation as an eigenvalue equation for a "mass operator"

$$\Lambda\equiv-\Pi\cdot(1+i\sigma)\cdot\Pi. \tag{1.3}$$

To be a useful concept Λ should in some sense be self-adjoint and have a complete orthogonal set of eigenfunctions. Such an eigenvalue problem has been investigated before in Ref. 1. There it has been shown that Λ is self-adjoint with respect to the not positive definite inner product

$$(\Phi_B;\Phi_A)\equiv\int d^4x\,\bar{\Phi}_B\Phi_A, \tag{1.4}$$

in which

$$\bar{\Phi}=\Phi^\dagger(-i\overleftarrow{\Pi}_4-\sigma\cdot\overleftarrow{\Pi}) \tag{1.5}$$

denotes the "dual" state associated with Φ. As discussed in Ref. 1, if the inner product (1.4) were positive definite; then all the conditions of the spectral theorem would be met and the existence of a complete orthogonal set of eigenfunctions of Λ would be guaranteed. It is shown in Ref. 1 that in the Coulomb case and in some other simple cases a complete orthogonal set of eigenfunctions exists in spite of the lack of positive definiteness. Such an eigenfunction expansion theorem can play a role in relativistic perturbation theory analogous to the role of energy eigenfunction expansions in nonrelativistic perturbation theory.

Mass eigenfunction expansions, the relativistic Kepler problem of the second-order Dirac equation, and the scalar formalism for QED have been treated in detail elsewhere. In the interest of brevity the reader is referred to this earlier work for background material, especially Refs. 1, 11, and 12.

Section II begins with the Feynman integral, Eq. (2.1), for the irreducible electron self-energy graph including the effects of the Coulomb interaction with the nucleus. A notation[15] is used in which the electron propagator $S_F(2,1)$ is visualized as the coordinate space representative of an abstract operator:

$$S_F(2,1)\equiv\langle2|1/((-i\partial_\mu)^2+m^2)|1\rangle.$$

A compact formal expression, Eq. (2.5), for the self-energy

 0022-2488/87/030720-08$02.50

graph ignoring "shift corrections" is obtained by performing the Feynman integrals explicitly. These expressions involve rather complicated operator functions of

$$\rho \equiv \Lambda/m^2. \tag{1.6}$$

An example is the function $\ln(1-\rho)/\rho$. Such operator functions of ρ are replaced by corresponding c-number expressions by inserting a complete orthogonal set of eigenfunctions of Λ.

The sum over bound states and integration over continuum states that arise in such an application of the mass eigenfunction expansion theorem are combined into a single integral in Sec. III A by use of contour integration techniques in the complex ρ' plane, where ρ' denotes the eigenvalue of ρ. An example of this is the integral representation [Eq. (3.3) repeated here for convenience]

$$\left[\frac{\ln(1-\rho)}{\rho}\right]' = -\int_1^\infty \frac{d\rho'}{\rho'}\left[\frac{1}{\rho'-\rho} - \epsilon_0 \frac{|\Phi_0\rangle\langle\bar{\Phi}_0|}{\rho'-1}\right].$$

This result coupled with the analogous equations (3.4) and (3.5) make possible an integral representation of the mass operator in which ρ appears essentially only through the relativistic Coulomb Green's function $1/(\rho'-\rho)$. Known representations of the relativistic Coulomb Green's function can thus be brought to bear on the problem. In Sec. III B a further transformation to a ζ representation is carried out. This transformation is designed to exhibit the $Z\alpha$ dependence of the integrand. This $Z\alpha$ dependence enters the integrand in a strikingly simple way when all is referred to a basis of relativistic Coulomb Sturmian functions. This is done in Sec. IV. The final integral representation of the factor $\ln(1-\rho)/\rho$ is given by Eq. (4.12). The nuclear charge enters the final integrand essentially only through two parameters: the $Z\alpha$ which appears in the "relativistic orbital angular momentum quantum number" $\gamma \equiv ((J+\frac{1}{2})^2 - (Z\alpha)^2)^{1/2}$, if $L = J + \frac{1}{2}$, and $\gamma \equiv ((J+\frac{1}{2})^2 - (Z\alpha)^2)^{1/2} - 1$, if $L = J - \frac{1}{2}$; and through a second parameter, q_0, where mq_0 corresponds to the classical relativistic electron momentum in the bound state being perturbed,

$$q_0 = \frac{Z\alpha/(\gamma_0 + n_0)}{[1 + (Z\alpha/(\gamma_0 + n_0))^2]^{1/2}}. \tag{1.7}$$

Using the ζ representation in a Sturmian basis, the integrals corresponding to the individual partial waves can be shown to be analytic functions of q_0 in the entire complex plane cut along the negative real axis, $-\infty < q_0 \leqslant 0$. The ζ representation in the Sturmian basis seems quite suitable for a future numerical evaluation of the Lamb shift. Also, the relatively simple form of the final integral representation in Sec. IV invites further analytical development. The method developed here would bear some resemblance to the earlier Lamb-shift calculation of Lieber.[16] However, in contrast to the method of Lieber, the method developed here could be fully relativistic. Other recent work on the Lamb shift that should be mentioned is that of Mohr and of Sapirstein.[17]

II. PRELIMINARY REDUCTIONS

The scalar formalism for quantum electrodynamics[12] gives the following expression for the electron self-energy operator Σ, defined through the equation $S_F' \equiv 1/(\Pi\cdot(1+i\sigma)\cdot\Pi + m^2 - \delta(m^2) + \Sigma)$:

$$\begin{aligned}\Sigma = 4\pi i\alpha \int \frac{d^4k}{(2\pi)^4}\frac{1}{k^2 - i\epsilon} \\ \times[\Pi\cdot(1+i\sigma) + (1+i\sigma)\cdot(\Pi - k)]_\mu \\ \times[(\Pi - k)\cdot(1+i\sigma)\cdot(\Pi - k) + m^2 - i\epsilon]^{-1} \\ \times[(\Pi - k)\cdot(1+i\sigma) + (1+i\sigma)\cdot\Pi]_\mu.\end{aligned} \tag{2.1}$$

In this equation

$$\Pi_4 \equiv i(E_0 + (Z\alpha/r)), \quad \vec{\Pi} \equiv -i\nabla - e\vec{A},$$

and

$$E_0 \equiv m/(1 + (Z\alpha/(\gamma_0 + n_0))^2)^{1/2}$$

equals the energy of the level whose level shift will be sought when Σ is used in a Lamb-shift calculation. Using the Feynman denominator combining equations in a standard way gives

$$\begin{aligned}\Sigma = 4\pi i\alpha \int \frac{d^4k}{(2\pi)^4}\int_0^1 d\lambda\,[\Pi\cdot(2-\lambda(1-i\sigma)) - (k-\lambda\Pi)\cdot(1-i\sigma)] \\ \times\{[(k-\lambda\Pi)\cdot(1+i\sigma)\cdot(k-\lambda\Pi) - \lambda(1-\lambda)\Lambda + \lambda m^2]^2\}^{-1} \\ \times[-(1-i\sigma)\cdot(k-\lambda\Pi) + (2-\lambda(1-i\sigma))\cdot\Pi].\end{aligned} \tag{2.2}$$

In this equation Λ is the operator of Eq. (1.3). In order to obtain closed analytic expressions to work with, an expansion in "shift corrections" is carried out as in the paper of Erickson and Yennie.[18] Accordingly, Σ is written $\Sigma = \Sigma_0 + R_0$, where

$$\begin{aligned}\Sigma_0 = 4\pi i\alpha \int \frac{d^4k}{(2\pi)^4}\int_0^1 d\lambda\,[\Pi\cdot(2-\lambda(1-i\sigma)) - k\cdot(1-i\sigma)] \\ \times[k^2 - \lambda(1-\lambda)\Lambda + \lambda m^2]^{-2}[-(1-i\sigma)\cdot k + (2-\lambda(1-i\sigma))\cdot\Pi],\end{aligned} \tag{2.3}$$

and

$$\begin{aligned}R_0 = 4\pi i\alpha \int \frac{d^4k}{(2\pi)^4}\int_0^1 d\lambda \int_0^1 d\xi \frac{\partial}{\partial\xi}[\Pi\cdot(2-\lambda(1-i\sigma)) - (k-\xi\lambda\Pi)\cdot(1-i\sigma)] \\ \times[(k-\xi\lambda\Pi)\cdot(1+i\sigma)\cdot(k-\xi\lambda\Pi) - \lambda(1-\lambda)\Lambda + \lambda m^2]^{-2} \\ \times[-(1-i\sigma)\cdot(k-\xi\lambda\Pi) + (2-\lambda(1-i\sigma))\cdot\Pi].\end{aligned} \tag{2.4}$$

In this preliminary study only the zero-order term Σ_0 is considered. Standard Feynman techniques can be applied to evaluate

the integral (2.3), with renormalization carried out as in scalar electrodynamics.[19,20] It has been convenient to use an intermediate renormalization, with subtraction point $p_\mu p_\mu = 0$. The result is

$$\Sigma_0 = \frac{-\alpha}{4\pi}\left[4\Pi\cdot(1-\rho)\frac{\ln(1-\rho)}{\rho}\cdot\Pi + \Pi\cdot(1-\rho)\frac{(1-\rho)\ln(1-\rho)+\rho}{\rho^2}(1-i\sigma)\cdot\Pi\right.$$
$$+\Pi\cdot(1-i\sigma)\frac{(1-\rho)\ln(1-\rho)+\rho}{\rho^2}(1-\rho)\cdot\Pi$$
$$+\frac{1}{3}\Pi\cdot(1-i\sigma)(1-\rho)\frac{(1-\rho)^2\ln(1-\rho)+\rho-\frac{3}{2}\rho^2}{\rho^3}(1-i\sigma)\cdot\Pi$$
$$+\frac{1}{12}(1-i\sigma)_{\mu\nu}m^2\rho(1-\rho)\frac{(1-\rho)^2\ln(1-\rho)+\rho-\frac{3}{2}\rho^2}{\rho^3}(1-i\sigma)_{\nu\mu}$$
$$\left.-\frac{1}{36}(1-i\sigma)_{\mu\nu}m^2(1-\rho)(1-i\sigma)_{\nu\mu}+\frac{31}{9}m^2(1-\rho)-3m^2+\frac{11}{9}e\sigma_{\mu\nu}F_{\mu\nu}\right]. \tag{2.5}$$

No special effort has been made to exhibit the explicit factor of $(Z\alpha)^4$ known to appear in the lowest non-zero-order terms contributing to the Lamb shift.[21] That factors of $Z\alpha$ are implicit in expression (2.5) when used in a Lamb-shift calculation is evidenced by the factors $(1-\rho)$, which can be converted into commutators after forming the expectation value in the unperturbed state $|\Phi_0\rangle$: the identities $\langle\bar{\Phi}_0|O(1-\rho) = \langle\bar{\Phi}_0|[\rho;O]$, and $(1-\rho)O|\Phi_0\rangle = -[\rho;O]|\Phi_0\rangle$, are a result of $|\Phi_0\rangle$ being an eigenstate of ρ with eigenvalue 1.

The Lamb shift may now be calculated in principle by inserting a complete set of mass eigenfunctions and taking the expectation value of Eq. (2.5). All expressions in Eq. (2.5) involving the operator ρ in complicated forms thereby become just c-number expressions. Equations for implementing such a program in a Sturmian representation are given in the Appendix. However, the further reductions to be carried out in Secs. III and IV are expected to lead to a more efficient Lamb-shift calculation.

III. INTEGRAL REPRESENTATIONS OF THE MASS OPERATOR

A. Analytic continuation in m^2

A direct use as described above of the mass eigenfunction expansion to evaluate the expectation value of Eq. (2.5) would entail summation over bound state eigenvalues $\rho' = (E_0/m)^2(1+(Z\alpha/(\gamma+n))^2)$ and integration over continuum state eigenvalues $\rho' = (E_0/m)^2 - (p/m)^2$, $0<p<\infty$ (see Ref. 1 for the derivation of this spectrum). Next, a method of combining these two types of contributions into a single integral will be described. This method involves applying the Cauchy integral formula to the operator structures $\ln(1-\rho)/\rho$, etc. appearing in Eq. (2.5). First note that the function $\ln(1-\rho')/\rho'$ of the c-number variable ρ' has a removable singularity at $\rho' = 0$. Accordingly, $\ln(1-\rho')/\rho'$ is analytic and single valued in the entire complex plane cut along the segment $1\leqslant\rho'<\infty$. Figure 1 shows this cut. The bound state eigenvalues of ρ are indicated by the dots in Fig. 1. The continuous spectrum of ρ is at the same time a branch cut for the function $\eta \equiv -ip$, $p\equiv((E_0)^2 - m^2\rho')^{1/2}$. This is the other cut in Fig. 1. Feynman boundary conditions are implemented as usual by assuming that m^2 has a small negative imaginary part. This small negative imaginary part places any bound state eigenvalues of ρ that lie on the segment $1\leqslant\rho'<\infty$ at an infinitesimal distance above the cut. Although only $\ln(1-\rho)/\rho$ will be treated explicitly, the same can be done for the other two types of functions, $((1-\rho)\ln(1-\rho)+\rho)/\rho^2$ and $((1-\rho)^2\ln(1-\rho)+\rho-\frac{3}{2}\rho^2)/\rho^3$, which appear in Eq. (2.5). The Cauchy integral formula is applied on an eigenvalue by eigenvalue basis. If ρ'' is one of the bound state eigenvalues other than $\rho'' = 1$, then

$$\frac{\ln(1-\rho'')}{\rho''} = \frac{1}{2\pi i}\oint_C d\rho'\,\frac{\ln(1-\rho')}{\rho'}\,\frac{1}{(\rho'-\rho'')}. \tag{3.1}$$

The integration contour, C, has been chosen to consist of a union of counterclockwise loops encircling all bound state eigenvalues of ρ, except for the eigenvalue $\rho' = 1$. With this way of choosing C, Eq. (3.1) remains valid for the same C when other bound state eigenvalues not equal to 1 are substituted for ρ''. Then when Eq. (3.1) is multiplied on both sides

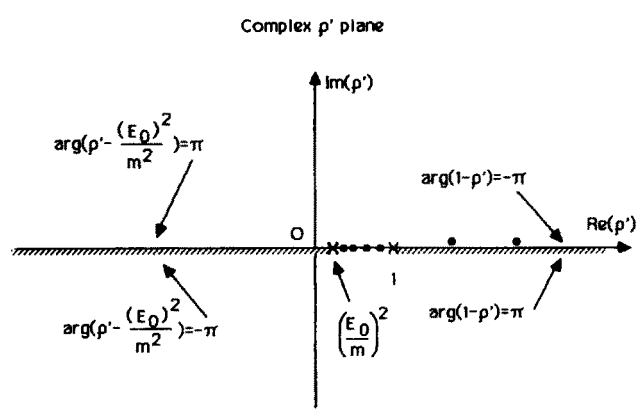


FIG. 1. Complex ρ' plane where ρ', corresponding to an eigenvalue of the operator $\rho \equiv -\Pi\cdot(1+i\sigma)\cdot\Pi/m^2$, is the dimensionless virtual mass squared of the electron. The branch cut $1\leqslant\rho'<\infty$ belongs to the function $\ln(1-\rho')/\rho'$. The other singular points are the singularities of the Coulomb Green's function $1/(\rho'-\rho)$ regarded as a function of ρ' while holding the energy fixed and equal to the energy E_0 of the unperturbed state.

by the projection operator[1] $\epsilon_{\rho'}|\Phi_{\rho'}\rangle\langle\bar{\Phi}_{\rho'}|$ and summed on bound states the summation is independent of C and can be taken under the integral sign. For an operator ρ with a discrete eigenvalue spectrum only, this sum forms on the left-hand side of Eq. (3.1) the spectral representation of $\ln(1-\rho)/\rho$. At the same time, the spectral representation of the Coulomb Green's function, $1/(\rho'-\rho)$, is built up under the integral sign on the right-hand side of Eq. (3.1). Of course, the operators lack the term in the spectral representation corresponding to the eigenvalue $\rho'=1$, a fact that is signaled by a prime notation. The result is the identity

$$\left[\frac{\ln(1-\rho)}{\rho}\right]' = \frac{1}{2\pi i}\oint_C d\rho' \frac{\ln(1-\rho')}{\rho'} \times\left[\frac{1}{\rho'-\rho} - \epsilon_0 \frac{|\Phi_0\rangle\langle\bar{\Phi}_0|}{\rho'-1}\right], \quad (3.2)$$

in which again the subscript 0 is used to refer to the state being perturbed. In view of the factors $(1-\rho)$ in Eq. (2.5), only the primed form of the operators contribute in Eq. (2.5). Since the operator ρ in the Kepler case does not have a bound state spectrum only, Eq. (3.2) requires a more careful proof. The result of a more careful study is that Eq. (3.2) holds also in the Coulomb case provided that to the contour C is added a loop encircling the continuous spectrum in the counterclockwise sense. But the combined effect of encircling all bound state eigenvalues counterclockwise and encircling the continuous spectrum counterclockwise is to encircle the branch cut $1\leqslant\rho'<\infty$ associated with $\rho'=1$ in the clockwise sense. This assumes sufficiently rapid vanishing of the function $\ln(1-\rho')/\{\rho'(\rho'-\rho'')\}$ at $\rho'=\infty$. Next, the integral around the cut attached to $\rho'=1$ is reduced to a simple real integral involving the discontinuity across the cut. This leads to the integral representation

$$\left[\frac{\ln(1-\rho)}{\rho}\right]' = -\int_1^\infty \frac{d\rho'}{\rho'}\left[\frac{1}{\rho'-\rho} - \epsilon_0\frac{|\Phi_0\rangle\langle\bar{\Phi}_0|}{\rho'-1}\right]. \quad (3.3)$$

Analogous simple formal representations can be written for the other functions $((1-\rho)\ln(1-\rho)+\rho)/\rho^2$ and $((1-\rho)^2\ln(1-\rho)+\rho-\frac{1}{2}\rho^2)/\rho^3$, which appear in Eq. (2.5):

$$\left[\frac{(1-\rho)\ln(1-\rho)+\rho}{\rho^2}\right]' = -\int_1^\infty d\rho' \frac{(1-\rho')}{(\rho')^2}\left[\frac{1}{\rho'-\rho} - \epsilon_0\frac{|\Phi_0\rangle\langle\bar{\Phi}_0|}{\rho'-1}\right], \quad (3.4)$$

$$\left[\frac{(1-\rho)^2\ln(1-\rho)+\rho-\frac{1}{2}\rho^2}{\rho^3}\right]' = -\int_1^\infty d\rho' \frac{(1-\rho')^2}{(\rho')^3}\left[\frac{1}{\rho'-\rho} - \epsilon_0\frac{|\Phi_0\rangle\langle\bar{\Phi}_0|}{\rho'-1}\right]. \quad (3.5)$$

These complicated operator structures are thereby expressed in terms of the relatively simple relativistic Coulomb Green's function, $1/(\rho'-\rho)$. There is an explicit subtraction in Eqs. (3.3)–(3.5) for the bound state with $\rho'=1$ at the branch point, and this subtraction renders the integrals convergent at the lower limit $\rho'=1$. For other bound states occurring in the integration region $1\leqslant\rho'<\infty$, the integral is interpreted in accordance with the requirement that the pole shall lie infinitesimally above the cut, as discussed earlier.

B. ζ representation

Equation (2.5) for Σ_0 is quite general and the integral representations (3.3)–(3.5) are expected to be applicable to other external potential problems in addition to the Coulomb problem. This section is specialized to the specific case of the Coulomb problem and a change of variables is made in order to eventually exhibit the $Z\alpha$ dependence of the integrands (3.3)–(3.5).

