Mixed supertableaux of the superunitary groups. II. SU($n|m$)

F. Delduc and M. Gourdin

Laboratoire de Physique Théorique et Hautes Energies,[a),b)] *Paris, France*

(Received 17 July 1984; accepted for publication 14 December 1984)

The Young supertableaux of the superunitary groups are studied. We give the conditions on the size of legal supertableaux, the type of atypicity for tensor representations which are irreducible, making the link between the Kac–Dynkin parameters and those of the supertableau. A discussion is given on nonfully reducible representations and generalized atypical supertableaux.

I. GENERAL CONSIDERATIONS

The Young tableaux play a key role in the study of the irreducible representations of the classical Lie groups. They offer a simple graphical description of the tensor representations of these groups and, in particular, they allow a quick reduction of the product of two irreducible representations into its irreducible components. This aspect of the Young tableaux theory is particularly appreciated by physicists.

The extension of the notion of Young tableaux to Z_2 graded Lie algebras is due to Balantekin and Bars.[1] These authors have defined the Young supertableaux for the superunitary and the orthosymplectic groups and they have given very elegant and compact formulas for computing characters, dimensions, and superdimensions. The important case of the superunitary groups SU($n|m$) has been considered by Bars, Morel, and Ruegg[2] who built the bridge between the Kac–Dynkin parameters of a representation[3,4] and the supertableau parameters. In particular a complete description of the fully covariant or fully contravariant supertableaux of SU($n|m$) can be found in Ref. 2. However, we know that mixed supertableaux with covariant boxes are present and the aim of our study is a complete analysis of these supertableaux. A first set of results concerning the supergroups SU($n|1$) or SU($1|m$) has already been published and we shall refer to this publication[5] as I.

Before going into details let us make a comparison between the SU(n) and SU($n|m$) cases on three specific points where the differences are nontrivial.

(1) For the unitary group SU(n) the covariant tensors describe all the finite-dimensional irreducible representations. Because of the unimodular character of the SU(n) transformations purely contravariant tensors and mixed tensors are equivalent to purely covariant ones.

For the superunitary groups SU($n|m$) such an equivalence between covariant and contravariant indices is no longer true. In particular, purely covariant and purely contravariant supertensors are unrelated. In the description of irreducible representations of SU($n|m$) we must use the two types of supertensors and, in addition, supertensors with both types of indices. Moreover the irreducible representations of the superunitary groups SU($n|m$) cannot all be described with supertensors when $n\neq 1$ and $m\neq 1$.

(2) The Young tableaux of SU(n) have, at most, n rows of arbitrary length, the highest rank of a fully skew symmetric tensor in an n-dimensional vector space being precisely n.

The same argument cannot be used for the Young supertableaux of SU($n|m$) because of the existence of bosonic and fermionic indices due to the grading of the space of representation. In fact a legal supertableau of SU($n|m$) has, at most, n rows and m columns of arbitrary length and the precise mathematical condition is given in Sec. IV.

(3) For the unitary group SU(n) all the finite-dimensional representations are fully reducible and each Young tableau is associated to one irreducible representation of SU(n).

For the superunitary group SU($n|m$) the same property extends to the typical representations, not to the atypical ones.[3] There exist nonfully reducible atypical representations and some of them can be associated to generalized atypical supertableaux which are collections of atypical supertableaux which cannot be divided.

The main original results of our study are the following: (1) a definition of the degeneracy of the atypicity of an irreducible representation of SU($n|m$) and of its minimal realization (Sec. II); (2) a definition of the legality of a Young supertableau of SU($n|m$) and a measure of its size (Sec. III); (3) a computation of the Kac–Dynkin parameters of a supertableau of SU($n|m$) (Sec. IV); (4) a relation between the size of a legal supertableau and the degeneracy of the atypicity of its highest weight (Sec. V); (5) a characterization of which irreducible representation of SU($n|m$) can be represented with a legal supertableau (Sec. VI); and (6) a description of the structure of the generalized atypical supertableaux (Sec. VII).