The change of variables needed to exhibit the $Z\alpha$ dependence of the integrands maps the ρ' plane onto the ζ plane, where

$$\zeta \equiv 2\eta_0/(\eta+\eta_0), \quad (3.6)$$

$$m^2\rho' \equiv (E_0)^2+\eta^2. \quad (3.7)$$

In these equations η_0 corresponds to the classical momentum of the electron in the bound state with energy E_0: $\eta_0 = mq_0$, where q_0 is the parameter (1.7). Solving for ρ' as a function of ζ, one first finds

$$\rho' = [4(q_0)^2 - 4(q_0)^2\zeta+\zeta^2]/\zeta^2, \quad (3.8)$$

and then

$$\left[\frac{\ln(1-\rho)}{\rho}\right]' = -\int_0^1 \frac{d\zeta(2-\zeta)}{\zeta}\frac{4(q_0)^2}{4(q_0)^2-4(q_0)^2\zeta+\zeta^2} \times\left[\frac{1}{\rho'-\rho} - \epsilon_0\frac{|\Phi_0\rangle\langle\bar{\Phi}_0|}{\rho'-1}\right]. \quad (3.9)$$

The ζ forms for the other operator structures appearing in Eq. (2.5) are

$$\left[\frac{(1-\rho)\ln(1-\rho)+\rho}{\rho^2}\right]' = \int_0^1 d\zeta \frac{(2-\zeta)(1-\zeta)}{\zeta}\left[\frac{4(q_0)^2}{4(q_0)^2-4(q_0)^2\zeta+\zeta^2}\right]^2 \times\left[\frac{1}{\rho'-\rho} - \epsilon_0\frac{|\Phi_0\rangle\langle\bar{\Phi}_0|}{\rho'-1}\right], \quad (3.10)$$

and

$$\left[\frac{(1-\rho)^2\ln(1-\rho)+\rho-\frac{1}{2}\rho^2}{\rho^3}\right]' = -\int_0^1 d\zeta \frac{(2-\zeta)(1-\zeta)^2}{\zeta} \times\left[\frac{4(q_0)^2}{4(q_0)^2-4(q_0)^2\zeta+\zeta^2}\right]^3 \times\left[\frac{1}{\rho'-\rho} - \epsilon_0\frac{|\Phi_0\rangle\langle\bar{\Phi}_0|}{\rho'-1}\right]. \quad (3.11)$$

The entire ρ' plane maps into the interior of the unit circle $|\zeta-1|=1$. This is shown in Fig. 2. The cut $1\leqslant\rho'<\infty$ maps into the cut $0<\zeta\leqslant1$ in Fig. 2 with branch point $\zeta=1$. Any bound state eigenvalues of ρ having ζ values in the integration region $0<\zeta<1$ are here to be placed at an infinitesimal distance *below* the cut. The cut $-\infty<\rho'\leqslant(E_0/m)^2$ in Fig. 1 for the continuous spectrum maps onto the boundary of the circle $|\zeta-1|=1$, with points above the cut going to the lower semicircle and points below the cut $-\infty<\rho'\leqslant(E_0/$

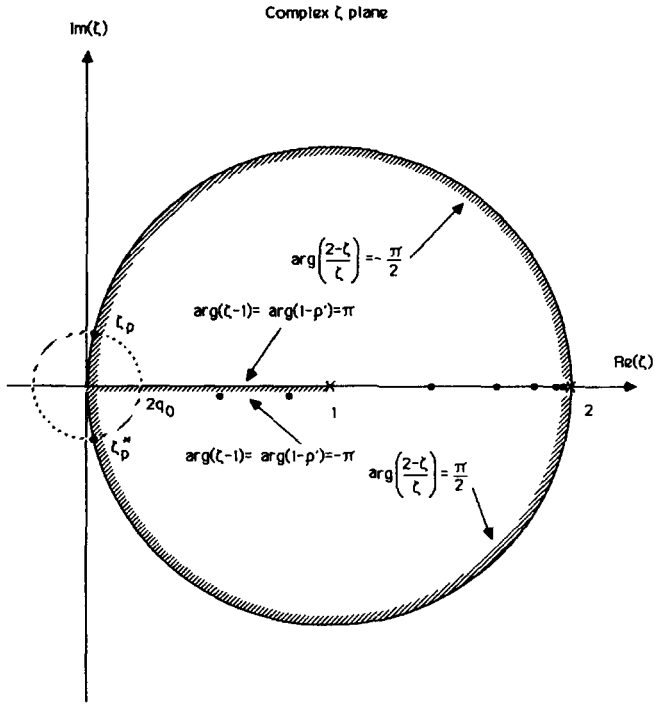


FIG. 2. This is Fig. 1 mapped into the complex ζ plane, where $\zeta \equiv 2\eta_0/(\eta+\eta_0)$. The entire ρ' plane maps into the interior of the unit circle $|\zeta - 1| = 1$. See the discussion following Eq. (3.11). The parameter $i\eta \equiv ((E_0)^2 - m^2\rho')^{1/2}$ is the wave number of the electron; $i\eta_0 = ((E_0)^2 - m^2)^{1/2}$ is the wave number in the unperturbed state of energy E_0.

$m)^2$ going to the upper semicircle. The points ζ_p and ζ_p^* in Fig. 2 are the two poles of the denominators $(4(q_0)^2 - 4(q_0)^2\zeta + \zeta^2)$ which appear in the integral representations (3.9)–(3.11),

$$\zeta_p \equiv 2q_0(q_0 + i(1-(q_0)^2)^{1/2}), \tag{3.12}$$

$$\zeta_p^* \equiv 2q_0(q_0 + i(1-(q_0)^2)^{1/2}). \tag{3.13}$$

It is an easy calculation to verify that the poles (3.12) and (3.13) lie at the intersection of the two circles $|\zeta| = 2q_0$ and $|\zeta - 1| = 1$.

IV. COULOMB STURMIAN BASIS

A. Survey of earlier work

A discrete expansion of the nonrelativistic Coulomb Green's function has been known for some time. The momentum space form of such an expansion was obtained by Schwinger,[22] who exploited the $O(4)$ invariance of the nonrelativistic Kepler problem. The coordinate space analog of Schwinger's nonrelativistic result has been discussed using a coupling constant eigenfunction concept.[23] Subsequently,[11] the same coupling constant eigenfunction concept was found to lead to a discrete expansion also for the Coulomb Green's function of the second-order Dirac equation, Eq. (1.1). Parallel results, referred to as "Sturmian" expansions, have been obtained[24–28] for the Coulomb Green's function of the conventional linear Dirac equation, to which Eq. (1.1) is equivalent.

Derivations of the following equations can be motivated by the coupling constant eigenfunction concept mentioned above, but for present purposes a purely mathematical treatment seems best. Let $E>0$ and $\eta>0$ be two arbitrary real parameters, for the moment assumed to be quite independent. The operator

$$O \equiv \left[\frac{r}{2E}\right]^{1/2}\left[-\frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r} + \eta^2 - \frac{1-(2\gamma+1)^2}{4r^2}\right]\left[\frac{r}{2E}\right]^{1/2} \tag{4.1}$$

is a self-adjoint operator on the Hilbert space of spinor functions $\Phi(\hat{r})$. In Eq. (4.1) γ can be viewed as a self-adjoint operator operating only on spinor and angular degrees of freedom and having the eigenvalues

$$\gamma = ((J+\tfrac{1}{2})^2 - (Z\alpha)^2)^{1/2}, \quad \text{if } L = J + \tfrac{1}{2},$$

and

$$\gamma = ((J=\tfrac{1}{2})^2 - (Z\alpha)^2)^{1/2} - 1, \quad \text{if } L = J - \tfrac{1}{2}.$$

The corresponding eigenfunctions are proportional to the spinor spherical harmonics, $Y_{LJM}(\hat{r})$. The eigenfunctions of O are found to form a discrete set. These eigenfunctions are denoted by $|\eta,A\rangle$ where the shorthand notation A is used to signify the complete set of quantum numbers $\{n,L,J,M\}$, $n = 1,2,3,\ldots$. The explicit expressions for the eigenfunctions are independent of E:

$$\langle \hat{r}|\eta,A\rangle = R_{\eta,n\gamma}(r)\,Y_{LJM}(\hat{r}), \tag{4.2}$$

$$R_{\eta,n\gamma}(r) = (2\eta)^{3/2}\left[\frac{(n-1)!}{(n+2\gamma)!}\right]^{1/2} \times (2\eta r)^{\gamma-1/2}e^{-\eta r}L_{n-1}^{2\gamma+1}(2\eta r). \tag{4.3}$$

The corresponding eigenvalues of O are $O' = (\gamma+n)\eta/E$. Physically, the eigenvalues are the requisite values of the coupling constant $Z\alpha$ needed to produce a Coulomb bound state of type γ,n when the parameters E,η have preassigned values.

The orthogonality and completeness relations for the eigenfunctions are

$$\langle \eta,B|\eta,A\rangle = \int d^3r \langle \eta,B|\hat{r}\rangle\langle \hat{r}|\eta,A\rangle = \delta_{A,B}, \tag{4.4}$$

and

$$\Sigma_A |\eta,A\rangle\langle\eta,A| = 1. \tag{4.5}$$

In this paper the Sturmian basis set (4.3) is used sometimes with one, sometimes with another, value of η. On the other hand, the parameter E always has the same value

$$E = E_0 = \frac{m}{[1+(Z\alpha/(\gamma_0+n_0))^2]^{1/2}}, \tag{4.6}$$

where E_0, γ_0, and n_0 refer to the level whose Lamb shift is to be sought.

An important result for applications is the overlap integral

$$\langle \eta_0,j\gamma|\,|\eta,n\gamma\rangle \equiv \int_0^\infty r^2\,dr\,R^*_{\eta_0,j\gamma}(r)R_{\eta,n\gamma}(r)$$

$$= (-1)^{n-1}\left[\frac{(n_>+2\gamma)!}{(n_>-1)!}\frac{(n_<+2\gamma)!}{(n_<-1)!}\right]^{1/2} \times \frac{(2\zeta-\zeta^2)^{\gamma+1}}{(2\gamma+1)!}(1-\zeta)^{n_>-n_<}$$

$$\times {}_2F_1(-(n_<-1),2\gamma+1+n_>,2\gamma+2;2\zeta-\zeta^2),$$
$$n_>=\max(j,n),\quad n_<=\min(j,n),$$
$$\zeta=2\eta_0/(\eta+\eta_0), \tag{4.7}$$

between eigenstates belonging to two different basis sets associated with different values of η. The hypergeometric function in Eq. (4.7) is a polynomial which could be expressed in terms of the Jacobi polynomial.

B. Sturmian representation of the relativistic Coulomb Green's function

For notational convenience a square bracket, [ ], notation shall signify matrix representatives with respect to the Sturmian basis (4.3). For example $[|\Phi\rangle]$ shall represent the infinite column matrix with matrix elements $[|\Phi\rangle]_A\equiv\langle\eta_0,A|\Phi\rangle$, and $[\rho]$ shall represent the infinite square matrix whose matrix elements are $[\rho]_{AB}\equiv\langle\eta_0,A|\rho|\eta_0,B\rangle$. The next goal will be to obtain the Coulomb Green's function $1/(\rho'-\rho)$ relative to the Sturmian basis. This calculation begins as follows:

$$\frac{1}{\rho'-\rho}=\frac{1}{\rho'-(-\Pi\cdot(1+i\sigma)\cdot\Pi/m^2)}=\frac{m^2}{D},$$
$$D=(m^2\rho'+\Pi\cdot(1+i\sigma)\cdot\Pi)$$
$$=m^2\rho'-(E_0)^2-\nabla^2-\frac{2E_0Z\alpha}{r}$$
$$-\frac{(Z\alpha)^2+iZ\alpha\vec{\sigma}\cdot\hat{r}}{r^2}$$
$$=\eta^2-\nabla^2-\frac{2E_0Z\alpha}{r}-\frac{(Z\alpha)^2+iZ\alpha\vec{\sigma}\cdot\hat{r}}{r^2},$$
$$\eta^2\equiv m^2\rho'-(E_0)^2.$$

At this point one can just follow Ref. 11 and find

$$D=S\left(\frac{2E_0}{r}\right)^{1/2}(O-Z\alpha)\left(\frac{2E_0}{r}\right)^{1/2}S^{-1},$$
$$\frac{1}{\rho'-\rho}=m^2S\left(\frac{r}{2E_0}\right)^{1/2}\frac{1}{(O-Z\alpha)}\left(\frac{r}{2E_0}\right)^{1/2}S^{-1},$$

where O is the operator (4.1). Inserting a complete set of eigenfunctions of O leads to

$$\frac{1}{\rho'-\rho}=m^2S\left(\frac{r}{2E_0}\right)^{1/2}$$
$$\times\Sigma_A\frac{|\eta,A\rangle\langle\eta,A|}{(\gamma+n)(\eta/E_0)-Z\alpha}\left(\frac{r}{2E_0}\right)^{1/2}S^{-1}. \tag{4.8}$$

This is the result of Ref. 11 except that here η has the value $\eta\equiv(m^2\rho'-(E_0)^2)^{1/2}$, as appropriate for use in Sec. III. In Eq. (4.8) $S=\cosh(\theta/2)+i\vec{\sigma}\cdot\hat{r}\sinh(\theta/2)$, $\theta=\tanh^{-1}(Z\alpha/(\vec{\sigma}\cdot\vec{L}+1))$, is the analog of an operator introduced by Biedenharn,[29] and by Martin and Glauber[30] to simplify the Kepler problem of the linear Dirac equation. A final transformation involves taking matrix elements of Eq. (4.8) with respect to the Sturmian basis (4.3) and going over to a description in terms of the parameter $\zeta=2\eta_0/(\eta+\eta_0)$. For this the relations

$$mq_0/E_0=Z\alpha/(\gamma_0+n_0),$$
$$\eta=mq_0(2-\zeta)/\zeta \tag{4.9}$$

are needed. The result of these changes is

$$\frac{1}{\rho'-[\rho]}=\frac{1}{4(q_0)^2}[S][r']^{1/2}$$
$$\times\Sigma_A\frac{\zeta[|\eta,A\rangle][\langle\eta,A|]}{2(\gamma+n)-\zeta(\gamma+n+\gamma_0+n_0)}$$
$$\times[r']^{1/2}[S]^{-1}, \tag{4.10}$$

in which $\vec{r}\equiv 2mq_0\vec{r}$ is a dimensionless coordinate vector. Relatively simple expressions for the matrix elements $[|\eta,A\rangle]$ and $[\langle\eta,A|]$ are provided by the overlap integrals (4.7). The matrix elements $[|\eta,A\rangle]$ and $[\langle\eta,A|]$ are seen to depend to $Z\alpha$ only through the parameter γ. If the second term in the factor $\{(1/(\rho'-\rho))-\epsilon_0|\Phi_0\rangle\langle\bar{\Phi}_0|/(\rho'-1)\}$ entering the integrals (3.9)–(3.11) is also transformed to the Sturmian basis the result is the identity[31]

$$\left(\frac{1}{\rho'-[\rho]}-\epsilon_0\frac{[|\Phi_0\rangle][\langle\bar{\Phi}_0|]}{(\rho'-1)}\right)$$
$$=\frac{[S][r']^{1/2}}{4(q_0)^2}\left(\Sigma_A\frac{\zeta[|\eta,A\rangle][\langle\eta,A|]}{2(\gamma+n)-\zeta(\gamma+n+\gamma_0+n_0)}\right.$$
$$\left.-\frac{\zeta^2[|\eta_0,0\rangle][\langle\eta_0,0|]}{2(\gamma_0+n_0)(1-\zeta)}\right)[r']^{1/2}[S]^{-1}, \tag{4.11}$$

in which $[|\eta_0,0\rangle]$ is the Sturmian representation of the unperturbed state, with matrix elements $[|\eta_0,0\rangle]_A=\langle\eta_0,A|\eta_0,0\rangle\equiv\delta_{A,0}$. Next Eq. (4.11) is substituted in the integral representations (3.9)–(3.11) to convert them over to the Sturmian basis. The result is

$$\left[\frac{\ln(1-[\rho])}{[\rho]}\right]'$$
$$=-[S][r']^{1/2}\int_0^1\frac{d\zeta(2-\zeta)}{\zeta}$$
$$\times\frac{4(q_0)^2}{(4(q_0)^2-4(q_0)^2\zeta+\zeta^2)}\frac{1}{4(q_0)^2}$$
$$\times\left(\Sigma_A\frac{\zeta[|\eta,A\rangle][\langle\eta,A|]}{2(\gamma+n)-\zeta(\gamma+n+\gamma_0+n_0)}\right.$$
$$\left.-\frac{\zeta^2[|\eta_0,0\rangle][\langle\eta_0,0|]}{2(\gamma_0+n_0)(1-\zeta)}\right)[r']^{1/2}[S]^{-1}, \tag{4.12}$$

with similar results for the other two integrals (3.10) and (3.11).

Aside from the factors $[S][r']^{1/2}$ and $[S]^{-1}[r']^{1/2}$ the nuclear charge is seen to appear in the integral representation (4.12) only through the two parameters q_0 and γ. The term involving Σ_A depends on $Z\alpha$ only through γ, a parameter which differs from the constant L by terms of order $(Z\alpha)^2$. The main $Z\alpha$ dependence of the integrand in Eq. (4.12) enters through the parameter q_0. But this q_0 dependence is exhibited explicitly in Eq. (4.12) and consists entirely of the factors $4(q_0)^2/(4(q_0)^2-4(q_0)^2\zeta+\zeta^2)$ and $(2q_0)^{-2}$. Similar statements hold with regard to the other two integral representations (3.10) and (3.11) in the Sturmian basis. By looking at the singularity structure of the integrands as a function of the complex variable q_0 holding γ fixed, the final integrals (3.9)–(3.11) for each partial wave in the Sturmian basis can be shown to define functions that

are analytic in the complex q_0 plane except for a branch cut along the negative axis $-\infty < q_0 \leqslant 0$.

The integral representation (4.12) and analogous representations corresponding to Eqs. (3.10) and (3.11) seem to be a convenient point of departure for a future numerical evaluation of the Lamb shift. In addition, their relatively simple forms invite further analytical study.

ACKNOWLEDGMENTS

This work was carried out over a period of time including my sabbatical leave in the fall of 1984 at which time I was a visiting professor at Yale University and including the summer of 1985 when I was a visiting scientist at Cornell University.

I am grateful to the physics departments at Yale University and Cornell University for their hospitality.

This research was supported in part by the National Science Foundation under Grant No. PHY-8415543.

APPENDIX: STURMIAN REPRESENTATION OF THE MASS EIGENFUNCTIONS

The eigenvalue problem $\Lambda\Phi = \Lambda'\Phi$, $\Lambda \equiv -\Pi\cdot(1 + i\sigma)\cdot\Pi$ [see the discussion in Sec. I following Eq. (1.3)] has been investigated in detail in Ref. 1 where the eigenvalues and eigenfunctions have been calculated. For the calculation of the Lamb shift described in Sec. II it seems convenient to refer all to a Sturmian basis. The Sturmian representations of the mass eigenfunctions are as follows.

Continuum states: Let the eigenfunctions be written in the form

$$\langle \vec{r}|\Phi_{pLJM}\rangle = S\chi_{p\gamma}(r)\,Y_{LJM}(\hat{r}),$$

where

$$\chi_{p\gamma}(r) = (dp/2\pi)^{1/2}\Gamma(1+\gamma-i\nu)e^{\pi\nu/2-1} \times \mathcal{M}_{i\nu;\gamma+1/2}(-2ipr), \quad \nu \equiv E_0 Z\alpha/p, \quad 0<p<\infty. \tag{A1}$$

The eigenvalue of Λ belonging to the eigenfunction $|\Phi_{pLJM}\rangle$ is $\Lambda' = (E_0)^2 - p^2 < (E_0)^2$. The operator Λ is constructed assuming the physical coupling constant $Z\alpha$ and assuming that the energy has the value E_0 given by Eq. (4.6). In the following a double bar notation is used to signify the radial part only of the overlap integrals. If the functions $\chi(r)\,Y_{LJM}(\hat{r})$ are multiplied by $r^{-1/2}$, then relatively simple expressions are obtained for their matrix elements with respect to the Sturmian basis, $|\eta_0,A\rangle$. For the continuum states the result is

$$\langle \eta_0,n\gamma|\left|\frac{1}{r^{1/2}}\right||\chi_{p\gamma}(r)\rangle \equiv \int_0^\infty r^2\,dr\,R^*_{\eta_0,n\gamma}(r)\frac{1}{r^{1/2}}\chi_{p\gamma}(r)$$
$$= \left(\frac{dp}{2\pi}\right)^{1/2}\frac{(-1)^{n-1}e^{-i\pi(1+\gamma+i\nu)/2}}{[(n-1)!(n+2\gamma)!]^{1/2}}\left(\frac{4\eta_0 p}{(\eta_0)^2+p^2}\right)^{\gamma+1}\left(\frac{\eta_0-ip}{\eta_0+ip}\right)^{n-1-i\nu}$$
$$\times\Gamma(n+\gamma-i\nu)\,{}_2F_1\left(-(n-1),\gamma+1+i\nu,1-n-\gamma+i\nu;\left(\frac{\eta_0+ip}{\eta_0-ip}\right)^2\right),$$
$$\nu = E_0 Z\alpha/p, \quad |\arg(\eta_0 \pm ip)| < \pi/2. \tag{A2}$$

Bound states: The bound state eigenfunctions $|\Phi\rangle$ calculated in Ref. 1 have the form

$$\langle \vec{r}|\Phi\rangle = S\chi_{n\gamma}(r)\,Y_{LJM}(\hat{r}),$$

where

$$\chi_{n\gamma}(r) = (\eta r/(n+\gamma))^{1/2}R_{\eta,n\gamma}(r), \tag{A3}$$

and $R_{\eta,n\gamma}(r)$ is the Sturmian function (4.3), with

$$\eta = E_0 Z\alpha/(n+\gamma). \tag{A4}$$

The matrix representative of $r^{-1/2}\chi_{n\gamma}(r)$ with respect to the Sturmian basis $|\eta_0,A\rangle$ is

$$\langle \eta_0,j\gamma|\left|\frac{1}{r^{1/2}}\right||\chi_{n\gamma}(r)\rangle \equiv \int_0^\infty r^2\,dr\,R^*_{\eta_0,j\gamma}(r)\frac{1}{r^{1/2}}\chi_{n\gamma}(r)$$
$$= \left[\frac{\eta}{(\gamma+n)}\right]^{1/2}\langle \eta_0,j\gamma|\,|\eta,n\gamma\rangle, \tag{A5}$$

where $\langle \eta_0,j\gamma|\,|\eta,n\gamma\rangle$ denotes the overlap integral (4.7).

[1] L. C. Hostler, J. Math. Phys. **26**, 124 (1985).
[2] O. Laporte and G. E. Uhlenbeck, Phys. Rev. **37**, 1380 (1931).
[3] R. P. Feynman and M. Gell-Mann, Phys. Rev. **109**, 193 (1958).
[4] L. M. Brown, Phys. Rev. **111**, 957 (1958).
[5] M. Tonin, Nuovo Cimento **14**, 1108 (1959).
[6] W. R. Theis, Fortschr. Phys. **7**, 559 (1959).
[7] H. Pietschmann, Acta Phys. Austriaca **14**, 63 (1961).
[8] L. M. Brown, "Two-component fermion theory," in *Lectures in Theoretical Physics* (Interscience, New York, 1962), Vol. IV.
[9] L. M. Brown, "Quantum electrodynamics at high energy," in *Topics in Theoretical Physics*, Proceedings of the Liperi Summer School in Theoretical Physics 1967, edited by C. Cronstrom (Gordon and Breach, New York, 1969), p. 113.
[10] L. C. Hostler, J. Math. Phys. **23**, 1179 (1982).
[11] L. C. Hostler, J. Math. Phys. **24**, 2366 (1983).
[12] L. C. Hostler, J. Math. Phys. **26**, 1348 (1985); **27**, 2208 (E) (1986).
[13] L. C. Hostler, J. Math. Phys. **27**, 2423 (1986).
[14] Natural units, defined by $\hbar = c = 1$, are used. Four-vectors have an imaginary time component, e.g., $x_\mu = (\mathbf{r},it)$. Accordingly, the Lorentz metric is the simple $\delta_{\mu\nu}$. Derivative operators acting to the left signify *minus* differentiation of the objects on the left, e.g., $\overline{\Phi}\overleftarrow{\partial}_\mu \equiv -\partial\overline{\Phi}/\partial x_\mu$.
[15] J. Schwinger, Proc. Nat. Acad. Sci. USA **37**, 455 (1951).
[16] M. Lieber, Phys. Rev. **174**, 2037 (1968).
[17] P. J. Mohr, Ann. Phys. (NY) **88**, 26 (1974); Phys. Rev. A **26**, 2338

(1982); J. Sapirstein, Phys. Rev. Lett. **47**, 1723 (1981); **51**, 985 (1983).
[18]G. W. Erickson and D. R. Yennie, Ann. Phys. (NY) **35**, 271, 447 (1965).
[19]F. Rohlich, Phys. Rev. **80**, 666 (1950).
[20]A. Salam, Phys. Rev. **86**, 731 (1952).
[21]See, for example, Ref. 18.
[22]J. Schwinger, J. Math. Phys. **5**, 1606 (1964).
[23]L. Hostler, J. Math. Phys. **11**, 2966 (1970).
[24]N. L. Manakov, L. P. Rapoport, and S. A. Zapryagaev, Phys. Lett. A **43**, 139 (1973).
[25]N. L. Manakov, L. P. Rapoport, and S. A. Zapryagaev, J. Phys. B **7**, 1076 (1974).
[26]Ya. I. Granovskii and V. I. Nechet, Teor. Mat. Fiz. **18**, 262 (1974).
[27]A. I. Mil'shtein and V. M. Strakhovenko, Phys. Lett. A **90**, 447 (1982).
[28]In addition to Refs. 24–27 on the relativistic Coulomb Green's function, see A. Maquet, Phys. Rev. A **15**, 1088 (1977), where nonrelativistic "Sturmian" expansions are applied extensively to atomic calculations.
[29]L. C. Biedenharn, Phys. Rev. **126**, 845 (1962).
[30]P. C. Martin and R. J. Glauber, Phys. Rev. **109**, 1307 (1958).
[31]The projection operator on the left-hand side of Eq. (4.11) is expressed in "standard form"; on the right-hand side of Eq. (4.11) it is expressed in the "short form" (see Ref. 1).

Rational functions of momentum as invariants for one-dimensional, time-dependent potentials: Basic theory

João Goedert[a] and H. Ralph Lewis
Los Alamos National Laboratory, MS-F642, Los Alamos, New Mexico 87545

(Received 19 November 1985; accepted for publication 29 October 1986)

A framework for the momentum-resonance formulation of Lewis and Leach [Ann. Phys. (NY) **164**, 47 (1985)] is presented that casts new light into the nature of exact, explicitly time-dependent invariants for one-dimensional, time-dependent potentials and produces additional examples of such invariants. The momentum-resonance formulation postulates that the invariant be a rational function of momentum with simple poles, which are called momentum resonances. It is shown that an invariant of resonance type can be written as a functional of the potential in terms of the solution of a system of linear algebraic equations; and a single necessary and sufficient condition for a potential to admit an invariant of resonance type is obtained. These results are obtained by reformulating the problem in terms of a set of discrete moments that satisfy two separate recursion formulas. Invariants for new time-dependent potentials can be obtained and previously known invariants are recovered.

I. INTRODUCTION

Exact invariants for Hamiltonian systems can be very useful for obtaining insight into the nature of solutions of the equations of motion and they can be helpful in computing solutions numerically. The search for invariants for specific systems has a long history. As an example, we mention the gravitational three-body problem, which is relevant to the motion of celestial bodies and has been studied extensively.[1] In 1887 Bruns showed that the ten so-called classical integrals of the three-body problem are the only invariants that exist that are algebraic functions of the coordinates, momenta, and time. In 1889 Poincaré showed for the restricted three-body problem that the only explicitly time-independent invariant that is periodic in the coordinates is a quantity known as the Jacobian energy. These important results illustrate that an explicitly time-dependent invariant is likely to be outside the class of algebraic functions and that explicitly time-independent invariants are uncommon, even for autonomous systems.

In the search for exact invariants, attention traditionally has been concentrated on functions that do not depend on time explicitly. Also, the systems considered have usually been autonomous; that is, the Hamiltonians usually have not depended on time explicitly. In this article, we consider the motion of a particle in a one-dimensional potential and allow the invariant as well as the potential to be *explicitly time dependent*. For an autonomous one-dimensional Hamiltonian system, all invariants that are functionally independent of the energy are explicitly time dependent. For any particular nonautonomous system, it may be that all global invariants are explicitly time dependent. In any event, any complete set of $2N$ invariants for an N-dimensional system must include at least one explicitly time-dependent invariant.

The equations of motion for any one-dimensional nonautonomous Hamiltonian system are equivalent to those of a corresponding autonomous two-dimensional Hamiltonian system. Therefore, it is always possible to treat one-dimensional nonautonomous systems by considering certain two-dimensional autonomous systems instead. The two-dimensional Hamiltonian depends linearly on a momentum variable that is associated with the energy of the one-dimensional system. In this regard, it should be noted that invariants derived by Darboux,[2] Whittaker,[1] Holt,[3] Hall,[4] and others for particle motion in a two-dimensional, time-independent potential cannot be used to obtain invariants for one-dimensional, time-dependent potentials. The reason is that the Hamiltonian for the motion of a particle in a two-dimensional, time-independent potential depends quadratically on each momentum variable.

An important area in which explicitly time-dependent invariants for time-dependent potentials play a crucial role is the self-consistent theory of collisionless plasma.[5] When there is only one spatial dimension, the governing equations, known as the Vlasov–Poisson equations, describe a continuum of particles that move in the electric field generated by the particles themselves. The electric field is to be determined self-consistently along with the motion of the particles. The phase-space distribution function for the particles, which is a solution of the Vlasov equation, is a function of invariants of the motion of a single particle in the electric field. An exact or approximate invariant can be useful in connection with the Vlasov–Poisson equations if it applies to a *class* of electric fields that approximate the field associated with the exact solution.

Exact invariants for particle motion in classes of one-dimensional explicitly time-dependent potentials have been found for both linear and nonlinear equations of motion. For linear, arbitrarily time-dependent oscillators, invariants are known that are homogeneous quadratic forms in the coordinate and momentum.[6] For nonlinear equations of motion, an exact invariant is known that is quadratic in the momentum when the potential has a certain form that involves an arbitrary function of a time-dependent linear function of the spatial variable.[7,8] This invariant has been used[9] to find new exact solutions of the Vlasov–Poisson equations.

In this article we describe an elaboration of the momentum-resonance ansatz of Lewis and Leach[10] to study exact invariants for time-dependent, one-dimensional potentials.

[a] Permanent address: Instituto de Física, Universidade Federal do Rio Grande do Sul, 90049-Porto Alegre, Rio Grande do Sul, Brazil.

 0022-2488/87/030728-08$02.50

Their ansatz provides a framework for studying invariants admitted by a larger class of time-dependent potentials than was known previously. For a potential that admits an exact invariant of the resonance type, we have shown that the invariant can be constructed as a functional of the potential in terms of the solution of a *linear algebraic* system of equations. In addition to this linearization theorem, we have found a necessary and sufficient condition for the existence of an invariant with a given number of resonances.

There exist more potentials that admit invariants with two resonances than were previously known and we have found examples of such potentials.[11] We have also examined the case of three resonances.[11] We have found examples of potentials that admit three-resonance invariants, but we have not found a generalization of the class of two-resonance cases that involve an arbitrary function of a time-dependent linear function of position.

The remainder of this article is organized as follows. In Sec. II we review the momentum-resonance ansatz. In Sec. III we present a discrete-moment description of invariants of the resonance type and derive the linearization theorem. In Sec. IV we derive a *single* necessary and sufficient condition for a potential to admit an invariant of the resonance type. In Sec. V we present some preliminary applications. We rederive the examples of Lewis, Leach, and Sarlet for one and two resonances and present two examples of potentials that admit invariants with three resonances. Additional examples are presented in Ref. 11. In Sec. VI we present some concluding remarks.

II. THE MOMENTUM-RESONANCE ANSATZ

We consider the Hamiltonian for a particle moving in a potential that depends on a coordinate q and time t,

$$H = \tfrac{1}{2}p^2 + V(q,t). \tag{1}$$

The momentum-resonance ansatz of Lewis and Leach[10] postulates that the invariant be a rational function of momentum expressed as a sum of terms whose singularities in momentum are only distinct simple poles ("momentum resonances"),

$$I(q,p,t) = c(q,t) + \sum_{n=1}^{N} \frac{v_n(q,t)}{p - u_n(q,t)}. \tag{2}$$

In view of the fact that any function of an invariant is also an invariant, the momentum-resonance ansatz is sufficient for considering a much wider class of invariants. For example, it includes the case of invariants that are polynomials in p with distinct zeros. A motivation for considering invariants that are rational functions of momentum is that a large class of functions can be well approximated by rational functions.[12] In a particular application, it may be possible to obtain a useful approximate invariant in terms of a potential that admits an invariant which is rational in the momentum with distinct simple poles.

The functions of position and time that appear in the expression for the invariant in resonance form satisfy conditions that are decoupled to a remarkable degree. The condition that $I(q,p,t)$ be an invariant is

$$\frac{dI}{dt} \equiv \frac{\partial I}{\partial t} + p\frac{\partial I}{\partial q} - \frac{\partial V}{\partial q}\frac{\partial I}{\partial p} = 0, \tag{3}$$

which implies that necessary and sufficient conditions on the functions c, v_n, and u_n such that $I(q,p,t)$ be an invariant are

$$\frac{\partial c}{\partial q} = 0, \tag{4a}$$

$$\frac{\partial c}{\partial t} + \sum_{n=1}^{N} \frac{\partial v_n}{\partial q} = 0, \tag{4b}$$

$$\frac{\partial v_n}{\partial t} + \frac{\partial}{\partial q}(u_n v_n) = 0, \tag{4c}$$

$$\frac{\partial u_n}{\partial t} + u_n\frac{\partial u_n}{\partial q} = -\frac{\partial V}{\partial q}. \tag{4d}$$

Lewis and Leach[10] introduced a set of N time-dependent transformations to Lagrangian coordinates to study these equations. So transformed, (4c) was satisfied identically, (4d) was an equation for the nth transformation function, and (4b) was a condition that related the set of N transformation functions. Lewis and Leach found all potentials that admit a one-resonance invariant ($N = 1$), they derived a class of two-resonance invariants, and they devised an approach for studying the general multiresonance case. The spatial derivative of a potential with a one-resonance invariant is a rational function of q with coefficients expressible in terms of three arbitrary functions of t. Those potentials were found in different contexts by Sarlet[13] and by Leach, Lewis, and Sarlet.[14] The two-resonance invariants found by Lewis and Leach are the reciprocals of invariants quadratic in the momentum. The potentials associated with invariants quadratic in the momentum have been derived earlier by Lewis and Leach[7] and by Sarlet and Ray.[8] They involve an arbitrary function of a time-dependent linear function of q. For a multiresonance case, it is necessary to find N solutions of (4d) for the same $\partial V/\partial q$. Lewis and Leach studied this question in the context of their transformations to Lagrangian coordinates. They were able to relate any two transformation functions that correspond to distinct solutions of (4d). However, they did not succeed in finding additional two-resonance examples or in showing that no further examples exist.

The structure of (4a)–(4d) is remarkable. Different values of n are coupled only through the single condition (4b); (4c) is of the form of the continuity equation for a fluid; and (4d) is of the form of the equation expressing conservation of momentum for a fluid. One would expect that the simplicity of the structure of the equations could be used to extract more information of a general nature about dynamical systems that admit invariants in resonance form. In addition, one might expect that further two-resonance examples exist and that examples of potentials that admit invariants with more than two resonances could be found. In the remainder of this article, we show that invariants in resonance form are related to the solution of a system of linear algebraic equations and we derive a necessary and sufficient condition for a potential to admit an invariant in resonance form. In addition, we derive the examples of Lewis and

Leach with a different formulation and present two examples of three-resonance invariants. We examine the two- and three-resonance cases in more detail in Ref. 11.

III. DISCRETE-MOMENT FORMULATION AND LINEARIZATION THEOREM

The momentum-resonance ansatz can be formulated in terms of certain discrete momentum moments. This formulation is characterized by the following features, which we discuss in this and the following section. Condition (4b) is satisfied identically. Aside from unspecified functions of t, which are "integration constants" associated with integrals with respect to q, the discrete moments can be calculated as functionals of the potential *a priori*, without solving (4c)–(4d). Any invariant that can be written in resonance form can be constructed explicitly from discrete moments. The foregoing features are derived in this section. Finally, in Sec. IV, we derive in terms of the discrete moments a necessary and sufficient condition for a potential to admit an invariant with N poles.

The k th moment $g_k(q,t)$ is defined by

$$g_k(q,t) = \sum_{n=1}^{N} u_n^k v_n. \tag{5}$$

For later use, in connection with (16), we exclude the case in which a function u_n is identically zero. If a function u_n were identically zero, then, from (4d), $\partial V/\partial q$ would be identically zero. Therefore we also assume that $\partial V/\partial q$ is not identically zero. If, for fixed q and t, we consider the quantities $v_n(q,t)$ to be the values of a function $v(q,p,t)$ that is defined at a discrete set of values of p given by $p = u_n(q,t)$ for $1 \leqslant n \leqslant N$, then $g_k(q,t)$ is the k th moment of $v(q,p,t)$ in that discrete space of values of p. By direct manipulation of (4c) and (4d), it can be shown that these moments satisfy the differential recursion relation

$$\frac{\partial g_k}{\partial q} = -\frac{\partial g_{k-1}}{\partial t} - (k-1)g_{k-2}\frac{\partial V}{\partial q}, \quad k \geqslant 1, \tag{6a}$$

with the initial condition

$$g_0(q,t) = -\dot{\alpha}_{-1}(t)q + \alpha_0(t), \tag{6b}$$

where $\alpha_{-1}(t) = c(t)$. Equation (6b) is the solution of (4b). Equation (6a) can be obtained by multiplying (4c) by u_n^{k-1}, using (4d), and summing over all k from 1 to N. Therefore the system of equations (4a)–(4d) implies the recursion relation (6a) and (6b). However, we stress that the converse is not true; the recursion relation (6a) and (6b) alone does not imply (4a)–(4d). Our interpretation of g_k as a moment is natural because (6a) is precisely the recursion relation satisfied by the continuum momentum moments of (3), which defines the invariant $I(q,p,t)$. In the context of collisionless plasma physics, (3) is also the Vlasov equation for a phase-space distribution function. Thus it is appropriate to interpret $v(q,p,t)$ as the representation of an invariant or a phase-space distribution function in a discrete space of momentum values given by $p = u_n(q,t)$ for $1 \leqslant n \leqslant N$.

We shall show that the moments g_k also satisfy an algebraic recursion relation in addition to the differential recursion relation (6a). The algebraic recursion relation is

$$g_l = -\sum_{n=1}^{N} a_n g_{l-n}, \quad l \leqslant N, \tag{7}$$

where the quantities a_k and u_n are related by

$$u_n^N + \sum_{k=1}^{N} a_k u_n^{N-k} = 0, \quad 1 \leqslant n \leqslant N. \tag{8}$$

Relation (7) plays an important role in numerical analysis of N th-order ordinary differential equations with constant coefficients.[15] In that context, the a_k are given and the u_n are determined from (8). When the v_n can be determined from the first N of Eqs. (5) by specifying the first N moments, then the remaining g_k that solve (7) are given by (5). For our purposes, we need instead that the definition (5) of the moments imply the relation (8) and the algebraic recursion relation (7). In order to establish (8), we define the quantities a_k to be the coefficients of the polynomial whose roots are the u_n. Let $D(p)$ be that polynomial in the variable p,

$$D(p) \equiv \prod_{k=1}^{N} (p - u_k). \tag{9}$$

By our definition of the a_k, $D(p)$ can also be written as

$$D(p) = p^N + \sum_{k=1}^{N} a_k p^{N-k}. \tag{10}$$

Thus (8) is simply the statement that each u_n is a root of $D(p)$,

$$D(u_n) = 0. \tag{11}$$

Now it is easy to use (5) and (8) to derive (7):

$$\begin{aligned} g_l = \sum_{k=1}^{N} u_k^N u_k^{l-N} v_k &= -\sum_{k=1}^{N}\sum_{n=1}^{N} a_n u_k^{N-n} u_k^{l-N} v_k \\ &= -\sum_{n=1}^{N} a_n \sum_{k=1}^{N} u_k^{l-n} v_k. \end{aligned} \tag{12}$$

If we take $l \geqslant N$, then (12) is the same as (7).