Our results contain, in particular, those obtained by Bars, Morel, and Ruegg for purely covariant and for purely contravariant supertableaux of SU($n|m$).[2]

II. KAC–DYNKIN PARAMETERS FOR SUPERUNITARY GROUPS

The graded Lie algebra of the superunitary group SU($n|m$) with $n\neq m$ is simple and its Bose sector is the ordinary Lie algebra of the group $G_0\equiv SU(n)\otimes SU(m)\otimes U(1)$. We have $n+m-1$ Cartan operators H_i interpreted as follows:

$H_1,\ldots,H_{n-1}$, $\quad n-1$ Cartan generators for SU(n);

$H_{n+1},\ldots,H_{n+m-1}$, $\quad m-1$ Cartan generators for SU(m).

The infinitesimal generator Q of the U(1) part of G_0 is defined by

---

[a)] Laboratoire Associé au C.N.R.S. n° 280.

[b)] Postal addresss: L.P.T.H.E.-Université P. et M. Curie, Tour 16, 1er étage, 4 Place Jussieu, 75230 Paris Cedex 05, France.

 0022-2488/85/081865-15$02.50

$$Q = \frac{1}{n}\sum_{1}^{n-1} sH_s + H_n - \frac{1}{m}\sum_{1}^{m-1}(m-t)H_{n+t}. \tag{1}$$

An irreducible representation of SU($n|m$) is determined by $n+m-1$ Kac–Dynkin parameters a_l which are the eigenvalues of the Cartan operators H_l for the highest weight Λ of the representation. This highest weight corresponds to the smallest eigenvalue of Q if $n>m$ and to the largest one if $n<m$ (see Refs. 3 and 4). The following notation will be used for an irreducible representation of SU($n|m$):

$$\Lambda \Rightarrow \{a_1, a_2,..., a_{n-1}|a_n|a_{n+1},..., a_{n+m-1}\}.$$

The Kac–Dynkin parameters except a_n are non-negative integers and *a priori* a_n is any real number. In the supertableau approach considered here a_n is restricted to be an algebraic integer.[2] The eigenvalue of the operator Q for the highest weight Λ is given from Eq. (1) by

$$Q_{\rm HW} = \frac{1}{n}\sum_{1}^{n-1} sa_s + a_n - \frac{1}{m}\sum_{1}^{m-1}(m-t)a_{n+t}. \tag{2}$$

The highest weight Λ of an irreducible representation of SU($n|m$) is also the highest weight of an irreducible representation of the ordinary Lie group G_0. As a consequence Λ can be described by pairs of ordinary Young tableaux (X, Y) and the explicit construction of these pairs is given in Appendix A.

An irreducible representation of SU($n|m$) is atypical when a_n takes one of the values A_{jk} defined by[2,3]

$$A_{jk} = \Big(k + \sum_{1}^{k} a_{n+r}\Big) - \Big(j + \sum_{1}^{j} a_{n-r}\Big), \tag{3}$$

with $0 \leqslant j \leqslant n-1$ and $0 \leqslant k \leqslant m-1$. For all the other values of a_n the representation is typical.

Because of the positivity of the SU(n) and SU(m) Kac–Dynkin parameters the discrete spectrum of the atypical values A_{jk} is located within the following bounds:

$$-\sum_{1}^{n-1}(1+a_{n-r}) \leqslant A_{jk} \leqslant +\sum_{1}^{m-1}(1+a_{n+r}). \tag{4}$$

For a given set of SU(n) and SU(m) Kac–Dynkin parameters we construct the $n\times m$ tableau of the atypical values A_{jk} defined in Eq. (3). The degree of degeneracy of the parameter a_n, $\delta(a_n)$, is the number of times a_n enters in this tableau. For a typical representation $\delta(a_n)=0$ and for an atypical one $\delta(a_n)\geqslant 1$. Using again the positivity of the SU(n) and SU(m) Kac–Dynkin parameters we find an upper bound for the function $\delta(a_n)$. Defining L as the minimum of the two integers n and m,

$$L = \min[n, m], \tag{5}$$

we easily get

$$0 \leqslant \delta(a_n) \leqslant L. \tag{6}$$

Let us consider an atypical irreducible representation with $\delta(a_n)=2$. By definition we have two pairs of indices (j, k), (j', k'), and two pairs only such that

$$A_{jk} = A_{j'k'}. \tag{7}$$

Again by positivity if $j'>j$ then $k'>k$ and the equality (7) implies a linear relation between SU(n) and SU(m) Kac–Dynkin parameters,

$$(j'-j-1) + \sum_{j+1}^{j'} a_{n-r} = (k'-k-1) + \sum_{k+1}^{k'} a_{n+r}. \tag{8}$$