The fact that the moments satisfy two distinct recursion relations is the basis for our remaining discussion. The algebraic recursion relation can be used to calculate the a_n in terms of the moments g_k directly, without using the u_n explicitly. Define square matrices Λ_K by

$$\Lambda_K = \begin{pmatrix} g_0 & g_1 & \cdots & g_K \\ g_1 & g_2 & \cdots & g_{K+1} \\ \vdots & \vdots & \vdots & \vdots \\ g_K & \cdot & \cdots & g_{2K} \end{pmatrix}, \tag{13}$$

with elements

$$(\Lambda_K)_{ij} = g_{i+j-2}, \quad 1 \leqslant i,j \leqslant K+1, \tag{13'}$$

and column matrices x_K and y_K by

$$x_K = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{pmatrix}, \quad y_K = \begin{pmatrix} g_K \\ g_{K+1} \\ \vdots \\ g_{2K-1} \end{pmatrix}, \tag{14}$$

with elements

$$(x_K)_j = a_j, \quad (y_K)_j = g_{K+j-1}, \quad 1 \leqslant j \leqslant K. \tag{14'}$$

Matrices of the form (13) are known as Hankel matri-

ces.[16–18] The first N of the recursion equations (7) are

$$\Lambda_{N-1}x_N = -y_N. \tag{15}$$

They can be considered as a system of N *linear algebraic* equations for the a_n. The $2N$ moments that occur in the definition of Λ_{N-1} and y_N can be calculated *a priori* from (6a) and (6b) in terms of the potential and $2N+2$ unspecified functions of t ("integration constants"). The system of equations (15) can be constructed explicitly in this way, without knowing the functions of u_n and v_n. The necessary and sufficient condition for (15) to have a solution is

$$\det \Lambda_{N-1} \neq 0. \tag{16}$$

If (16) did not hold, then the columns of Λ_{N-1} would be linearly dependent, which would imply that there existed an algebraic recursion relation of the form (7) with N replaced by $N-1$. However, if there existed a relation of the form (7) with N replaced by $N-1$, then, to be consistent with (7) and (5), one u_n would have to be identically zero, contrary to assumption.

We now demonstrate that the invariant (2) can be written explicitly in terms of g_0 through g_{N-1} and the coefficients a_n by the formula

$$I(q,p,t) = c(t) + \frac{\sum_{n=1}^{N} p^{N-n} \sum_{k=1}^{n} a_{k-1} g_{n-k}}{p^N + \sum_{n=1}^{N} a_n p^{N-n}}, \tag{17}$$

where we have defined

$$a_0 \equiv 1 \tag{18}$$

and used (4a), which states that c depends on t only. The expressions for the invariant given by (2) and (17) are identically equal because of the algebraic recursion relation (7). We first show that $D(p)$, defined by (9), can be factored according to the formula

$$D(p) = (p-u_n) \sum_{k=1}^{N} p^{N-k} \sum_{s=1}^{k} a_{k-s} u_n^{s-1} \tag{19}$$

for any n satisfying $1 \leqslant n \leqslant N$. By carrying out the multiplication by the first factor in (19) we have

$$D(p) = \sum_{k=1}^{N} p^{N-k+1} \sum_{s=1}^{k} a_{k-s} u_n^{s-1} - \sum_{k=1}^{N} p^{N-k} \sum_{s=1}^{k} a_{k-s} u_n^{s}. \tag{20}$$

We now change the summation indices k to $\kappa = k-1$ and s to $\sigma = s-1$ in the first summation in (20) and rewrite (20) as

$$\begin{aligned} D(p) &= \sum_{\kappa=0}^{N} a_\kappa p^{N-\kappa} - \sum_{\sigma=0}^{N} a_{N-\sigma} u_n^{\sigma} \\ &= \sum_{k=0}^{N} a_k p^{N-k} - \sum_{k=0}^{N} a_k u_n^{N-k} \\ &= \sum_{k=0}^{N} a_k p^{N-k} - D(u_n), \end{aligned} \tag{21}$$

which is the definition (10) of $D(p)$ because $D(u_n)$ is zero by assumption. Thus (19) is proved. In order to complete the proof of the equivalence of (2) and (17), we rewrite (2) as

$$I(q,p,t) = c(t) + \frac{1}{D(p)} \sum_{n=1}^{N} \frac{D(p) v_n}{p-u_n}. \tag{22}$$

Now substitute (19) for $D(p)$ in the numerator of (22), thus canceling the factor $(p-u_n)$. The resulting expression is

$$I(q,p,t) = c(t) + \frac{1}{D(p)} \sum_{n=1}^{N} v_n \sum_{k=1}^{N} p^{N-k} \sum_{s=1}^{k} a_{k-s} u_n^{s-1}. \tag{23}$$

Performing the summation over n first and using the definition (5), we obtain

$$I(q,p,t) = c(t) + \frac{1}{D(p)} \sum_{k=1}^{N} p^{N-k} \sum_{s=1}^{k} a_{k-s} g_{s-1}. \tag{24}$$

If we now change the summation index s to $\sigma = k-s+1$ we obtain (17), which proves our assertion of the equivalence of (2) and (17). We can now state the following theorem, which provides a means of constructing an N-resonance invariant for a given potential if one exists.

Linearization Theorem: If $V(q,t)$ $(\partial V/\partial q \not\equiv 0)$ admits an invariant with N resonances, then this invariant can be expressed by (17) in terms of moments g_k and coefficients a_n. The g_k can be calculated from (6a) and (6b) in terms of the potential and $2N+2$ unspecified functions of t ("integration constants"). The a_n are the solution of the system of N linear algebraic equations (15).

For a specified potential, the entire q dependence of the moments can be determined *a priori* from (6a) and (6b). Only the functions of t that arise as the "integration constants" for (6a) are undetermined from (6a) and (6b).

IV. NECESSARY AND SUFFICIENT CONDITION

In this section we complement the linearization theorem by deriving a necessary and sufficient condition for an N-resonance invariant to exist. The condition is stated in the following theorem.

Theorem: An N-resonance invariant exists for the Hamiltonian (1) with potential $V(q,t)$ $(\partial V/\partial q \not\equiv 0)$ if, and only if, there exist moments $g_k(q,t)$, $1 \leqslant k \leqslant 2N$, such that

$$\det \Lambda_N = 0, \tag{25}$$

where Λ_N is a Hankel matrix defined by (13) and where the moments g_k in Λ_N satisfy the differential recursion relation (6a) and (6b). If the invariant exists, then it can be expressed by (17) by choosing the "integration constants" in the moments $g_k(q,t)$ such that the moments satisfy (25). Stated briefly,

$$\frac{dI}{dt} = 0 \quad \Leftrightarrow \quad \det \Lambda_N = 0. \tag{26}$$

We now prove (26) directly. There may exist a more concise proof, perhaps by induction. However, our direct proof suffices and is interesting in its own right.

To begin the proof, we extend the range of the indices of g_k and a_k by defining

$$g_{-k} = a_{-k} = a_{N+k} = 0, \quad k > 0. \tag{27}$$

We also define the set of auxiliary functions

$$A_s = \sum_{n=0}^{s} a_{n-1} g_{s-n}, \tag{28}$$

where the lower limit $n = 0$ has been chosen instead of $n = 1$ for later convenience. We observe that (27) trivially implies

$$A_0 = A_{-k} = 0, \quad k > 0. \tag{29}$$

In addition, (27) and (7) imply

$$A_{N+k} = 0, \quad 1 \leqslant k \leqslant N. \tag{30}$$

Because of (27), A_{2N+1} can be written as

$$A_{2N+1} = g_{2N} + \sum_{k=1}^{N} a_k g_{2N-k}. \tag{31}$$

This equation along with the N equations given by (15) can be considered as a system of $N+1$ equations for A_{2N+1} and the N quantities a_n,

$$\delta_{N,k} A_{2N+1} - \sum_{n=1}^{N} a_n g_{N+k-n} = g_{N+k}, \quad 0 \leqslant k \leqslant N. \tag{32}$$

The solution for A_{2N+1} is

$$A_{2N+1} = (\det \Lambda_N)/(\det \Lambda_{N-1}). \tag{33}$$

In terms of a_k and A_k the invariant (17) is

$$I(q,p,t) = c(t) + \frac{1}{D(p)} \sum_{k=0}^{N} p^{N-k} A_k, \tag{34}$$

where $D(p)$ is defined by (10). The lower limit $k = 0$ has been chosen in (34) instead of $k = 1$ for later convenience.

For dI/dt to vanish identically, it is necessary and sufficient that $D^2(dI/dt)$ vanish identically,

$$D^2 \frac{\partial I}{\partial t} + D^2 p \frac{\partial I}{\partial q} - D^2 \frac{\partial V}{\partial q} \frac{\partial I}{\partial p} = 0. \tag{35}$$

In order to satisfy (35), which is a polynomial equation, the coefficient of each distinct power of p must vanish. Each of the three terms on the left-hand side of (35) can be easily evaluated. A convenient way of expressing the results is

$$D^2 \frac{\partial I}{\partial t} = \sum_{s=0}^{N} \sum_{k=0}^{N} p^{2N-s-k} (a_s \dot{A}_k - A_k \dot{a}_s + a_s a_k \dot{c}), \tag{36a}$$

$$pD^2 \frac{\partial I}{\partial q} = \sum_{s=0}^{N} \sum_{k=0}^{N} p^{2N-s-k} (a_{s+1} A'_k - A_k a'_{s+1}) + \sum_{k=0}^{N} p^{2N-k} A'_{k+1}, \tag{36b}$$

and

$$D^2 \frac{\partial I}{\partial p} = \sum_{s=0}^{N} \sum_{k=0}^{N} p^{2N-s-k} (s-k-1) a_{s-1} A_k + \sum_{k=0}^{N} p^{N-k} (N-k+1) a_N A_{k-1}. \tag{36c}$$

In the equations (36), and in what follows, we use prime for partial derivative with respect to q and dot for partial derivative with respect to t. Substitution of equations (36) into the invariance condition (35) leads to the following equivalent condition:

$$\sum_{s=0}^{N} \sum_{k=0}^{N} p^{2N-s-k} \left[a_s \dot{A}_k + a_{s+1} A'_k - A_k \hat{a}_s - g'_0 a_s a_k + (k-s+1) a_{s-1} A_k \frac{\partial V}{\partial q} \right] + \sum_{k=0}^{N} p^{N-k} (k-N-1) a_N A_{k-1} \frac{\partial V}{\partial q} + \sum_{k=0}^{N} p^{2N-k} A'_{k+1} = 0, \tag{37}$$

where we have defined

$$\hat{f}_s \equiv f'_{s+1} + \dot{f}_s \tag{38}$$

for any functions f_k that depend on q and t.

Equation (37) can be organized in increasing powers of p. This is achieved by use of the identity

$$\sum_{s=0}^{N} \sum_{k=0}^{N} Q_{s,k} = \sum_{s=0}^{N} \sum_{k=s}^{N} Q_{N+s-k,k} + \sum_{s=0}^{N-1} \sum_{k=0}^{s} Q_{s-k,k}, \tag{39}$$

where $Q = \{Q_{s,k}\}$ is an arbitrary square matrix. Formula (39) is obtained by reorganizing the sum of the elements of Q along diagonals instead of along rows. In view of (39), Eq. (37) can be transformed into

$$\sum_{s=0}^{N} p^{N-s} \sum_{k=s-1}^{N} \left[a_{N+s-k} \hat{A}_k - A_k \hat{a}_{N+s-k} + (2k-s-N+1) a_{N+s-k-1} \times A_k \frac{\partial V}{\partial q} - g'_0 a_k a_{N+s-k} \right] + \sum_{s=0}^{N-1} p^{2N-s} \sum_{k=0}^{s} \left[a_{s-k} \hat{A}_k - A_k \hat{a}_{s-k} + (2k-s+1) a_{s-k-1} A_k \frac{\partial V}{\partial q} - g'_0 a_k a_{s-k} \right] = 0. \tag{40}$$

We now notice that the first summation over k can be made to start at zero instead of $s-1$, whereas the second summation over k can be extended from s to N. These changes, in view of (27), only add zeros to the original sums. In addi-

tion, in the first sum over s we change the summation index from s to $s+N$; and in both summations over k we change the summation index from k to $s-k$. With these transformations, (40) reduces to the form

$$\sum_{\sigma=0}^{2N} p^{2N-\sigma}\Gamma_\sigma = 0, \tag{41}$$

where

$$\Gamma_\sigma = \sum_{k=\sigma-N}^{\sigma}\left[a_k\hat{A}_{\sigma-k} - A_{\sigma-k}\hat{a}_k - g_0' a_k a_{\sigma-k} + (\sigma-2k+1)a_{k-1}A_{\sigma-k}\frac{\partial V}{\partial q}\right], \tag{42}$$

valid for $0\leqslant\sigma\leqslant 2N$. We now notice that the lower limit of the sum in (42) can be set to zero whenever $\sigma\neq 2N$. This is true because either $\sigma\leqslant N$, causing all additional $\hat{a}_k, a_k$, and a_{k-1} to be zero, or $\sigma > N$, in which case all additional $\hat{A}_{\sigma-k}$, $A_{\sigma-k}$, and $a_{\sigma-k}$ will be zero. When $\sigma = 2N$ the lower limit of the sum can also be set to zero provided we remove the only nonzero term generated by this transformation. That is, we can write

$$\Gamma_\sigma = \sum_{k=0}^{\sigma}\left[a_k\hat{A}_{\sigma-k} - A_{\sigma-k}\hat{a}_k - g_0' a_k a_{\sigma-k} + (\sigma-2k+1)a_{k-1}A_{\sigma-k}\frac{\partial V}{\partial q}\right] - A'_{2N+1}\delta_{2N,\sigma}, \tag{43}$$

where $\delta_{k,s}$ is the Kronecker delta.

We next compute the sum involving $\hat{A}_{\sigma-k}$ in (43):

$$\sum_{k=0}^{\sigma} a_k\hat{A}_{\sigma-k} = \sum_{k=0}^{\sigma} a_k\left[\sum_{s=0}^{\sigma-k}(g_{\sigma-k-s}\hat{a}_{s-1} + a_{s-1}\hat{g}_{\sigma-k-s}) + a_{\sigma-k}g_0'\right]. \tag{44}$$

Equation (6a) can be used to transform the term with $\hat{g}_{\sigma-k-s}$ on the right-hand side of (44) into a term proportional to $g_{\sigma-k-s-1}$. Then we transform the first term in (44) by use of the algebraic relation

$$\sum_{k=0}^{m}\sum_{s=0}^{m-k} B_{s,k} = \sum_{s=0}^{m}\sum_{k=0}^{m-s} B_{s,k} = \sum_{s=0}^{m}\sum_{k=0}^{m-s} B_{k,s}, \quad m\geqslant 0, \tag{45}$$

which can easily be shown to hold for any matrix $B = \{B_{k,s}\}$. Using (45) and (27) we can transform (44) into

$$\sum_{k=0}^{\sigma}(a_k\hat{A}_{\sigma-k} - a_{\sigma-k}a_k g_0') = \sum_{s=0}^{\sigma}\sum_{k=0}^{\sigma-s}\left[a_{k-1}g_{\sigma-k-s}\hat{a}_s - a_k a_{s-1}g_{\sigma-k-s-1}(\sigma-k-s)\frac{\partial V}{\partial q}\right]. \tag{46}$$

By noticing that the first term on the right-hand side of (46) involves $A_{\sigma-k}$, we can further write

$$\sum_{k=0}^{\sigma}\left[a_k\hat{A}_{\sigma-k} - a_{\sigma-k}a_k g_0'\right] = \sum_{k=0}^{\sigma}\left[A_{\sigma-k}\hat{a}_k - \sum_{s=0}^{\sigma-k} a_k a_{s-1}g_{\sigma-k-s-1} \times(\sigma-k-s)\frac{\partial V}{\partial q}\right]. \tag{47}$$

Substitution of (47) into (43) results in

$$\Gamma_\sigma + \delta_{2N,\sigma}A'_{2N+1} = \sum_{k=0}^{\sigma}\sum_{s=0}^{\sigma-k} a_{s-1}\left[(\sigma-2k+1)a_{k-1}g_{\sigma-k-s} - (\sigma-k-s)a_k g_{\sigma-k-s-1}\right]\frac{\partial V}{\partial q}. \tag{48}$$

We now change the summation index k to $k+1$ in the second term in the double summation on the right-hand side of (48). Additional use of (27) allows us to readjust the limits of the summations to obtain

$$\Gamma_\sigma + \delta_{2N,\sigma}A'_{2N+1} = \sum_{k=0}^{\sigma}\sum_{s=0}^{\sigma-k} a_{s-1}a_{k-1}(s-k)g_{\sigma-k-s}\frac{\partial V}{\partial q} \equiv 0, \tag{49}$$

where we have used (45) again. Therefore all Γ_σ are zero except for $\sigma = 2N$.

Equations (49) and (33) imply

$$D^2\frac{dI}{dt} = -\frac{\partial}{\partial q}\left\{\frac{\det\Lambda_N}{\det\Lambda_{N-1}}\right\}. \tag{50}$$

Therefore $\det\Lambda_N = 0$ implies $dI/dt = 0$. On the other hand, if $dI/dt = 0$, then $[(\det\Lambda_N)/(\det\Lambda_{N-1})]' = 0$ implies $\det\Lambda_N = \psi(t)(\det\Lambda_{N-1})$, where $\psi(t)$ is an arbitrary function of time. Since g_{2N} contains an arbitrary additive function of time, $\psi(t)$ can be chosen to be zero without any loss of generality. This can be seen by expanding $\det\Lambda_N$ in minors along the last row or the last column. That is, $dI/dt = 0$ implies that g_{2N} can be chosen such that $\det\Lambda_N = 0$. Therefore the possibility of choosing the moments g_k such that $\det\Lambda_N = 0$ is both necessary and sufficient for $dI/dt = 0$. This completes the proof of the theorem associated with statement (26).

V. SOME PRELIMINARY APPLICATIONS

A direct application of condition (25) for the case $N = 1$ leads to a first-order linear differential equation for $V(q,t)$ that can be completely integrated to yield the previously known result for potentials that admit an invariant with only one resonance. For $N = 2$, Eq. (25) is a nonlinear integrodifferential equation. In the special case for which $g_0 = 0$, the condition again becomes a linear partial differential equation for the potential that can be integrated. The result is the class of potentials found by Lewis and Leach.[10] However, condition (25) allows a wider class of potentials that admit an exact invariant with two resonances. In the following article,[11] we exhibit solutions of (25) with $N = 2$ for $g_0\neq 0$ and we look for a class of $N = 3$ examples. Our approach there combines features of the moment formulation with the original resonance ansatz. In this section, we use (25) directly to obtain the results of Lewis and Leach for

$N=1$ and $N=2$ and to find two examples with three resonances.

In order to condense the notation and speed the calculation we define

$$P_n=\sum_{k=0}^{n}(-1)^k\alpha_{n-k-1}^{(k)}\frac{q^k}{k!},\quad n\geqslant 0, \tag{51}$$

where the superscripts inside parentheses represent multiple time derivatives,

$$\alpha_{-1}\equiv c(t), \tag{52}$$

where $c(t)$ is the additive function of time in (2), and α_s for $s\geqslant 0$ is an arbitrary function of time. Notice that this definition implies

$$P_0=c(t)\quad\text{and}\quad P_s'=-\dot{P}_{s-1},\quad\text{for}\quad s\geqslant 1. \tag{53}$$

Also define the symbols V_s^k to represent

$$V_s^k\equiv\frac{\partial^k}{\partial t^k}\int^q\int^{x_s}\cdots\int^{x_2}V(x_1,t)dx_1\cdots dx_s. \tag{54}$$

By using (51)–(54) we can calculate the first few g_s in the form

$$g_0=P_1, \tag{55a}$$

$$g_1=P_2, \tag{55b}$$

$$g_2=P_3-P_1V-P_0^{(1)}V_1^0, \tag{55c}$$

$$g_3=P_4-2P_2V-P_1^{(1)}V_1^0+P_1V_1^1+2P_0^{(1)}V_2^1, \tag{55d}$$

$$g_4=P_5-3P_3V-P_2^{(1)}V_1^0+2P_2V_1^1+2P_1^{(1)}V_2^1 -3P_0^{(1)}V_3^2+\tfrac{3}{2}P_1VV+3P_0^{(1)}VV_1^0 -\frac{3}{2}P_0^{(1)}\int^q V(x,t)V(x,t)dx. \tag{55e}$$

We do not proceed beyond g_4 because nonlinearities, which already appear in g_4, quickly become more severe in the higher-order moments and are exacerbated in the determinantal condition. For the three-resonance case we shall use g_5 and g_6, but with very simplifying assumptions. Formulas (55) suffice for $N=1$ and $N=2$ in general. We now consider the cases of one, two, and three resonances in succession.