The particular solutions of this equality

$$\sum_{j+1}^{j'} a_{n-r} = k'-k-1, \quad \sum_{k+1}^{k'} a_{n+r} = j'-j-1 \tag{9}$$

are called, by definition, minimal realizations of the degree of degeneracy $\delta(a_n)=2$ exhibited by Eq. (7). The solution of Eq. (9) is, in general, not unique; but we must pay attention, in solving Eqs. (9), to avoid the appearance of unwanted degeneracies.

When $\delta(a_n)>2$ we have $\delta-1$ linear relations of the type (8) and these relations are independent because they involve different SU(n) and SU(m) Kac–Dynkin parameters. As a consequence a degree of degeneracy $\delta>2$ can equivalently be viewed as a set of $\delta-1$ degeneracies of order 2. In particular a minimal realization of a degeneracy $\delta>2$ is nothing but a simultaneous minimal realization of $\delta-1$ degeneracies of degree 2.

The reduction of an irreducible representation of SU($n|m$) with respect to the subgroup G_0 is the description of the SU(n) $\otimes$ SU(m) content of the various levels associated to the possible eigenvalues of the U(1) operator Q. These levels can be labeled by a non-negative integer ν,

$$O_\nu = Q_{\rm HW} + \nu[1/m - 1/n], \quad \nu = 0,1,...,\nu_M. \tag{10}$$

the maximal value ν_M is bounded by the product nm and for typical representations we have $\nu_M = nm$.

For the two extreme levels $\nu=0$ and $\nu=\nu_M$ we have only one SU(n) $\otimes$ SU(m) component defining, respectively, the highest weight and the lowest weight of the considered representation. For typical representations the highest weight and the lowest weight have the same SU(n) $\otimes$ SU(m) component and the eigenvalues of the operator Q are related by

$$Q_{\rm LW} = Q_{\rm HW} + n - m. \tag{11}$$

To our knowledge for an atypical representation the explicit relation between the Kac–Dynkin parameters of the highest and lowest weights and the formula giving the dimension of the representation in terms of Kac–Dynkin parameters are not known.

Two irreducible representations R and $\overline{R}$ of SU($n|m$) are contragradient representations if their reduction with respect to the subgroup G_0 produces contragradient SU(n) $\otimes$ SU(m) components with opposite eigenvalues of the operator Q. In particular the highest weight Λ of $\overline{R}$ is the contragradient weight of the lowest weight Λ_0 of R. If R is typical then $\overline{R}$ is also typical with a highest weight $\overline{\Lambda}$ related to Λ by

$$\overline{\Lambda} \Rightarrow \{a_{n-1}, a_{n-2},..., a_1|\bar{a}_n|a_{n+m-1},..., a_{n+1}\}$$

with

$$\bar{a}_n = -a_n + \sum_{1}^{m-1}(1+a_{n+r}) - \sum_{1}^{n-1}(1+a_{n-r}).$$

This last equality is just a consequence of Eqs. (2) and (11). If R is atypical then $\overline{R}$ is also atypical and the following properties will become obvious in the supertableau approach of the coming sections.

(i) The degree of degeneracy of the atypicity is the same for R and $\bar{R}$ $\delta(a_n) = \delta(\bar{a}_n)$.

(ii) If R is a minimal realization of the degeneracy $\delta(a_n)$ then $\bar{R}$ is also a minimal realization of the atypicity $\delta(\bar{a}_n)$.

However, the determination of the Kac–Dynkin parameters of $\bar{\Lambda}$ in terms of those of Λ is still an open problem for atypical representations. Obviously this problem is identical, by contragradience, to that pointed out at the end of the previous paragraph.

III. CLASSIFICATION OF SU($n|m$) SUPERTABLEAUX

For a measure of the number of boxes of a Young supertableau we introduce the usual notations[2]

$$\left|\begin{matrix} b_j \\ \bar{b}_j \end{matrix}\right| \text{ counts the } \left|\begin{matrix} \text{covariant} \\ \text{contravariant} \end{matrix}\right| \text{ boxes of the row } j,$$

$$\left|\begin{matrix} c_k \\ \bar{c}_k \end{matrix}\right| \text{ counts the } \left|\begin{matrix} \text{covariant} \\ \text{contravariant} \end{matrix}\right| \text{ boxes of the column } k,$$

with the positivity constraints

$$\begin{aligned} &b_1 \geqslant b_2 \geqslant \cdots \geqslant b_j \geqslant \cdots \geqslant 0, \quad c_1 \geqslant c_2 \geqslant \cdots \geqslant c_k \geqslant \cdots \geqslant 0, \\ &\bar{b}_1 \geqslant \bar{b}_2 \geqslant \cdots \geqslant \bar{b}_j \geqslant \cdots \geqslant 0, \quad \bar{c}_1 \geqslant \bar{c}_2 \geqslant \cdots \geqslant \bar{c}_k \geqslant \cdots \geqslant 0. \end{aligned} \tag{12}$$

We introduce two sets of constraints B_l and C_l, where l is a non-negative integer: (i) $n - l + 1$ constraints B_l,

$$\bar{b}_{\alpha+1} + b_{n-l-\alpha+1} \leqslant m - l, \quad \alpha = 0, 1, \ldots, n - l; \tag{13}$$

(ii) $m - l + 1$ constraints C_l,

$$c_{\beta+1} + \bar{c}_{m-l-\beta+1} \leqslant n - l, \quad \beta = 0, 1, \ldots, m - l. \tag{14}$$

Such a construction is possible as long as $n - l$ and $m - l$ are non-negative. Using the minimal value L of n and m already defined in Eq. (5) we introduce a family J of indices $J = \{0, 1, \ldots, L\}$. The sets B_l and C_l are defined provided $l \in J$.

By definition a Young supertableau belongs to the set S_l if and only if at least one of the constraints B_l is satisfied. It is straightforward to check that equivalently S_l can be defined as the set of supertableaux satisfying at least one of the constraints C_l.

As a consequence of the structure of the constraints B_l and C_l we get the following relations of inclusion:

$$S_0 \supset S_1 \supset S_2 \supset \cdots \supset S_l \supset \cdots \supset S_L. \tag{15}$$

As pointed out in I the set S_0 is the set of the legal supertableaux of class one of SU($n|m$) and for that reason we have limited l to non-negative values. The cleanest proof of this result lies on the possibility for the supertableaux of S_0 and for these supertableaux only to perform a reduction with respect to the subgroup SU(n) ⊗ SU(m) ⊗ U(1) as explained in Ref. 2. This aspect will be discussed in the next section. A different argument based on the computation of the dimension of illegal supertableaux is proposed in the Appendix B.

The subset Δ_l of the set S_l is defined as the complement of S_{l+1}. By extension we put $\Delta_L \equiv S_L$. The classes of supertableaux Δ_l realize a partition of the set S_0 of legal supertableaux in nonempty disjoint subsets whose union is S_0 and where l belongs to the family J of indices. For instance, in the cases of SU($n|1$) or SU($1|m$), $L = 1$, and we only have two classes Δ_0 and Δ_1 as discussed in I.

The size σ of a supertableau belonging to the class Δ_l is simply defined by

$$\sigma(l) = L - l. \tag{16}$$

As a consequence

$$0 \leqslant \sigma \leqslant L. \tag{17}$$

The structure of the two sets of constraints B_l and C_l and the general positivity constraints (12) imply two Lemmas.