$N=1$: All the potentials that admit an invariant with one pole are easily determined from (25) using (55a)–(55c). For this case (25) reads

$$g_0g_2=g_1^2, \tag{56}$$

which is a linear ordinary differential equation for $V(q,t)$. Despite its solution being now a well-known result, we shall present it once again for two reasons: (i) to display a compact representation of the potential and the associated invariant in terms of the discrete moments; and (ii) to exhibit an extremely simple and elegant manner of obtaining the result. Both these goals are achieved in one stroke by combining the present formalism with the original formulation of Lewis and Leach. Instead of solving the differential equation (56) directly, we make the simple observation that, for $N=1$, the definition of the moments (5) yields

$$v_1=g_0\quad\text{and}\quad u_1=g_1/g_0. \tag{57}$$

We now can use the representations of the invariant and the potential given by (2) and (4d) to immediately write

$$I(q,p,t)=c(t)+g_0/(p-g_1/g_0), \tag{58}$$

and

$$\frac{\partial V}{\partial q}=-\frac{\partial}{\partial t}\frac{g_1}{g_0}-\frac{g_1}{g_0}\frac{\partial}{\partial q}\frac{g_1}{g_0}, \tag{59}$$

which, as expected, agree with the known result for $N=1$ (see Refs. 10 and 14).

$N=2$: The results for $N=1$ are deceptively simple, as will become clear by studying the $N=2$ case. For the most general situation with $N=2$, we have to make use of all the moments from g_0 up to g_4. This last moment is a quadratic functional of V and the determinantal condition is a cubic functional of V. A close analysis of this condition reveals that it can be recast into a partial differential equation for V_2^0 with quartic nonlinearity. This equation is of third order in q and of second order in t with coefficients that are polynomials in q.

A solution of condition (25) is any set of functions $\{\alpha_{-1},\alpha_0,\ldots,\alpha_{2N+1},V(q,t)\}$ which make it an identity. Keeping this definition of a solution in mind, we can always start by imposing particular values for some or all of the α_s and/or V in an attempt to simplify the equation enough to make it tractable. Following this philosophy, we impose $\alpha_{-1}=\alpha_0=0$, implying $P_1=P_0=0$ and reducing the system (55) to

$$g_0=0, \tag{60a}$$

$$g_1=Q_0, \tag{60b}$$

$$g_2=Q_1, \tag{60c}$$

$$g_3=Q_2-2Q_0V, \tag{60d}$$

$$g_4=Q_3-3Q_1V-Q_0^{(1)}V_1^0+2Q_0V_1^1, \tag{60e}$$

where

$$Q_s=\sum_{k=0}^{s}(-1)^k\alpha_{s-k+1}^{(k)}\frac{q^k}{k!}. \tag{61}$$

By this choice of the two initial α_s, the determinantal condition is reduced to

$$Q_1V+2Q_0V_1^1-Q_0^{(1)}V_1^0=2\frac{Q_1Q_2}{Q_0}-\frac{Q_1^3}{Q_0^2}-Q_3. \tag{62}$$

Equation (62) can be viewed as a first-order linear partial differential equation for V_1^0, which is the indefinite integral of $V(q,t)$ with respect to q. It can be integrated exactly without any further constraints on the remaining arbitrary functions of time. The results fully coincide with those found by Lewis and Leach[10] if the functions ρ and α are related to α_1 and α_2 by

$$\rho(t)=\frac{1}{\sqrt{\alpha_1(t)}} \tag{63}$$

and

$$\alpha(t)=-\frac{1}{2\sqrt{\alpha_1(t)}}\int^t\frac{\alpha_2(t')}{\sqrt{\alpha_1(t')}}dt'.$$

$N=3$: To illustrate the use of our condition (25) further, we derive two potentials for which invariants with three resonances exist and we construct the invariants by using (17). These examples are the result of a preliminary study of the $N=3$ case.

We consider $N=3$ and choose $g_0=g_1=g_3=0$ and $g_2=1$. This is consistent with (6a) and (6b). All other g_k

can be calculated from (6a). The result is

$$g_4 = -3V(q,t) - \tfrac{3}{2}V_1(t), \tag{64a}$$

$$g_5 = 3\frac{\partial}{\partial t}\int^q V(x,t)dx + \frac{3}{2}\frac{dV_1}{dt} + V_2(t), \tag{64b}$$

$$g_6 = -3\frac{\partial^2}{\partial t^2}\int^q dx\int^x dy\, V(y,t) - 3\frac{d^2V_1}{dt^2}q^2 - \frac{dV_2}{dt}q - V_3(t) + \tfrac{15}{2}V_1V(q,t) + \tfrac{15}{2}V^2(q,t), \tag{64c}$$

where $V_1(t)$, $V_2(t)$, and $V_3(t)$ are unspecified functions of time. Condition (25) for this case is simply $g_6 - g_4^2 = 0$, which is the following integrodifferential equation for the potential:

$$V^2(q,t) + V_1(t)V(q,t) + 2\frac{\partial^2}{\partial t^2}\int^q\int^y V(x,t)dx\,dy = -\phi(q,t), \tag{65}$$

where

$$\phi(q,t) \equiv 2\frac{d^2V_1}{dt^2}q^2 + \frac{2}{3}\frac{dV_2}{dt}q + \frac{2}{3}V_3 + \frac{3}{2}V_1^2. \tag{66}$$

The search for solutions is simplified by starting with the equation

$$2\frac{\partial^2 V}{\partial t^2} + \frac{\partial^2 V^2}{\partial q^2} + V_1(t)\frac{\partial^2 V}{\partial q^2} = -4\frac{d^2V_1}{dt^2}, \tag{67}$$

which is obtained by taking the second spatial derivative of (65). Any solution of (67) can then be substituted into (65), which becomes an equation for the remaining unknowns. Particular solutions of (65)–(67) can be found easily. Two solutions are

$$V(q,t) = At\sqrt{q}, \quad V_1 = 0, \quad V_2 = -\tfrac{1}{2}A^2t^3 + B, \quad V_3 = 0, \tag{68a}$$

and

$$V(q,t) = -(A/2) \pm (A^2 + 4C + 4Bq)^{1/2},$$
$$V_1 = A, \quad V_2 = -6Bt + D, \quad V_3 = -\tfrac{9}{4}A^2 - 6C, \tag{68b}$$

where A, B, C, and D are arbitrary constants. Using those solutions in (64) to determine the g_k and, with the g_k calculating the a_n from (15), we obtain the corresponding invariants

$$I = 1/(p^3 + 3Apt\sqrt{q} - 2Aq\sqrt{q} + \tfrac{1}{2}A^2t^3 - 3B) \tag{69a}$$

and

$$I = 1/[p^3 \pm 3p(A^2 + 4C + 4Bq)^{1/2} + 6Bt], \tag{69b}$$

respectively. It is interesting to notice that, although both invariants are explicitly time dependent, the first system is nonautonomous, while the second is autonomous. The latter possesses the energy invariant in addition to I and, therefore, we have a complete set of two invariants. The trajectories could be determined by, for example, solving the energy invariant for the momentum p and substituting this into I.

VI. FINAL REMARKS

We have elaborated the momentum-resonance ansatz in an advantageous way by introducing discrete momentum moments. We proved a linearization theorem, which shows that an N-resonance invariant is associated with the solution of a system of linear algebraic equations; and we derived a necessary and sufficient condition for the existence of an N-resonance invariant. Preliminary applications of the expanded formulation include deriving two examples of three-resonance invariants and rederiving the previously known examples with one and two resonances. We present a more detailed study of the two- and three-resonance cases in the subsequent article in this journal.[11]

It is likely that further understanding of invariants of time-dependent dynamical systems can be achieved by using ideas described here. An attractive line of investigation would be to combine the ideas here with the Lagrangian approach developed by Lewis and Leach.[10] Application to the Vlasov–Poisson equations of plasma physics and study of quantum mechanical systems and systems with more than one spatial dimension may also be fruitful.

ACKNOWLEDGMENTS

We thank Gene H. Golub and Daniel C. Barnes for helpful discussions. Professor Golub pointed out that our discrete moments have the form of solutions of algebraic recursion relations that arise in numerical analysis of ordinary differential equations with constant coefficients. Dr. Barnes noticed the connection between the Vlasov equation and the differential recursion relation satisfied by the discrete moments.

This work was performed during a visit to the Controlled Thermonuclear Research Division of the Los Alamos National Laboratory by João Goedert. Financial support of his visit was provided by the Conselho Nacional de Pesquisas (CNPq) of Brazil and the Universidade Federal do Rio Grande do Sul (UFRS) in Porto Alegre, Brazil.

[1] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies* (Cambridge U. P., Cambridge, 1937), 4th ed.
[2] G. Darboux, Arch. Neerlandaise (ii) **6**, 371 (1901); see also, G. Thompson, J. Phys. A **17**, 985 (1984).
[3] C. R. Holt, J. Math. Phys. **23**, 1037 (1982).
[4] L. S. Hall, Physica D **8**, 90 (1983).
[5] G. Schmidt, *Physics of High Temperature Plasmas* (Academic, New York, 1979), 2nd ed.
[6] H. R. Lewis, J. Math. Phys. **9**, 1976 (1968).
[7] H. R. Lewis and P. G. L. Leach, J. Math. Phys. **23**, 2371 (1982).
[8] W. Sarlet and J. R. Ray, J. Math. Phys. **22**, 2504 (1981); see Sec. 4.2.
[9] H. R. Lewis and K. R. Symon, Phys. Fluids **27**, 192 (1984).
[10] H. R. Lewis and P. G. L. Leach, Ann. Phys. (NY) **164**, 47 (1985).
[11] H. R. Lewis and J. Goedert, J. Math. Phys. **28**, 736 (1987).
[12] D. J. Newman, "Approximation with rational functions," *Conference Board of the Mathematical Sciences*, Regional Conference Series in Mathematics No. 41 (Am. Math. Soc., Providence, RI, 1970).
[13] W. Sarlet (private communication).
[14] P. G. L. Leach, H. R. Lewis, and W. Sarlet, J. Math. Phys. **25**, 486 (1984).
[15] F. B. Hildebrand, *Finite-Difference Equations and Simulations* (Prentice-Hall, Englewood Cliffs, NJ, 1968).
[16] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications* (Univ. California P., Berkeley, 1958).
[17] G. Kowalewsky, *Einführung in die Determinantentheorie* (Walter de Gruyter, Berlin, 1954).
[18] A. S. Householder, *The Theory of Matrices in Numerical Analysis* (Blaisdell, New York, 1964).

Rational functions of momentum as invariants for one-dimensional, time-dependent potentials: Two- and three-resonance cases

H. Ralph Lewis and João Goedert[a)]
Los Alamos National Laboratory, MS-F642, Los Alamos, New Mexico 87545

(Received 19 November 1985; accepted for publication 29 October 1986)

The momentum-moment formulation of Goedert and Lewis [J. Math. Phys. 28, 728 (1987)] and the momentum-resonance formulation of Lewis and Leach [Ann. Phys. (NY) **164**, 47 (1985)] are used to study one-dimensional, time-dependent potentials that admit invariants which are rational functions of momentum with two or three simple poles. New examples are presented.

I. INTRODUCTION

In the preceding article in this journal,[1] Goedert and Lewis present a formulation for invariants for one-dimensional, time-dependent potentials in terms of discrete momentum moments. Their discussion is applicable to the case where the invariant is a function of a rational function of momentum with simple poles. It extends and supplements the analysis by Lewis and Leach[2] of a momentum-resonance ansatz. That ansatz postulates the invariant to be a rational function of momentum with simple poles; the analysis by Lewis and Leach makes particular use of an expression for the invariant as a sum of "resonance" terms of the form $v_n(q,t)/(p-u_n(q,t))$, where q, p, and t are position, momentum, and time, respectively. There are three major results in the article by Goedert and Lewis[1]: (i) a linearization theorem,which relates the q dependence of the invariant to the solution of a linear system of algebraic equations; (ii) a necessary and sufficient condition for a potential to admit an invariant that is a rational function of p with simple poles; and (iii) a derivation of certain relations between the discrete moments of Goedert and Lewis and the quantities $u_n(q,t)$ and $v_n(q,t)$. In the present article, we use these results to study the cases of two and three resonances.

The case of one resonance has been treated earlier by Sarlet,[3] Lewis and Leach,[2] and Leach, Lewis, and Sarlet.[4] All potentials that admit an invariant with one resonance are known. The associated forces are rational functions of q with coefficients expressible in terms of three arbitrary functions of t. The treatment of this case is particularly simple in the discrete-moment formulation and is presented in the preceding article.[1] It is also possible to derive a second invariant for these potentials.[5]

A class of two-resonance cases with potentials involving an arbitrary function of a time-dependent linear function of q has been obtained by Lewis and Leach[2,6] and Sarlet and Ray.[7] This class can also be obtained easily in the framework of the discrete-moment formulation of Goedert and Lewis.[1] An outstanding question has been whether this class exhausts the examples with two resonances. We have answered the question by finding examples with two resonances outside this class. Our analysis is based on a characterization of all two-resonance cases in terms of an interesting implicit change of independent variables. One example of two-resonance cases is for precisely the same class of potentials for which there exists a one-resonance invariant. The two-resonance invariant for a potential in this class is not simply a function of the one-resonance invariant. In the other examples, the force and invariant are either in a parametric form or in terms of the solution of a nonlinear first-order differential equation.

We have examined the three-resonance case by using a certain generalization of our treatment of the two-resonance case. The objective was to obtain a class of three-resonance examples that involve an arbitrary function of some function of q and t, in analogy with the previously known class of two-resonance examples. That generalization only led to the potential for a driven, time-dependent linear oscillator.

In Sec. II we summarize some basic formulas of the momentum-resonance ansatz[2] and the discrete-moment formulation.[1] We also develop the framework within which we treat the two- and three-resonance cases in this article. In Secs. III and IV we analyze the two-resonance case. In Sec. V we consider the three-resonance case. In Sec. VI we make some concluding remarks, including some probably fruitful directions for further work.

II. BACKGROUND AND PRELIMINARY CONSIDERATIONS

We begin by summarizing some aspects of the work by Lewis and Leach[2] and by Goedert and Lewis.[1] We consider a system described by a Hamiltonian

$$H=\tfrac{1}{2}p^2+V(q,t) \tag{1}$$

that admits an invariant (constant of the motion) of the form

$$I(q,p,t)=c(t)+\sum_{n=1}^{N}\frac{v_n(q,t)}{p-u_n(q,t)}, \tag{2}$$

where q, p, and t stand for position, momentum, and time, respectively. We denote an invariant of this type as an invariant with N momentum resonances. The functions c, u_n, and v_n satisfy

$$\frac{dc}{dt}+\frac{\partial}{\partial q}\sum_{n=1}^{N}v_n=0, \tag{3a}$$

$$\frac{\partial v_n}{\partial t}+\frac{\partial}{\partial q}(u_n v_n)=0, \tag{3b}$$

$$\frac{\partial u_n}{\partial t}+u_n\frac{\partial u_n}{\partial q}=-\frac{\partial V}{\partial q}. \tag{3c}$$

[a)] Permanent address: Instituto de Física, Universidade Federal do Rio Grande do Sul, 90049-Porto Alegre, Rio Grande do Sul, Brazil.

 0022-2488/87/030736-08$02.50

From the functions u_n and v_n we construct a set of N discrete moments, $g_k(q,t)$, defined by

$$g_k(q,t) = \sum_{n=1}^{N} u_n^k v_n. \tag{4}$$

They satisfy an algebraic recursion relation,

$$g_l = -\sum_{n=1}^{N} a_n g_{l-n}, \quad l \geqslant N, \tag{5}$$

where the a_n are the coefficients of a polynomial whose roots are the u_n,

$$u_n^N + \sum_{k=1}^{N} a_k u_n^{N-k} = 0, \quad 1 \leqslant n \leqslant N. \tag{6}$$

The invariant (2) can be expressed in terms of the discrete moments $g_k(q,t)$ and the coefficients $a_n(q,t)$ by the formula

$$I(q,p,t) = c(t) + \frac{\Sigma_{n=1}^{N} p^{N-n} \Sigma_{k=1}^{n} a_{k-1} g_{n-k}}{\Sigma_{n=0}^{N} a_n p^{N-n}}, \tag{7}$$

where, by definition,

$$a_0 \equiv 1. \tag{8}$$

The moments satisfy a differential recursion relation in addition to the algebraic recursion relation (5),

$$\frac{\partial g_k}{\partial q} = -\frac{\partial g_{k-1}}{\partial t} - (k-1) g_{k-2} \frac{\partial V}{\partial q}, \quad k \geqslant 1, \tag{9a}$$

with initial condition

$$g_0(q,t) = -\dot{\alpha}_{-1}(t) q + \alpha_0(t), \quad \alpha_{-1}(t) \equiv c(t). \tag{9b}$$

The quantities and relations defined in (1)–(9b) have been discussed by Lewis and Leach[2] and by Goedert and Lewis.[1]

As is indicated in the foregoing paragraph, the u_n are the roots of an Nth-degree polynomial whose coefficients figure importantly in the structure of an N-resonance invariant. Motivated by this fact, for $N \leqslant 4$, we introduce a representation of the u_n that displays clearly the relationship which the roots of the polynomial must have among themselves. That relationship can be expressed by

$$u_k = A_1 + \sum_{j=2}^{N} \gamma_{kj} A_j, \quad 1 \leqslant k \leqslant N, \tag{10}$$

where the γ_{kj} are numbers that only depend on the value of N and any particular nonzero γ_{kj} can be chosen to be of modulus unity. For $N \leqslant 4$, relation (10) can be verified and numbers γ_{kj} determined by examining the formulas for the roots of quadratic, cubic and quartic equations.[8] We choose the numbers γ_{kj} such that the u_k can be written as follows:

For quadratic equations,

$$u_1 = A_1 + A_2, \quad u_2 = A_1 - A_2. \tag{11}$$

For cubic equations,

$$\begin{aligned} u_1 &= A_1 + A_2 + A_3, \\ u_2 &= A_1 + \omega A_2 + \omega^* A_3, \\ u_3 &= A_1 + \omega^* A_2 + \omega A_3, \end{aligned} \tag{12}$$

where

$$\omega = e^{i2\pi/3} = -\tfrac{1}{2} + i\sqrt{3}/2, \quad \omega^2 = 1/\omega = \omega^*. \tag{13}$$

For quartic equations,

$$\begin{aligned} u_1 &= A_1 - A_2 + iA_3, \\ u_2 &= A_1 - A_2 - iA_3, \\ u_3 &= A_1 + A_2 + iA_4, \\ u_4 &= A_1 + A_2 - iA_4. \end{aligned} \tag{14}$$

The functions v_n are related to the u_n and the moments g_k through the first N of Eqs. (4):

$$g_k = \sum_{n=1}^{N} u_n^k v_n, \quad 0 \leqslant k \leqslant N-1, \tag{15}$$

where the g_k are calculated from the differential recursion relation (9a) and (9b) in terms of the potential and $2N+2$ unspecified functions of time ("integration constants"). The first of Eqs. (15), which is the definition of g_0, is the solution of (3a).