Lemma I: if the constraint α in B_l is not satisfied or simply saturated then the constraint α in $B_{l+\delta_1}$ for $0 < \delta_1 \leqslant n - l - \alpha$ and the constraint $\alpha - \delta_2$ in $B_{l+\delta_2}$ for $0 < \delta_2 \leqslant \alpha$ are not satisfied.

Proof: By assumption,

$$\bar{b}_{\alpha+1} + b_{n-l-\alpha+1} \geqslant m - l.$$

We use the positivity requirements (12)

$$\begin{aligned} &b_{n+l-\alpha+1} \leqslant b_{n-l-\alpha+1} - \delta_1, && \text{for } 0 < \delta_1 \leqslant n - l - \alpha, \\ &\bar{b}_{\alpha+1} \leqslant \bar{b}_{\alpha+1} - \delta_2, && \text{for } 0 < \delta_2 \leqslant \alpha, \end{aligned}$$

and we get

$$\begin{aligned} \bar{b}_{\alpha+1} + b_{n-l-\alpha+1-\delta_1} &\geqslant \bar{b}_{\alpha+1} + b_{n-l-\alpha+1} \\ &\geqslant m - l > m - l - \delta_1, \\ \bar{b}_{\alpha+1-\delta_2} + b_{n-l-\alpha+1} &\geqslant \bar{b}_{\alpha+1} + b_{n-l-\alpha+1} \\ &\geqslant m - l > m - l - \delta_2. \end{aligned}$$

These inequalities are the content of the Lemma I.

Lemma II: If the constraint β in C_l is not satisfied or simply saturated then the constraint β in $C_{l+\delta_1}$ for $0 < \delta_1 \leqslant m - l - \beta$ and the constraint $\beta - \delta_2$ in $C_{l+\delta_2}$ for $0 < \delta_2 \leqslant \beta$ are not satisfied.

The proof of Lemma II is entirely similar to that of Lemma I. As a consequence of these lemmas we get the following interesting result.

Result: If the supertableau T satisfies only one constraint B_l or only one constraint C_l then it belongs to the class Δ_l.

Let us point out that this sufficient condition is not a necessary one.

The operation of contragradience for supertableaux is defined as the exchange of covariant and contravariant boxes

$$b \leftrightarrow \bar{b}, \quad c \leftrightarrow \bar{c}.$$

In such a transformation the sets of contraints B_l and C_l remain globally invariant. As a consequence a supertableau T and its contragradient $\bar{T}$ have the same size and they belong to the same class Δ.

IV. HIGHEST WEIGHT OF A SUPERTABLEAU OF SU($n|m$)

A legal supertableau is entirely determined by the lengths b and $\bar{b}$ of its rows satisfying at least one constraint B_0. Equivalently it is entirely defined by the lengths c and $\bar{c}$ of its columns satisfying at least one constraint C_0. However the determination of the highest weight of a representation Λ refers to the subgroup SU(n) ⊗ SU(m) ⊗ U(1) and it is more convenient to use a description of the supertableau with both row and column parameters. In order to avoid a double counting we must decide if a given box belongs to a row or to a column. A third description of a legal supertableau of $SU(n|m)$ is then given by n parameters b and $\bar{b}$ and m column parameters c and $\bar{c}$.

Let us define as J the smallest value of α for which one constraint B_0 is satisfied and by K the smallest value of β for which one constraint C_0 is satisfied. We obviously have

$$0 \leqslant J \leqslant n, \quad 0 \leqslant K \leqslant m. \tag{18}$$

The considered tableau T belongs to the class T_J^K and we have realized another partition of the set S_0 of legal supertableaux of class one of $SU(n|m)$. From Eq. (18) the total number of these classes is $(n+1)(m+1)$. A supertableau of T_J^K is then described by

J row parameters $\bar{b}$ $\quad$ K column parameters c

$n-J$ row parameters b $\quad$ $m-K$ column parameters $\bar{c}$.