Under the assumption that the u_n satisfy (3c), the N Eqs. (3b) are equivalent to

$$\begin{aligned} &\frac{\partial}{\partial q} \sum_{n=1}^{N} u_n^k v_n + \frac{\partial}{\partial q} \sum_{n=1}^{N} u_n^{k-1} v_n \\ &\quad + (k-1) \frac{\partial V}{\partial q} \sum_{n=1}^{N} u_n^{k-2} v_n, \quad 1 \leqslant k \leqslant N. \end{aligned} \tag{16}$$

This can be obtained by adding the product of (3b) with u_n^{k-1} to the product of (3c) with $(k-1)u_n^{k-2}v_n$ and summing over n. For k in (16) in the range $1 \leqslant k \leqslant N-1$, we can use (15) to eliminate the summations in (16), thereby obtaining (9a). Because the g_k in (15) satisfy (9a) by assumption, Eq. (16) for $1 \leqslant k \leqslant N-1$ are satisfied identically when the v_n are determined from (15). When $k = N$, Eq. (16) is not satisfied identically. It is the single condition that must be satisfied in order that functions u_n and v_n satisfying (3c) and (15) be a solution of the system (3a)–(3c). By using the algebraic recursion relation (5), we can write (16) for $k = N$ as

$$\frac{\partial}{\partial q} \sum_{k=1}^{N} a_k g_{N-k} = \frac{\partial g_{N-1}}{\partial t} + (N-1) g_{N-2} \frac{\partial V}{\partial q}. \tag{17}$$

Because the a_k are the coefficients of the polynomial (6) whose roots are the u_n, the a_k can be expressed as certain symmetric functions of the u_n. By so expressing the a_k, we transform (17) into an equation relating the u_n and the first N moments.

In the next two sections, we express the u_n for the two- and three-resonance cases as in (11) and (12) and base our analysis on (3c), (15), and (17).

III. TWO RESONANCES

Substitute (11) into (3c) to obtain the following equations for A_1 and A_2:

$$\frac{\partial A_1}{\partial t} + \frac{\partial A_2}{\partial t} + \frac{\partial}{\partial q}\left[\frac{1}{2}(A_1^2 + A_2^2) + A_1 A_2\right] = -\frac{\partial V}{\partial q}, \tag{18a}$$

$$\frac{\partial A_1}{\partial t} - \frac{\partial A_2}{\partial t} + \frac{\partial}{\partial q}\left[\frac{1}{2}(A_1^2 + A_2^2) - A_1 A_2\right] = -\frac{\partial V}{\partial q}. \tag{18b}$$

The difference and sum of these equations yield

$$\frac{\partial A_2}{\partial t}+\frac{\partial}{\partial q}(A_1A_2)=0, \tag{19}$$

$$\frac{\partial A_1}{\partial t}+\frac{1}{2}\frac{\partial}{\partial q}(A_1^2+A_2^2)=-\frac{\partial V}{\partial q}. \tag{20}$$

We treat (19) by introducing an implicitly defined variable x. Let $f(q,t)$ be a particular solution of

$$\frac{\partial f}{\partial t}+A_1\frac{\partial f}{\partial q}=0. \tag{21}$$

Define x by

$$x=f(q,t) \tag{22}$$

and the inverse by

$$q=r(x,t). \tag{23}$$

From the identity

$$x=f[r(x,t),t] \tag{24}$$

it is easy to establish the relations

$$\frac{\partial f}{\partial q}=\frac{1}{\partial r/\partial x},\quad \frac{\partial f}{\partial t}=-\frac{\partial r/\partial t}{\partial r/\partial x}. \tag{25}$$

From (21) and (25), we see that $A_1(q,t)$ is expressible as a function of x and t by

$$A_1(q,t)=\frac{\partial r}{\partial t}. \tag{26}$$

Equation (19), written in terms of (x,t) instead of (q,t), is an ordinary differential equation in t whose solution gives $A_2(q,t)$ in terms of x and t:

$$A_2(q,t)=\frac{W(x)}{\partial r/\partial x}, \tag{27}$$

where $W(x)$ is arbitrary.

Having expressed A_1 and A_2 in terms of the as yet unknown transformation function $r(x,t)$ and the arbitrary function $W(x)$, we shall view (20) as a specification of $\partial V/\partial q$ in terms of (x,t) and proceed to write (17) in terms of (x,t) as well. This will lead to an interesting and usable characterization of all two-resonance cases in terms of the transformation function.

By using (20), we express $-\partial V/\partial q$ in terms of $r(x,t)$ as

$$-\frac{\partial V}{\partial q}=\frac{\partial^2 r}{\partial t^2}+\frac{1}{2}\frac{1}{\partial r/\partial x}\frac{\partial}{\partial x}\left[\left(\frac{W(x)}{\partial r/\partial x}\right)^2\right]. \tag{28}$$

The coefficients a_k and moments g_k required for (17) can be written as

$$a_1=-(u_1+u_2)=-2A_1,\quad a_2=u_1u_2=A_1^2-A_2^2, \tag{29}$$

$$g_0=\epsilon_1(t)q+\epsilon_2(t),\quad g_1=-\tfrac{1}{2}\dot\epsilon_1(t)q^2-\dot\epsilon_2(t)q-\dot\epsilon_3(t), \tag{30}$$

where a dot denotes differentiation with respect to t and where ϵ_1, ϵ_2, and ϵ_3 are arbitrary functions of t. Condition (17) can be written as

$$\frac{\partial g_1}{\partial t}-\frac{\partial}{\partial q}[u_1u_2g_0-(u_1+u_2)g_1]+g_0\frac{\partial V}{\partial q}=0. \tag{31}$$

It can be transformed to

$$\frac{\partial}{\partial t}\left[\frac{g_1-g_0A_1}{A_2}\right]+\frac{\partial}{\partial q}\left[\frac{A_1(g_1-g_0A_1)}{A_2}+g_0A_2\right]=0 \tag{32}$$

by multiplying by $2/(u_1-u_2)=1/A_2$. We define

$$h_0(x,t)=g_0[r(x,t),t],\quad h_1(x,t)=g_1[r(x,t),t], \tag{33}$$

and write (32) in terms of (x,t) as

$$\frac{\partial}{\partial t}\left[\frac{(\partial r/\partial x)^2(h_1-h_0(\partial r/\partial t))}{W(x)}\right]+\frac{\partial}{\partial x}\left[\frac{h_0W(x)}{\partial r/\partial x}\right]=0. \tag{34}$$

Condition (34), which is the expression of (17) in terms of (x,t), has the following significance. For a two-resonance invariant to exist, it is necessary and sufficient that there exist functions $r(x,t)$ and $W(x)$ that satisfy (34), where $h_0(x,t)$ and $h_1(x,t)$ are defined by (33). The remainder of this section is devoted to exploring some properties of (34) and to writing two-resonance invariants and their associated forces in terms of solutions of (34). In the next section, we proceed to obtain new examples with two resonances.

The second derivative of $r(x,t)$ with respect to t is related to derivatives of $f(q,t)$ by

$$r_{tt}=-\left[\frac{\partial}{\partial t}-\frac{f_t}{f_q}\frac{\partial}{\partial q}\right]\frac{f_t}{f_q}, \tag{35}$$

where we have used the subscript notation for derivatives. Thus we can express $-\partial V/\partial q$ from (28) in terms of $f(q,t)$ as

$$-\frac{\partial V}{\partial q}=-\left[\frac{\partial}{\partial t}-\frac{f_t}{f_q}\frac{\partial}{\partial q}\right]\frac{f_t}{f_q}+\frac{1}{2}\frac{\partial}{\partial q}\{f_qW[f(q,t)]\}^2. \tag{36}$$

The class of potentials that admit an invariant which is quadratic in p is known[2,6] and is a subset of the potentials that admit a two-resonance invariant. (This is because the reciprocal of a quadratic invariant is an invariant with two poles.) Those potentials can be obtained by taking $f(q,t)$ to be a function of a linear function of q,

$$f(q,t)=F[(q-\alpha)/\rho], \tag{37}$$

where $\alpha(t)$ and $\rho(t)$ are arbitrary functions of t. Direct calculation shows that this choice of $f(q,t)$ yields

$$-\frac{\partial V}{\partial q}=\frac{\ddot\rho}{\rho}(q-\alpha)+\ddot\alpha+\frac{1}{2\rho^2}\frac{\partial}{\partial q}\left[F'\left(\frac{q-\alpha}{\rho}\right)W\left(\frac{q-\alpha}{\rho}\right)\right]^2, \tag{38}$$

where $F'(x)=dF/dx$ and a dot denotes differentiation with respect to t. Because W and F are arbitrary, (38) gives the class of potentials with invariants quadratic in p. It should be noted that the form of F is unimportant for obtaining this result. Since any function of a solution of (21) is also a solution, we could, for example, choose F to be linear. In that case, we must be able to satisfy condition (34) by taking $r(x,t)$ to be linear in x. That is indeed possible for arbitrary $W(x)$ by choosing, for example, $h_0(x,t)$ to be identically zero and suitably choosing $h_1(x,t)$, which is then only a function of t.

In order to obtain more general solutions of condition (34), it may be useful to view it as an ordinary differential equation for $W^2(x)$ given $r(x,t)$, $h_0(x,t)$, and $h_1(x,t)$:

$$\frac{1}{2}\frac{h_0}{r_x}\frac{dW^2}{dx}+\left[\frac{\partial}{\partial x}\left(\frac{h_0}{r_x}\right)\right]W^2=-\frac{\partial}{\partial t}\left[r_x^2(h_1-h_0r_t)\right]. \tag{39}$$

The solution is

$$W^2(x)=\left(\frac{r_x}{h_0}\right)^2\left\{D(t)-2\int_{x_0}^{x}dx'\,\frac{h_0}{r_{x'}}\frac{\partial}{\partial t}\left[r_{x'}^2(h_1-h_0r_t)\right]\right\}, \tag{40}$$

where

$$D(t)=W^2(x)/(r_x/h_0)^2\big|_{x=x_0}. \tag{40'}$$

As long as $W^2(x)$ calculated from (40) is indeed independent of t, then (40) represents a valid solution of (34). Thus a characterization of two-resonance examples is that they can be derived from any functions $r(x,t)$, $h_0(x,t)$, and $h_1(x,t)$ if $W^2(x)$ calculated from those functions *via* (40) is indeed not a function of t. [It is to be remembered that $h_0(x,t)$ and $h_1(x,t)$ are linear and quadratic functions of $r(x,t)$ as given by (33) and (30).]

It might be thought that new examples with two resonances could be found by requiring that the ratios of the coefficients in (39) be independent of t, so that $W^2(x)$ would satisfy an equation that does not involve t. The two equations that express this requirement, under the assumption that $h_0(x,t)$ not vanish identically, are

$$\frac{\partial}{\partial t}\left[\frac{\partial}{\partial x}\log\left(\frac{h_0}{r_x}\right)\right]=0, \tag{41}$$

$$\frac{\partial}{\partial t}\left\{\frac{r_x}{h_0}\frac{\partial}{\partial t}\left[r_x^2(h_1-h_0r_t)\right]\right\}. \tag{42}$$

Analysis of these equations leads to the conclusion that $r(x,t)$ can be written as a linear function of a function of x. However, since this is the same as taking $f(q,t)$ to be a function of a linear function of q, the examples are again from the class given by (38).

The failure of the approach described in the preceding paragraph to yield new examples is associated with the fact that $W(x)$ enters the determination of $\partial V/\partial q$ from (36) in an irrelevant way except when $h_0(x,t)$ vanishes identically. If $h_0(x,t)$ vanishes identically,then $W(x)$ can be removed from (34) by multiplication and formula (36) for $\partial V/\partial q$ contains $W(x)$ in a nontrivial way. Now suppose that $h_0(x,t)$ does not vanish identically. Define y and $s(y,t)$ by

$$y=\int^{x}W(x')dx', \tag{43}$$

$$s(y,t)=r(x,t)=q, \tag{44}$$

and the inversion of (44) to give y as a function of q and t by

$$y=\phi(q,t). \tag{45}$$

In terms of y and t, condition (34) is

$$\frac{\partial}{\partial t}\left\{\left(\frac{\partial s}{\partial y}\right)^2\left[g_1(s,t)-g_0(s,t)\frac{\partial s}{\partial t}\right]\right\}+\frac{\partial}{\partial y}\left[\frac{g_0(s,t)}{\partial s/\partial y}\right]=0, \tag{46}$$

which does not involve the function W. Equation (36) can be written in terms of $\phi(q,t)$ as

$$-\frac{\partial V}{\partial q}=-\left[\frac{\partial}{\partial t}-\frac{\phi_t}{\phi_q}\frac{\partial}{\partial q}\right]\frac{\phi_t}{\phi_q}+\frac{1}{2}\frac{\partial}{\partial q}\phi_q^2, \tag{47}$$

which also does not involve W. Therefore, if $h_0(x,t)=g_0[r(x,t),t]$ is not identically zero, then $\partial V/\partial q$ is independent of the function W.

In fact, we can go further and transform (46) to a form that does not involve the functions g_0 and g_1 at all. Again assume that $g_0(q,t)$ does not vanish identically and define

$$K(y,t)=\int^{s(y,t)}dq'\,g_0(q',t). \tag{48}$$

The derivatives of $K(y,t)$ are

$$\frac{\partial K}{\partial y}=g_0(s,t)\frac{\partial s}{\partial y}, \tag{49}$$

$$\frac{\partial K}{\partial t}=g_0(s,t)\frac{\partial s}{\partial t}+\int^{s(y,t)}dq'\,\dot g_0(q',t)=g_0(s,t)\frac{\partial s}{\partial t}-g_1(s,t), \tag{50}$$

where we have chosen the "integration constant" in (48) such that

$$K(y,t)+\tfrac{1}{2}\epsilon_1(t)s^2+\epsilon_2(t)s+\epsilon_3(t). \tag{51}$$

Then (46) can be written as

$$\frac{\partial}{\partial t}\left[\left(\frac{\partial s}{\partial y}\right)^2\frac{\partial K}{\partial t}\right]=\frac{\partial}{\partial y}\left[\frac{\partial K/\partial y}{(\partial s/\partial y)^2}\right]. \tag{52}$$

As will be apparent shortly, it is useful to define a variable τ and a function $J(y,\tau)$ by

$$\tau=\begin{cases}2\int^{t}dt'\,\epsilon_1(t'), & \text{for } \epsilon_1(t)\not\equiv 0,\\ \int^{t}dt'\,\epsilon_2^2(t'), & \text{for } \epsilon_1(t)\equiv 0;\end{cases} \tag{53}$$

and

$$J(y,\tau)=\begin{cases}[1/(2\epsilon_1)](\epsilon_1 s+\epsilon_2)^2=K(y,t)+b(\tau), & \\ \qquad\qquad\qquad\qquad \text{for } \epsilon_1(t)\not\equiv 0,\\ \epsilon_2 s+\epsilon_3=K(y,t), \qquad \text{for } \epsilon_1(t)\equiv 0,\end{cases} \tag{54}$$

where

$$b(\tau)=\frac{\epsilon_2^2(t)}{2\epsilon_1(t)}-\epsilon_3(t). \tag{55}$$

Then (52) becomes

$$\frac{\partial}{\partial\tau}\left[\frac{J_y^2}{J}\frac{\partial}{\partial\tau}(J-b)\right]=\frac{\partial}{\partial y}\left[\frac{J}{J_y}\right],\quad \text{for } \epsilon_1(t)\not\equiv 0, \tag{56a}$$

$$\frac{\partial}{\partial\tau}\left[J_y^2J_\tau\right]=\frac{\partial}{\partial y}\left[\frac{1}{J_y}\right],\quad \text{for } \epsilon_1(t)\equiv 0. \tag{56b}$$

Equation (56a) or (56b) is the necessary and sufficient condition for a two-resonance invariant to exist if the moment $g_0(q,t)$ does not vanish identically.

The introduction of $J(y,\tau)$ is especially convenient because the force $-\partial V/\partial q$ and the invariant can both be expressed directly in terms of $J(y,\tau)$. To calculate the force, we

use (44) and (45) to eliminate y from (54),

$$J[\phi(q,t),\tau] = \begin{cases} \{1/[2\epsilon_1(t)]\}[\epsilon_1(t)q+\epsilon_2(t)]^2, & \epsilon_1(t)\neq 0, \\ \epsilon_2(t)q+\epsilon_3(t), & \epsilon_1(t)\equiv 0. \end{cases} \tag{57}$$

From the two equations obtained from differentiating (57) with respect to q or t, we can solve for ϕ_q and ϕ_t and use them in (47) to calculate $-\partial V/\partial q$. We can express the result conveniently in terms of

$$X \equiv \epsilon_1(t)q+\epsilon_2(t) \quad [=g_0(q,t)], \tag{58}$$

$$\rho(t) = 1/\epsilon_2(t), \quad \alpha(t) = -\epsilon_3(t)/\epsilon_2(t). \tag{59}$$

The derivatives ϕ_q and ϕ_t are

$$\phi_q = \begin{cases} X/J_y, & \epsilon_1(t)\neq 0, \\ \epsilon_2/J_y, & \epsilon_1(t)\equiv 0, \end{cases} \tag{60}$$

$$\phi_t = \begin{cases} \dfrac{(\partial/\partial t)\{[1/(2\epsilon_1)]X^2\} - 2\epsilon_1 J_\tau}{J_y}, & \epsilon_1(t)\neq 0, \\ \dfrac{\dot\epsilon_2 q + \dot\epsilon_3 - J_\tau \epsilon_2^2}{J_y}, & \epsilon_1(t)\equiv 0. \end{cases} \tag{61}$$

The expressions for the force are

$$-\frac{\partial V}{\partial q} = -\frac{J_{yy}}{J_y^4}X^3 + \left[\frac{1}{2}\left(\frac{1}{\epsilon_1}\right)^{\cdot\cdot} - \frac{1}{4}\epsilon_1\left(\frac{1}{\epsilon_1}\right)^{\cdot 2} + \frac{\epsilon_1}{J_y^2}\right]X - \left(\frac{\epsilon_2}{\epsilon_1}\right)^{\cdot\cdot} + 4\epsilon_1^2 J_{\tau\tau}\frac{1}{X} - 4\epsilon_1^3 J_\tau^2 \frac{1}{X^3}, \quad \epsilon_1(t)\neq 0, \tag{62}$$

$$-\frac{\partial V}{\partial q} = \frac{\ddot\rho}{\rho}q + \frac{\rho\ddot\alpha - \alpha\ddot\rho}{\rho} - \frac{2}{\rho^3}\left[\frac{J_{y\tau}J_\tau}{J_y} + \frac{J_{yy}}{J_y^4}\right], \quad \epsilon_1(t)\equiv 0. \tag{63}$$

In order to calculate the invariant from (2), we need to calculate $u_1(q,t)$, $u_2(q,t)$, $v_1(q,t)$, and $v_2(q,t)$. We use (54) as an explicit relation between $J(y,\tau)$ and $s(y,t)$:

$$J(y,\tau) = \begin{cases} \{1/[2\epsilon_1(t)]\}[\epsilon_1(t)s(y,t)+\epsilon_2(t)]^2, & \epsilon_1(t)\neq 0, \\ \epsilon_2(t)s(y,t)+\epsilon_3(t), & \epsilon_1(t)\equiv 0. \end{cases} \tag{64}$$

From the two equations obtained from differentiating (64) with respect to y or t, we can solve for s_y and s_t. From (26), (27), (43), and (44), we can express A_1 and A_2 as

$$A_1 = s_t \quad \text{and} \quad A_2 = 1/s_y. \tag{65}$$

The result for A_1 and A_2 is

$$A_1 = \begin{cases} \dfrac{2\epsilon_1 J_\tau}{X} + \dfrac{1}{2}\left(\dfrac{1}{\epsilon_1}\right)^{\cdot}X - \left(\dfrac{\epsilon_2}{\epsilon_1}\right)^{\cdot}, & \epsilon_1(t)\neq 0, \\ \epsilon_2 J_\tau + \left(\dfrac{1}{\epsilon_2}\right)^{\cdot}(\epsilon_2 q+\epsilon_3) - \left(\dfrac{\epsilon_3}{\epsilon_2}\right)^{\cdot}, & \epsilon_1(t)\equiv 0, \end{cases} \tag{66}$$

$$A_2 = \begin{cases} X/J_y, & \epsilon_1(t)\neq 0, \\ \epsilon_2/J_y, & \epsilon_1(t)\equiv 0. \end{cases} \tag{67}$$

The functions u_1 and u_2 are given in terms of A_1 and A_2 by (11).

The functions $v_1(q,t)$ and $v_2(q,t)$ are obtained by solving Eqs. (15). The result is

$$v_1 = (g_0u_2 - g_1)/(u_2-u_1), \quad v_2 = (g_1 - g_0u_1)/(u_2-u_1). \tag{68}$$

Finally, the invariant is expressed in terms of u_1, u_2, v_1, and v_2 by (2).

In the next section, we find solutions of (56a) and (56b) that give two-resonance examples of invariants for potentials outside the class given by (38).