Necessary conditions of existence of such a supertableau are

$$\begin{aligned} &\bar{b}_1 \geqslant \cdots \geqslant \bar{b}_j \geqslant m-K, \quad c_1 \geqslant \cdots \geqslant c_K \geqslant n-J, \\ &b_1 \geqslant \cdots \geqslant b_{n-J} \geqslant K, \quad \bar{c}_1 \geqslant \cdots \geqslant \bar{c}_{m-K} \geqslant J. \end{aligned} \tag{19}$$

Let us point out that other nontrivial positivity restrictions for the parameters of a supertableau of the class T_J^K originate in the fact that, by definition of J and K, the constraints B_0 for $0 \leqslant \alpha < J$ and the constraints C_0 for $0 \leqslant \beta < K$ are not satisfied. We shall come back to this point later.

The method of determination of the highest weight Λ of a supertableau T has been introduced in I. We have to make a reduction of the supertableau T with respect to the subgroup $G_0 \equiv SU(n) \otimes SU(m) \otimes U(1)$ and the highest weight Λ is associated to a pair of Young tableaux (X, Y) which describes an irreducible representation of G_0 as explained in Sec. II and Appendix A. In the language of I we associate to the supertableau T a point $P(T)$ in a $n+m$ dimensional lattice space and the coordinates $x_1, \ldots, x_n; y_1, \ldots, y_m$ of P define the Young tableau X and Y. The Kac–Dynkin parameters are then computed from the coordinates of P by the formulas given in Appendix A. The relation between the coordinates of $P(T)$ and the supertableau parameters b, $\bar{b}$, c, and $\bar{c}$ is discussed in Appendix C.

We first consider the case $0 < J < n$, $0 < K < m$ where the four types of parameters b, $\bar{b}$, c, and $\bar{c}$ are present. By definition of J and K the constraints $\alpha = J-1$ of B_0 and $\beta = K-1$ of C_0 are not satisfied:

$$\bar{b}_J + b_{n-J+2} \geqslant m+1, \quad c_K + \bar{c}_{m-K+2} \geqslant n+1. \tag{20}$$

Using conditions (19) we also have

$$b_{n-J+2} \leqslant b_{n-J+1} \leqslant K, \quad \bar{c}_{m-K+2} \leqslant \bar{c}_{m-K+1} \leqslant J, \tag{21}$$

and therefore we obtain for $\bar{b}_j$ and c_K positivity constraints which are more restrictive than conditions (19):

$$\bar{b}_J \geqslant m-K+1, \quad c_K \geqslant n-J+1. \tag{22}$$

The simultaneous lower limit for $\bar{b}_J$ and c_K is also excluded. The coordinates of the point $P(T)$ are given in terms of the row and column parameters by the expressions derived in Appendix C:

$$\begin{aligned} x_s &= b_s, & s &= 1,2,\ldots,n-J, \\ x_s &= m - \bar{b}_{n+1-s}, & s &= n-J+1,\ldots,n; \\ y_t &= c_t - n, & t &= 1,2\ldots,K, \\ y_t &= -\bar{c}_{m+1-t}, & t &= K+1,\ldots,m. \end{aligned} \tag{23}$$

The $SU(n)$ Kac–Dynkin parameters are computed by using Eq. (A2)

$$\begin{aligned} a_s &= b_s - b_{s+1}, & s &= 1,2,\ldots,n-J-1, \\ a_{n-J} &= b_{n-J} + \bar{b}_J - m, & & \\ a_s &= \bar{b}_{n-s} - \bar{b}_{n-s+1}, & s &= n-J+1,\ldots,n-1; \end{aligned} \tag{24}$$

and the $SU(m)$ Kac–Dynkin parameters by using Eq. (A6)

$$\begin{aligned} a_{n+t} &= c_t - c_{t+1}, & t &= 1,2,\ldots,K-1, \\ a_{n+K} &= c_k + \bar{c}_{m-K} - n, & & \\ a_{n+t} &= \bar{c}_{m-t} - \bar{c}_{m-t+1}, & t &= K+1,\ldots,m-1. \end{aligned} \tag{25}$$