IV. PARTICULAR SOLUTIONS WITH TWO RESONANCES

In this section we find solutions of (56a) and (56b) that determine potentials with invariants of two-resonance form. We begin with (56a) and assume a solution of the form

$$J(y,\tau) = \delta(\tau)y + \psi(\tau). \tag{69}$$

Substitution of (69) into (56a) and integration once with respect to τ leads to

$$\delta^2\frac{\partial}{\partial\tau}(\delta y + \psi - b) = (\tau-\tau_0)(\delta y+\psi). \tag{70}$$

In order for (70) to hold, both the linear and constant terms in y have to be equal separately:

$$\delta\frac{\partial\delta}{\partial\tau} = \tau-\tau_0 \tag{71a}$$

and

$$\frac{\partial\psi}{\partial\tau} - \frac{\tau-\tau_0}{\delta^2}\psi = \frac{db}{d\tau}. \tag{71b}$$

Equation (71a) has the general solution

$$\delta^2 = \tau^2 - 2\tau_0\tau - \tau_1. \tag{72}$$

Equation (71b) can be transformed by use of (71a) into

$$\frac{\partial\psi}{\partial\tau} - \frac{1}{\delta}\frac{\partial\delta}{\partial\tau}\psi = \frac{db}{d\tau}. \tag{73}$$

This last equation can be integrated exactly,

$$\psi(\tau) = \delta_0\delta + \delta\int^\tau \frac{1}{\delta}\frac{db}{d\tau'}d\tau', \tag{74}$$

where δ_0 is a constant.

The $J(y,\tau)$ defined by (69), (72), and (74) determines a force that possesses a two-resonance invariant. The result, according to (62), is

$$-\frac{\partial V}{\partial q} = -\frac{V_{-3}}{X^3} - \frac{V_{-2}}{X^2} - \frac{V_{-1}}{X} - V_0 - V_1X, \tag{75}$$

where

$$V_{-3} = 4\epsilon_1^3 b_\tau^2, \tag{76a}$$

$$V_{-2} = 0, \tag{76b}$$

$$V_{-1} = -4\epsilon_1^2 b_{\tau\tau}, \tag{76c}$$

$$V_0 = (\epsilon_2/\epsilon_1)^{\cdot\cdot}, \tag{76d}$$

$$V_1 = \frac{\ddot\epsilon_1}{2\epsilon_1^2} - \frac{3}{4}\frac{\dot\epsilon_1^2}{\epsilon_1^3} + 3(\tau_0^2+\tau_1)\frac{\epsilon_1}{\delta^4}. \tag{76e}$$

Here we make the surprising observation that the force given by (75) and (76) has exactly the same q dependence as that associated with a one-resonance invariant.[2] In fact, as

we shall demonstrate, the class of forces represented by (75) and (76) is the same as the class for which there exists a one-resonance invariant.

Let us denote the potential for the one-resonance case by $W(q,t)$ and write $-\partial W/\partial q$ in the form[2]

$$-\frac{\partial W}{\partial q} = -\frac{W_{-3}}{Q^3} - \frac{W_{-1}}{Q} - W_0 - W_1 Q, \tag{77}$$

where

$$Q = \alpha q + \beta, \tag{78}$$

and

$$W_{-3} = \dot{\gamma}^2/(4\alpha), \tag{79a}$$

$$W_{-1} = (\alpha/2)(\dot{\gamma}/\alpha^2)^{\cdot}, \tag{79b}$$

$$W_0 = (\beta/\alpha)^{\cdot\cdot}, \tag{79c}$$

$$W_1 = (\dot{\alpha}/\alpha^2)^{\cdot}, \tag{79d}$$

where α, β, and γ are arbitrary functions of time.

We now show that the class of forces represented by (75) and (76) and (77)–(79) are the same. We assume that the forces are not simply linear in q. Therefore, $\epsilon_1(t)$ and $\alpha(t)$ are nonzero and the forces are singular in q. For the forces to be equal, their singularities must be located at the same value of q, which implies

$$\epsilon_1(t) = \lambda(t)\alpha(t), \quad \epsilon_2(t) = \lambda(t)\beta(t), \tag{80}$$

where $\lambda(t)$ is an as yet unspecified function of t. An immediate consequence of (80) is

$$X = \lambda Q. \tag{81}$$

Therefore the necessary and sufficient conditions for the forces to be equal are

$$V_{-3} = \lambda^3 W_{-3}, \tag{82a}$$

$$V_{-1} = \lambda W_{-1}, \tag{82b}$$

$$V_0 = W_0, \tag{82c}$$

$$\lambda V_1 = W_1. \tag{82d}$$

By virtue of (80), condition (82c) is satisfied. Define

$$\tilde{\alpha}(\tau) = \alpha(t), \ \tilde{\epsilon}_1(\tau) = \epsilon_1(t), \ \tilde{\gamma}(\tau) = \gamma(t), \ \tilde{\lambda}(\tau) = \lambda(t). \tag{83}$$

Then condition (82b) reduces to

$$b_{\tau\tau} = -\frac{1}{2}\frac{d}{d\tau}\left[\frac{\tilde{\lambda}\tilde{\gamma}_\tau}{\tilde{\alpha}}\right], \tag{84}$$

with the solution

$$b_\tau = -\tfrac{1}{2}\tilde{\lambda}\tilde{\gamma}_\tau/\tilde{\alpha} + c_1, \tag{85}$$

where c_1 is a constant. Condition (82a) is then equivalent to

$$c_1 = 0. \tag{85'}$$

Condition (82d) can be written as

$$\frac{d^2}{d\tau^2}\left(\frac{\tilde{\epsilon}_1^{1/2}}{\tilde{\alpha}}\right) + \frac{3}{4}(\tau_0^2 + \tau_1)\frac{1}{\delta^4}\left(\frac{\tilde{\epsilon}_1^{1/2}}{\tilde{\alpha}}\right) = 0. \tag{86}$$

This equation has a two-parameter family of solutions for each pair (τ_0,τ_1). For any pair (τ_0,τ_1) and any function $\tilde{\epsilon}_1(\tau)$, we can choose a particular $\tilde{\alpha}(\tau)$ to satisfy (86). For that $\tilde{\alpha}(\tau)$, we can then choose any other pair (τ_0,τ_1) and choose another function $\tilde{\epsilon}_1(\tau)$ to again satisfy (86). Thus varying (τ_0,τ_1) does not vary the class of forces that can be represented by (75) and (76). The pair (τ_0,τ_1) can be chosen for convenience and any particular solution of (86) can be used for making the forces (75) and (77) equal. Given either $\tilde{\alpha}(\tau)$ or $\tilde{\epsilon}_1(\tau)$, the solution of (86) then determines $\tilde{\lambda}(\tau)$. Thus we have demonstrated that the class of forces represented by (75) and (76) is the same as the class with a one-resonance invariant.

Our two-resonance invariant for the force given by (75) is obtained by calculating u_1, u_2, v_1, and v_2 from (11), (66), (67), and (68). The result is

$$u_{\substack{1\\2}} = \frac{2\epsilon_1 b_\tau}{X} - \left(\frac{\dot{\epsilon}_1}{2\epsilon_1^2} - \frac{\delta_\tau}{\delta} \mp \frac{1}{\delta}\right)X - \left(\frac{\epsilon_2}{\epsilon_1}\right)^{\cdot}, \tag{87}$$

$$v_{\substack{1\\2}} = (1 \mp \delta_\tau)X/2. \tag{88}$$

The invariant is

$$I(q,p,t) = -\int^t \epsilon_1\, dt' + \frac{X}{2}\left(\frac{1-\delta_\tau}{p-u_+} + \frac{1+\delta_\tau}{p-u_-}\right). \tag{89}$$

This invariant is functionally independent of an $N=1$ invariant for the same force for nearly all values of τ_0 and τ_1. An $N=1$ invariant can be written as[1,2]

$$I_1 = c(t)\,\frac{p - (g_1/g_0 - (g_0/c))}{p - (g_1/g_0)}, \tag{90}$$

where

$$c(t) = -\int^t dt'\,\epsilon_1(t'). \tag{90'}$$

The only class of $N=2$ invariants that can be constructed from I_1 is

$$I_2 = k_1 I_1 + k_2/I_1, \tag{91}$$

where k_1 and k_2 are any constants. The poles for I_2 are

$$u_1 = g_1/g_0 \quad \text{and} \quad u_2 = u_1 - g_0/c. \tag{92}$$

An examination of (87) shows that the poles for the two-resonance invariant defined by (89) satisfy

$$u_2 - u_1 = -g_0/c = -X/c \tag{93}$$

only if $\tau_0 = \tau_1 = 0$. A similar consideration shows that nearly all of the $N=2$ invariants derived in Ref. 2 for this force are also functionally independent of I_1.

New two-resonance examples can be found by applying the Lie theory of extended groups to (56a) and (56b) to identify similarity solutions. Following Bluman and Cole[9] with the help of a MACSYMA computer code written by J. L. Schwarzmeier, the only similarity variable that we found is

$$\chi = \tau y. \tag{94}$$

In terms of this variable, there are similarity solutions of the form

$$J(\tau,y) = F(\chi) = F(\tau y). \tag{95}$$

Using (95) as an ansatz in (56a), we find that $b(\tau)$ has the form

$$b(\tau) = k_0 + k_1 \log \tau, \tag{96}$$

where k_0 and k_1 are constants. Assuming $b(\tau)$ to be of this form and substituting (95) into (56a), we get a second-order ordinary differential equation for $F(\chi)$. That equation

can be integrated once with the result

$$\chi F'^2(\chi F' - k_1)/F = 1/F' + F_0, \tag{97}$$

where F_0 is an integration constant. Any solution of this equation will lead to a potential with a two-resonance invariant. These examples are new.

We now turn to a consideration of (56b). An ansatz like (69) would not lead to anything new because then y could be expressed as a linear function of q, which would imply that the force is in the class defined by (38). Therefore we adopt the similarity ansatz (95).

Substituting (95) into (56b) and integrating once with respect to χ yields

$$\chi^2\left(\frac{dF}{d\chi}\right)^4 - F_0\frac{dF}{d\chi} = 1, \tag{98}$$

where F_0 is an integration constant. Here we make the observation that if the solution of (98) were a homogeneous function of χ then the similarity solution would be separable and the resulting potential would belong to the same category represented by (38). In fact, by assuming $F = \chi^\epsilon$ we find that ϵ would have to satisfy

$$4\epsilon - 2 = \epsilon - 1 = 0, \tag{99}$$

which is impossible. If we took $F_0 = 0$, then the solution of (98) would be

$$F(\chi) = k\sqrt{\chi} + F_1, \tag{100}$$

where k and F_1 are constants. But, from (54), (44), (43), and (22), this solution implies that $f(q,t)$ is a function of a linear function of q; therefore the potential would again fall into the category represented by (38). The conclusion is that we have to find a solution of (98) with $F_0 \neq 0$ in order to obtain a potential that is not already included in (38). This is possible to achieve in parametric form because (98) does not depend explicitly on F (see Ref. 10).

To find the parametric solution of (98), we first solve (98) for χ:

$$\chi = \Phi\left(\frac{dF}{d\chi}\right) = \frac{(1 + F_0\, dF/d\chi)^{1/2}}{(dF/d\chi)^2}. \tag{101}$$

Then

$$F(\chi) = F_1 + \lambda\chi - \int^\lambda \Phi(\sigma)d\sigma, \tag{102a}$$

$$\chi = \Phi(\lambda) = (1 + F_0\lambda)^{1/2}/\lambda^2, \tag{102b}$$

where F_1 is a constant, is a solution in terms of the parameter λ. The parameter λ is given implicitly as a function of χ by (102b). Performing the integral in (102a) we get

$$F(\chi) = F_1 + 2\frac{(1 + F_0\lambda)^{1/2}}{\lambda} + \frac{F_0}{2}\log\left(\frac{(1 + F_0\lambda)^{1/2} + 1}{(1 + F_0\lambda)^{1/2} - 1}\right). \tag{103}$$

The $J(y,\tau)$ defined by (103), (102b), and (95) determines a force that possesses a two-resonance invariant. The result according to (63) is

$$-\frac{\partial V}{\partial q} = \frac{\ddot\rho}{\rho}q + \frac{\rho\ddot\alpha - \alpha\ddot\rho}{\rho} - \frac{1}{\rho^3}\frac{1}{\tau^2}\frac{2F_0(1 + F_0\lambda)^{1/2}}{(4 + 3F_0\lambda)}. \tag{104}$$

Were it not for the factor of $1/\tau^2$, this force would be in the class represented by (38). This is a result of (103), (95), (54), and (44), which show that λ is a function of $\epsilon_2 q + \epsilon_3 = (q - \alpha)/\rho$. Because of the factor $1/\tau^2$, the force given by (104) is a new example of a force with a two-resonance invariant.

Our two-resonance invariant for the force given by (104) is obtained by calculating u_1, u_2, v_1, and v_2 from (11), (66), (67), and (68) using

$$J_y = \lambda\tau, \quad J_\tau = \lambda y, \quad y = (1 + F_0\lambda)^{1/2}/(\tau\lambda^2). \tag{105}$$

Because ϵ_1 is zero, $c(t)$ will be a constant. We incorporate it into the invariant and write

$$I(q,p,t) = v_1/(p - u_1) + v_2/(p - u_2). \tag{106}$$

V. THREE RESONANCES

The three-resonance case can be formulated in analogy to the formulation for two resonances that was presented in Sec. III. Although we have not been able to carry the analysis through in detail, we present the formulation in order to provide a basis for further study.

The following equations for A_1, A_2, and A_3 are obtained by substituting (12) into (3c):

$$\frac{\partial A_1}{\partial t} + \frac{\partial A_2}{\partial t} + \frac{\partial A_3}{\partial t} + \frac{\partial}{\partial q}\left[\frac{1}{2}(A_1^2 + A_2^2 + A_3^2) + A_1A_2 + A_2A_3 + A_1A_3\right] = -\frac{\partial V}{\partial q}, \tag{107a}$$

$$\frac{\partial A_1}{\partial t} + \omega\frac{\partial A_2}{\partial t} + \omega^*\frac{\partial A_3}{\partial t} + \frac{\partial}{\partial q}\left[\frac{1}{2}(A_1^2 + \omega^*A_2^2 + \omega A_3^2) + \omega A_1A_2 + A_2A_3 + \omega^*A_1A_3\right] = -\frac{\partial V}{\partial q}, \tag{107b}$$

$$\frac{\partial A_1}{\partial t} + \omega^*\frac{\partial A_2}{\partial t} + \omega\frac{\partial A_3}{\partial t} + \frac{\partial}{\partial q}\left[\frac{1}{2}(A_1^2 + \omega A_2^2 + \omega^*A_3^2) + \omega^*A_1A_2 + A_2A_3 + \omega A_1A_3\right] = -\frac{\partial V}{\partial q}. \tag{107c}$$

Subtract (107a) from (107b) and (107c) and add (107a) through (107c) to get

$$\epsilon\frac{\partial A_2}{\partial t} + \epsilon^*\frac{\partial A_3}{\partial t} + \frac{\partial}{\partial q}\left[\frac{1}{2}(\epsilon^*A_2^2 + \epsilon A_3^2) + \epsilon A_1A_2 + \epsilon^*A_1A_3\right] = 0, \tag{108a}$$

$$\epsilon^* \frac{\partial A_2}{\partial t} + \epsilon \frac{\partial A_3}{\partial t} + \frac{\partial}{\partial q}\left[\frac{1}{2}(\epsilon A_2^2 + \epsilon^* A_3^2) + \epsilon^* A_1 A_2 + \epsilon A_1 A_3\right] = 0, \tag{108b}$$

$$\frac{\partial A_1}{\partial t} + \frac{\partial}{\partial q}\left[\frac{1}{2}A_1^2 + A_2 A_3\right] = -\frac{\partial V}{\partial q}, \tag{109}$$

where

$$\epsilon = \omega - 1. \tag{110}$$

In analogy with the treatment for two resonances, we try to satisfy (108a) and (108b) in a general way and we take (109) as a formula for calculating $\partial V/\partial q$ in terms of A_1, A_2, and A_3.

The difference of (108a) and (108b) is equivalent to

$$\frac{\partial}{\partial t}(A_2 - A_3) + \frac{\partial}{\partial q}\left\{(A_2 - A_3)\left[A_1 - \frac{1}{2}(A_2 + A_3)\right]\right\} = 0. \tag{111}$$

This is of the same form as (19). We again define a variable x and functions $f(q,t)$ and $r(x,t)$. The function $f(q,t)$ is some particular solution of

$$\frac{\partial f}{\partial t} + \left[A_1 - \frac{1}{2}(A_2 + A_3)\right]\frac{\partial f}{\partial q} = 0. \tag{112}$$

The variable x is defined by

$$x = f(q,t) \tag{113}$$

and the inverse is

$$q = r(x,t). \tag{114}$$

In analogy with (26) and (27) we have

$$A_1 - \frac{1}{2}(A_2 + A_3) = \frac{\partial r}{\partial t}, \tag{115}$$

$$A_2 - A_3 = \frac{W(x)}{\partial r/\partial x}, \tag{116}$$

where $W(x)$ is arbitrary. Thus $A_2(q,t)$ and $A_3(q,t)$ can be written in terms of $r(x,t)$ and $W(x)$ as

$$A_2 = A_1 - \frac{\partial r}{\partial t} + \frac{1}{2}\frac{W(x)}{\partial r/\partial x}, \tag{117a}$$

$$A_3 = A_1 - \frac{\partial r}{\partial t} - \frac{1}{2}\frac{W(x)}{\partial r/\partial x}. \tag{117b}$$

The sum of (108a) and (108b) yields

$$\frac{\partial}{\partial t}(A_2 + A_3) + \frac{\partial}{\partial q}\left[\frac{1}{2}(A_2^2 + A_3^2) + A_1(A_2 + A_3)\right] = 0, \tag{118}$$

which, by using (117), can be written in terms of x and t as

$$2\frac{\partial}{\partial t}(B_1 - r_t) - 2\frac{r_t}{r_x}\frac{\partial}{\partial x}(B_1 - r_t) + \frac{1}{r_x}\frac{\partial}{\partial x}\left[(B_1 - r_t)^2 + \frac{1}{4}\left(\frac{W(x)}{r_x}\right)^2 + 2B_1(B_1 - r_t)\right] = 0, \tag{119}$$

where

$$B_1(x,t) = A_1(q,t). \tag{120}$$

We have not succeeded in solving (119) and, therefore, cannot proceed to study (17) with this approach. We attempted to get a class of solutions by viewing (119) as an ordinary differential equation for $W^2(x)$ and requiring that the ratios of the coefficients be independent of t. That analysis led to the conclusion that the potential would have to be restricted to the case of a driven linear oscillator. Since that problem is well understood, it was unnecessary to study (17) in this case.

VI. DISCUSSION

We have succeeded in establishing a fairly clear picture of two-resonance invariants and the potentials which admit them. Our examples extend the previously known examples of two-resonance invariants in an interesting way. The case of three resonances remains poorly developed.

An attractive avenue for further research, which holds the prospect for progress when there are more than two resonances, is to return to the general considerations presented by Lewis and Leach in Sec. VI of Ref. 2. If our results for the two-resonance case were understood in their context, then it would seem likely that the case of three or more resonances could be tackled more successfully.

It would also be valuable to investigate the consequences of removing the restriction of simple momentum poles in the original ansatz for the form of an invariant.

ACKNOWLEDGMENTS

We thank James L. Schwarzmeier for his assistance with MACSYMA in finding the determining equations for the treatment of (56a) and (56b) with the Lie theory of extended groups. We also thank Paul R. Stein for his advice concerning the roots of quartic algebraic equations.

[1] J. Goedert and H. R. Lewis, J. Math. Phys. **28**, 728 (1987).
[2] H. R. Lewis and P. G. L. Leach, Ann. Phys. (NY) **164**, 47 (1985).
[3] W. Sarlet (private communication).
[4] P. G. L. Leach, H. R. Lewis, and W. Sarlet, J. Math. Phys. **25**, 486 (1984).
[5] H. R. Lewis and J. Goedert, in preparation.
[6] H. R. Lewis and P. G. L. Leach, J. Math. Phys. **23**, 2371 (1982).
[7] W. Sarlet and J. R. Ray, J. Math. Phys. **22**, 2504 (1981).
[8] H. W. Turnbull, *Theory of Equations* (Interscience, New York, 1952).
[9] G. W. Bluman and J. D. Cole, *Similarity Methods for Differential Equations* (Springer, New York, 1974).
[10] E. L. Ince, *Integration of Ordinary Differential Equations* (Interscience, New York, 1952).