Using the inequalities (19) and (22) we obtain positivity constraints on the Kac–Dynkin parameters a_{n-J} and a_{n+K} entering in the highest weight of a supertableau of the class T_J^K

$$a_{n-J} \geqslant 1, \quad a_{n+K} \geqslant 1. \tag{26}$$

Inverting the formulas (24) we compute the n row parameters b and $\bar{b}$ in terms of the $n-1$ $SU(n)$ Kac–Dynkin parameters a_s and one free parameter it is convenient to choose as the coordinate x_n

$$\begin{aligned} b_s &= x_n + \sum_1^{n-s} a_{n-r}, \quad s = 1,2,\ldots,n-J, \\ \bar{b}_s &= m - x_n - \sum_1^{s-1} a_{n-r}, \quad s = 1,2,\ldots,J. \end{aligned} \tag{27}$$

Inverting the formulas (25) we compute the m column parameters c and $\bar{c}$ in terms of the $m-1$ $SU(m)$ Kac–Dynkin parameters a_{n+t} and one free parameter, it is convenient to choose as the coordinate y_1:

$$\begin{aligned} c_t &= n + y_1 - \sum_1^{t-1} a_{n+r}, \quad t = 1,2,\ldots,K; \\ \bar{c}_t &= -y_1 + \sum_1^{m-t} a_{n+r}, \quad t = 1,2,\ldots,m-K. \end{aligned} \tag{28}$$

The ranges of variation of the parameters x_n and y_1 are governed by the positivity constraints (19) and (22) on $b_{n-J}, \bar{b}_j, c_K$ and $\bar{c}_{m-K}$. We get

$$\begin{aligned} &K - \sum_1^J a_{n-r} \leqslant x_n \leqslant K - 1 - \sum_1^{J-1} a_{n-r}, \\ &-J + 1 + \sum_1^{K-1} a_{n+r} \leqslant y_1 \leqslant -J + \sum_1^K a_{n+r} \end{aligned} \tag{29}$$

Let us point out that the ranges of variation (29) are not empty as soon as the constraints (26) are satisfied.

Finally the last Kac–Dynkin parameter a_n is given by Eq. (A11)

$$a_n = x_n + y_1.$$

As a consequence of the inequalities (29) we immediately compute a lower bound and an upper bound for the possible values of the Kac–Dynkin parameter a_n of the supertableau of the class T_J^K. Using Eq. (3) we get

$$A_{JK-1} + 2 \leqslant a_n \leqslant A_{J-1K} - 2. \tag{30}$$

TABLE I. Positivity constraints for the row and column parameters of a supertableau of class T_J^K.

	$J=0$		$0<J<n$		$J=n$	
$K=0$	–	$b_n \geqslant 0$	$\bar{b}_j \geqslant m+1$	$b_{n-J} \geqslant 0$	$\bar{b}_n \geqslant m+1$	–
	–	$\bar{c}_m \geqslant 0$	–	$\bar{c}_m \geqslant J$	–	$\bar{c}_m \geqslant n$
$0<K<m$	–	$b_n \geqslant K$	$\bar{b}_j \geqslant m-K+1$	$b_{n-J} \geqslant K$	$b_n \geqslant m-K+1$	–
	$c_K \geqslant n+1$	$\bar{c}_{m-K} \geqslant 0$	$c_K \geqslant n-J+1$	$\bar{c}_{m-K} \geqslant J$	$c_K \geqslant 1$	$\bar{c}_{m-K} \geqslant n$
$K=m$	–	$b_n \geqslant m$	$\bar{b}_j \geqslant 1$	$b_{n-J} \geqslant m$	$b_n \geqslant 1$	–
	$c_m \geqslant n+1$	–	$c_m \geqslant n-J+1$	–	$c_m \geqslant 1$	–

The limiting cases $J=0$, $J=n$, $K=0$, $K=m$ can be discussed in an analogous way. We only give here the main results collected in the tables. The positivity constraints on the length of the row and column parameters are given in the Table I. The positivity constraints (26) on the SU(n) and SU(m) Kac–Dynkin parameters extend trivially. They are indicated on the Table II with the range of variation of the Kac–Dynkin parameter a_n. We observe that for the supertableaux of the classes T_0^0 and T_n^m no restriction occurs on the Kac–Dynkin parameters. Finally, by using the bounds (4) we easily see that the supertableaux of the classes T_n^0 and T_0^m are all typical.