On unusual statistics and quantum point vortices

Gerald A. Goldin
Departments of Mathematics and Physics, Rutgers University, New Brunswick, New Jersey 08903
Ralph Menikoff and David H. Sharp
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 1 July 1986, accepted for publication 29 October 1986)

In an ideal two-dimensional incompressible fluid, θ statistics for a pair of identical quantum point vortices is not well defined.

I. INTRODUCTION

The classical equations of motion for point vortices in an effectively two-dimensional incompressible fluid (e.g., a thin film) are well known to form a Hamiltonian system. The relative motion of two vortices of equal circulation is described by the equation $dx/dt = \{x,H\} = \partial H/\partial y$ and $dy/dt = \{y,H\} = -\partial H/\partial x$, where, after subtracting the individual vortex self-energies, $H = -\pi^{-1}\ln(x^2+y^2)$, and where the Cartesian coordinates $x = x_1 - x_2$ and $y = y_1 - y_2$ are canonically conjugate; i.e., $\{x,y\} = 1$. For simplicity we have set the circulation $\kappa = 2$, a scale factor $a = 1$, and $\rho\delta = 1$ (where ρ is the density and δ the thickness of the fluid film). Using the fact that the phase space for this system has a nontrivial topology (that of the plane $\mathbb{R}^2$ without the origin), it was recently argued in an interesting paper[1] that the quantum vortices obtained by quantizing these classical equations obey the θ statistics introduced earlier for two-dimensional particle systems by several physicists,[2] with either $\theta = \pi/2$ or $3\pi/2$. This conclusion is based on a change of variables between Cartesian and polar coordinates which appears to allow the introduction of a polar angle ϕ and an associated winding number (of one vortex about the other), in terms of which the θ statistics can be defined. The allowed values of θ are determined from the eigenvalues of the angular momentum operator. Subsequently the conclusion was generalized to $\theta(N) = \pi/N$ or $\pi/N + \pi$ for the case of N identical point vortices.[3]

In this paper, we conclude that it is in fact not possible to assign θ statistics consistently to identical point vortices in an incompressible fluid within the framework of quantizing the above classical equations. Mathematically, the statistics of a quantum system depends on the topology of its configuration space, not its phase space.[4] In this sense statistics is kinematical rather than dynamical. We shall see below that although the classical phase space for this problem is multiply connected, *no multiply connected quantum configuration space can be introduced that is consistent with the interpretation of the angular momentum operator as the generator of physical rotations.* In particular, a mathematical argument taking account of operator domains demonstrates that this cannot be accomplished by transforming to polar coordinates as in Ref. 1. We also consider other proposed phase operators,[5] and examine why they cannot be used in this context to define θ statistics.

The physical reason underlying our result is the following. Because x and y are canonically conjugate, the problem is effectively one dimensional. The uncertainty principle implies that the winding number can be measured only for trajectories where the vortices are well separated. If the vortices are sufficiently close (so that the region of uncertainty includes the origin), one cannot observe whether one vortex has passed "above" or "below" the other. Consequently θ statistics cannot be defined in terms of the winding number.

II. QUANTIZATION IN CARTESIAN AND POLAR COORDINATES

Let us see why in the context of the present model the transformation from Cartesian to polar coordinates cannot be implemented quantum mechanically. Classically we have $x = r\cos\phi$ and $y = r\sin\phi$, with $\{r^2/2,\phi\} = 1$ and $H = -\pi^{-1}\ln r^2$. To construct an analogous quantum-mechanical transformation, let the operators $x, y = (1/i)d/dx$, and $H = -\pi^{-1}\ln(x^2+y^2)$ be defined on the "Cartesian" Hilbert space $\mathscr{H}_c = L^2(R)$, where x and y satisfy $[x,y] = i$ (setting $\hbar = 1$). The spectrum of $(x^2+y^2)/2$ is the usual "harmonic oscillator" spectrum $\{n+\frac{1}{2}; n = 0,1,2,\ldots\}$. Likewise let the operators ϕ and $r^2/2 = -(1/i)d/d\phi$ be defined on the "polar" Hilbert space $\mathscr{H}_p = L^2(S^1)$, where the commutator $[r^2/2,\phi] = i$ formally holds. Next consider the domains on which these operators are defined. Since ϕ is bounded, it is defined everywhere; $D(\phi) = \mathscr{H}_p$. There is, however, a one-parameter family of distinct self-adjoint operators which can be defined from the differential operator $-(1/i)d/d\phi$, on the domains D_θ of suitably smooth functions $\Phi(\phi)$ such that $\Phi(2\pi) = \exp(-i\theta)\Phi(0)$. Defining $r^2/2$ on D_θ, its eigenfunctions are $\exp[-i(n+\theta/2\pi)\phi]$ with eigenvalues $n+\theta/2\pi$ for $n = 0, \pm 1, \pm 2, \ldots$, which are unbounded above and below.[6]

Now a symmetric form of the coordinate transformation is needed to preserve the self-adjointness of the operators, which led [Ref. 1, Eqs. (18) and (19)] to the definition

$$\begin{aligned}\tilde{x} &= [\exp(i\phi/2)r\exp(i\phi/2) \\ &\quad + \exp(-i\phi/2)r\exp(-i\phi/2)]/2, \\ \tilde{y} &= [\exp(i\phi/2)r\exp(i\phi/2) \\ &\quad - \exp(-i\phi/2)r\exp(-i\phi/2]/2i.\end{aligned} \tag{1}$$

We have introduced tildes in these equations to label operators that act in $\mathscr{H}_p$ rather than $\mathscr{H}_c$. For the physics to be independent of the choice of coordinate system, there must be an isometry U: $\mathscr{H}_c \to \mathscr{H}_p$ such that $x = U^{-1}\tilde{x}U$ and $y = U^{-1}\tilde{y}U$. In order for the transformation of Eq. (1) to be defined, the operator r must exist. But r can be defined only on the subspace of $\mathscr{H}_p$ where the spectrum of r^2 is

 0022-2488/87/030744-03$02.50

positive. Consequently the best one can do in setting up a correspondence between Cartesian and polar coordinates is to define U to map $\mathscr{H}_c$ onto a subspace $\mathscr{H}_p^+$ of $\mathscr{H}_p$.

To determine the appropriate subspace $\mathscr{H}_p^+$ and construct U, we consider the angular momentum operator L_z in $\mathscr{H}_c$, defined to be the infinitesimal generator of physical rotations. Thus

$$R(\alpha)^{-1}xR(\alpha) = x\cos\alpha - y\sin\alpha$$

and

$$R(\alpha)^{-1}yR(\alpha) = y\cos\alpha + x\sin\alpha,$$

where $R(\alpha) = \exp(-i\alpha L_z)$, leading to the formula $L_z = -(x^2+y^2)/2 + c$, where c is a constant. In Ref. 1 the choice $c = 0$ is made. In any case, the spectrum of L_z is $\{-(n+\frac{1}{2}) + c;\ n = 0,1,2,\ldots\}$. Let

$$\tilde{L}_z = -(\tilde{x}^2+\tilde{y}^2)/2 + c$$

be the corresponding operator in $\mathscr{H}_p$. From Eqs. (1) we have

$$\tilde{L}_z = -[\exp(i\phi/2)r^2\exp(-i\phi/2) + \exp(-i\phi/2)r^2\exp(i\phi/2)]/4 + c. \quad (2)$$

For the negative eigenvalues of $\tilde{L}_z - c$ to coincide with the spectrum of $L_z - c$, it is *necessary* in Eq. (2) for the domain of r^2 to be $D_{\theta=0}$ and the domain of $\tilde{L}_z$ to be $D_{\theta=\pi}$. Then we must choose $\mathscr{H}_p^+$ to be the closed subspace of $\mathscr{H}_p$ spanned by $\{\exp[-i(n+\frac{1}{2})\phi];\ n = 0,1,2,\ldots\}$. Now Eq. (2) makes sense in $\mathscr{H}_p$, while $\tilde{L}_z$ maps vectors in $\mathscr{H}_p^+ \cap D_{\theta=\pi}$ to $\mathscr{H}_p^+$. The commutation relation $[r^2/2,\phi] = i$ is satisfied on the dense domain of vectors $\Phi\in D_{\theta=0}$ such that $\phi\Phi\in D_{\theta=0}$ (which is not, however, a very large domain—its elements are wave functions that vanish at $\phi = 0$ and $\phi = 2\pi$, and thus belong to D_θ for all θ). Moreover in Eqs. (1), $\tilde{x}$ and $\tilde{y}$ map vectors in $\mathscr{H}_p^+ \cap D_{\theta=\pi}$ to $\mathscr{H}_p^+$, and the transformation makes sense in $\mathscr{H}_p$ with

$$r\colon \exp(-in\phi) \to (2n)^{1/2}\exp(-in\phi) \quad \text{for } n \geqslant 0.$$

We emphasize, however, that the domain of $\tilde{L}_z$ is not the same as the domain of r^2. Thus the classical relation $\tilde{x}^2+\tilde{y}^2 = r^2$ is not valid as an operator equation, nor can $\tilde{x}^2+\tilde{y}$ and r^2 be made equal by adding a constant, although they are both given by the same differential operator $-(2/i)d/d\phi$ on their respective domains.

Nevertheless, we can use the correspondence between eigenfunctions of L_z and $\tilde{L}_z$ to define U. The eigenfunctions of L_z are normalized Hermite functions $\Psi_n(x)$, $n = 0,1,2,\ldots$. The eigenfunctions with corresponding eigenvalues of $\tilde{L}_z$ in $\mathscr{H}_p^+$ are

$$\Phi_n(\phi) = (2\pi)^{-1/2}\exp[-i(n+\tfrac{1}{2})\phi].$$

The isometry U is now generated by linearity and continuity from $U\Psi_n = \Phi_n$. Thus the image of U is $\mathscr{H}_p^+$, and $UL_zU^{-1} = \tilde{L}_z$ on the domain of $\tilde{L}_z$ in $\mathscr{H}_p^+$. A simple computation verifies that Φ_n satisfies the same recursion relation for $\tilde{x}$ as Ψ_n does for x, namely

$$\tilde{x}\Phi_n = [(n+1)^{1/2}\Phi_{n+1} + n^{1/2}\Phi_{n-1}]/2.$$

Hence $UxU^{-1} = \tilde{x}$ on the domain of $\tilde{x}$ in $\mathscr{H}_p^+$, and by a similar argument $UyU^{-1} = \tilde{y}$. This construction of U parallels that used in discussing action-angle variables.[7]

Since the vortices are identical, the wave functions that describe them must undergo a fixed phase change when rotated by π (not just by 2π). The equation $\exp(-i\pi L_z)\Psi_n = \exp(i\theta)\Psi_n$ in $\mathscr{H}_c$, or equivalently $\exp(-i\pi\tilde{L}_z)\Phi_n = \exp(i\theta)\Phi_n$ in $\mathscr{H}_p^+$, gives $\exp[i\pi(n+1/2-c)] = \exp(i\theta)$. Thus either $n = 0,2,4,\ldots$ with $\theta = (\frac{1}{2}-c)\pi$, or $n = 1,3,5,\ldots$, with $\theta = (\frac{3}{2}-c)\pi$.

But it is now apparent that U is incompatible with the operator ϕ. For any nonzero Φ in $\mathscr{H}_p^+$, $\phi\Phi$ is not in $\mathscr{H}_p^+$; thus ϕ is not an operator in $\mathscr{H}_p^+$. This means the polar angle operator on $\mathscr{H}_c$, $U^{-1}\phi U$, does not exist because the range of ϕU meets the domain of U^{-1} only at $\Phi = 0$. In short it is not possible to choose the domains of self-adjoint operators x,y,r^2, and ϕ so that the Cartesian and polar coordinate descriptions are equivalent, with Eqs. (1) and the commutation relations $[x,y] = i$ and $[r^2/2,\phi] = i$ simultaneously satisfied on appropriate subdomains. *Thus the polar angle ϕ is not an observable, and cannot be used to measure a winding number for the quantum vortices.* While one can talk mathematically about a phase change of $\exp(i\theta)$ in a wave function under the action of the rotation operator $\exp(-i\pi L_z)$, one cannot perform a physical measurement of the phase. As a result the constant c in L_z is physically indeterminate, and therefore arbitrary.

For the case of a coherent state describing widely separated vortices, such as an asymptotic state in a semiclassical approximation, a winding number could be defined by localizing in x and y, and defining the phase angle as $\tan^{-1}(\langle y\rangle/\langle x\rangle)$. But because high angular momentum values enter, a small uncertainty in the angle results in a large uncertainty in the phase shift. Thus c remains indeterminate even in the semiclassical approximation.

III. OTHER CANDIDATES FOR THE PHASE OBSERVABLE

We have seen in detail why the polar angle ϕ is not an observable. Let us now consider other candidates for the phase observable. The operator $\phi' = U^{-1}P\phi U$, where P denotes orthogonal projection onto $\mathscr{H}_p^+$, defines a bounded self-adjoint operator (and thus an observable) in $\mathscr{H}_c$. It turns out that ϕ' is equivalent to the phase observables defined to be canonically conjugate to the number operator $N_{\rm op}$ in Ref. 5 (with L_z equivalent to $-N_{\rm op} - \frac{1}{2} + c$). Thus L_z and ϕ' are canonically conjugate on a certain domain; however, the domain is not sufficiently large to permit exponentiation of the commutation relation. Since ϕ' is self-adjoint, one can transform to a Hilbert space $\mathscr{H}' = L^2(S^1)$ on which ϕ' is diagonal; i.e., $\phi'\Phi'(\omega) = \omega\Phi'(\omega)$ for all $\Phi'\in\mathscr{H}'$. To describe the effect of a physical rotation on ϕ', define $\phi'_\alpha = \exp(i\alpha L_z)\phi'\exp(-i\alpha L_z)$. For statistics to be associated with measurement of the winding of vortices, it would have to be the case that

$$\phi'_\alpha\Phi'(\omega) = [(\omega+\alpha)\bmod 2\pi]\Phi'(\omega),$$

whence ϕ' and ϕ'_α would commute. But it can be shown explicitly that $[\phi',\phi'_\alpha] \neq 0$. Indeed, for ϕ'_α to act as indicated would require L_z to be an infinitesimal generator of rotation

in ω space. Then L_z would necessarily be a self-adjoint operator $(1/i)d/d\omega$ on a domain D_θ of suitably smooth functions $\Phi(\omega)$ such that $\Phi'(2\pi) = \exp(-i\theta)\Phi'(0)$, with eigenvalue spectrum unbounded above and below—and could not be the physical rotation operator for our problem.

IV. COMMENTS ON RELATED MODELS

A. Vortex dipoles

We have seen that θ statistics is ill-defined for identical quantum point vortices. The same argument applies to a pair of distinguishable vortices, as long as the relative coordinates are canonically conjugate, which is true unless the vortices have equal and opposite circulations. In that case the x and y relative coordinates commute.[8] This occurs because the circulation enters both the Hamiltonian and the commutation relations. When the vortices are identical, x and y commute with $X = (x_1 + x_2)/2$ and $Y = (y_1 + y_2)/2$, respectively, i.e., $[x,X] = [y,Y] = 0$, while $[x,y] = 4[X,Y] = i$. When the circulations are opposite, x and y commute with each other and are conjugate to X and Y, respectively. Thus x and y can be observed simultaneously by giving up information about X and Y, and the *configuration space*, $\mathbb{R}^2$ without the origin, is multiply connected. When one vortex circles the other by 2π, the wave function can be multiplied by an arbitrary *observable* phase $\exp(i\theta)$, with any value of θ between 0 and 2π kinematically allowed. Such a "vortex dipole" system is of interest because of its relation to the Kosterlitz–Thouless model.[9] As has been remarked elsewhere, in two dimensions "unusual statistics" makes sense even for particles which are not identical.[4]

B. Microscopic model for a compressible superfluid

Haldane and Wu,[10] starting from a particular microscopic model for a *compressible* superfluid, have also argued that θ statistics for identical quantum point vortices is ill defined. While we agree with their conclusion, we stress that the results established here are rigorous, do not depend on a particular microscopic model, and follow directly from the assumptions of the model of Ref. 1, irrespective of its intended domain of application.

C. Quantum hydrodynamics

The authors have proposed that theories of quantum hydrodynamics based on *infinitely many* degrees of freedom can be obtained from the classical configuration space.[11] For a two-dimensional, incompressible fluid, this space can be identified[12] with the group $\mathrm{Diff}_v(R^2)$ of all volume-preserving diffeomorphisms of $\mathbb{R}^2$. Quantization of such theories leads to the study of unitary representations of $\mathrm{Diff}_v(\mathbb{R}^2)$. This in fact first led us to the conclusions presented here. It turns out that the idealization of pure point quantum vortices is inconsistent in this picture. Quantum vortex systems that do exist (e.g., extended objects) possess degrees of freedom which have the effect of permitting the relative x and y coordinates to be simultaneously observable! Thus we anticipate that in effectively two-dimensional vortex systems, statistics can be observed. The determination of Bose, Fermi, or θ statistics is then an experimental question. Quantum vortices from this point of view are discussed by the authors in Ref. 13.

ACKNOWLEDGMENTS

The authors thank R. Y. Chiao for stimulating and helpful discussions, and for thoughtful comments on an earlier draft of this paper. One of us (G. G.) acknowledges hospitality from the Department of Mathematics, University of California (Berkeley) and the Theoretical Division, Los Alamos National Laboratory.

We are grateful to the U. S. Department of Energy for financial support.

[1] R. Y. Chiao, A. Hansen, and A. A. Moulthrop, Phys. Rev. Lett. **54**, 1339 (1985).
[2] J. M. Leinaas and J. Myrheim, Nuovo Cimento B **37**, 1 (1977); G. A. Goldin, R. Menikoff, and D. H. Sharp, J. Math. Phys. **22**, 1664 (1981); F. Wilczek, Phys. Rev. Lett. **49**, 957 (1982); G. A. Goldin and D. H. Sharp, Phys. Rev. D **28**, 830 (1983).
[3] A. Hansen, A. A. Moulthrop, and R. Y. Chiao, Phys. Rev. Lett. **55**, 1431 (1985).
[4] G. A. Goldin, R. Menikoff, and D. H. Sharp, J. Math. Phys. **21**, 650 (1980); Phys. Rev. Lett. **54**, 603 (1985).
[5] J. C. Garrison and J. Wong, J. Math. Phys. **11**, 2242 (1970); A. Galindo, Lett. Math. Phys. **8**, 495 (1984).
[6] See, e.g., F. Riesz and B. Sz.-Nagy, *Functional Analysis* (Ungar, New York, 1955), pp. 308–313.
[7] R. G. Newton, Ann. Phys. (NY) **124**, 327 (1980).
[8] J. L. McCauley, Jr., J. Phys. A **12**, 1999 (1979).
[9] J. M. Kosterlitz and D. V. Thouless, J. Phys. C **6**, 1181 (1973).
[10] F. D. M. Haldane and Y. S. Wu, Phys. Rev. Lett. **55**, 2887 (1985).
[11] G. A. Goldin, R. Menikoff, and D. H. Sharp, Phys. Rev. Lett. **51**, 2246 (1983); G. A. Goldin, Contemp. Math. **28**, 189 (1984).
[12] J. Marsden and A. Weinstein, Physica D **7**, 305 (1983).
[13] G. A. Goldin, R. Menikoff, and D. H. Sharp, "Diffeomorphism groups, coadjoint orbits, and the quantization of classical fluids" and "Quantizied vortex elements in incompressible fluids," in *Proceedings of the First International Conference on the Physics of Phase Space*, edited by Y. S. Kim and W. W. Zachary (Springer, Berlin, in press).

Erratum: Spherically symmetric perfect fluid solutions in isotropic coordinates [J. Math. Phys. 27, 1363 (1986)]

Joseph Hajj-Boutros
Physics Department, Lebanese University, Mansourieh El-Metn, P.O. Box 72, Lebanon

(Received 5 May 1986; accepted for publication 7 May 1986)

On p. 1364, the left-hand side of Eq. (3.2) should read $8\pi p$.

 0022-2488/87/030747-01$02.50