V. SIZE AND ATYPICITY OF SUPERTABLEAUX

We consider a supertableau T of the set S_l defined in Sec. III. Because of the inclusion property (15) the supertableau T belongs also to the sets S_λ with the integer λ in the range

$$0 \leqslant \lambda \leqslant l. \tag{31}$$

We then define as j_λ (k_λ) the smallest value of $\alpha(\beta)$ for which one constraint B_λ (C_λ) is satisfied. Of course $J=j_0$ and $K=k_0$. Using Lemmas I and II we easily see that the two sets of integers j_λ and k_λ are ordered as follows:

$$\begin{aligned} &0 \leqslant J \leqslant j_1 \leqslant \cdots \leqslant j_\lambda \leqslant \cdots \leqslant j_l \leqslant n-l, \\ &0 \leqslant K \leqslant k_1 \leqslant \cdots \leqslant k_\lambda \leqslant \cdots \leqslant k_l \leqslant m-l. \end{aligned} \tag{32}$$

The supertableau T belongs to the class T_J^K with J and K constrained by the inequalities (32). Conversely a supertableau T of the class T_J^K can belong only to sets S_l for which

$$0 \leqslant l \leqslant L(J,K), \tag{33}$$

where

$$L(J,K) = \min[n-J, m-K]. \tag{34}$$

In particular when $J=n$ or when $K=m$ we have $L(n,K) = L(J,m) = 0$ and the supertableaux of the classes T_n^K with $0 \leqslant K \leqslant m$ and T_J^m with $0 \leqslant J \leqslant n$ cannot belong to the set S_1. Therefore they have the maximal size $\sigma = L$ and they all are in the class Δ_0.

Theorem I: If a supertableau T belongs to the set S_1 then its highest weight is atypical with $a_n = A_{j_1 k_1}$.

Let us simply, in the text, give the proof in the case $j_1 = J$ and $k_1 = K$. By assumption the supertableau T satisfies the additional constraints of B_1 and C_1:

$$\bar{b}_{J+1} + b_{n-J} \leqslant m-1, \tag{35}$$

$$c_{K+1} + \bar{c}_{m-K} \leqslant n-1. \tag{36}$$

On the other hand, the supertableau T being in the class T_J^K, we have the relations of positivity

$$b_{n-J} \geqslant K, \quad \bar{c}_{m-K} \geqslant J, \quad \bar{b}_{J+1} \leqslant m-K, \quad c_{K+1} \leqslant n-J,$$

and the only possible solution of the constraints (35) and (36) turns out to be

$$b_{n-J} = K, \quad \bar{c}_{m-K} = J. \tag{37}$$

The equalities (37) determine the parameters x_n and y_1. But using Eqs. (27) and (28) we get

$$x_n = K - \sum_1^J a_{n-r}, \quad y_1 = -J + \sum_1^K a_{n+r}, \tag{38}$$

and for the Kac–Dynkin parameter a_n we obtain an atypical value

$$a_n = x_n + y_1 = A_{JK}. \tag{39}$$

The general proof of Theorem I is given in Appendix D.

Corollary I: A typical supertableau has $\delta(a_n) = 0$. From Theorem I, a typical supertableau cannot belong to the set S_1 and therefore it belongs to the class Δ_0.

Theorem II: If a supertableau T belongs to the set S_l with $2 \leqslant l \leqslant L$ then it is atypical from Theorem I with

TABLE II. Constraints on the Kac–Dynkin parameters for a supertableau of class T_J^K.

	$J=0$	$0<J<n$	$J=n$
$K=0$		$a_{n-J} \geqslant 1$ $a_n \leqslant A_{J-1\,0} - 2$	$a_n \leqslant A_{n-1\,0} - 2$
$0<K<m$	$a_{n+K} \geqslant 1$ $A_{0\,K-1} + 2 \leqslant a_n$	$a_{n-J} \geqslant 1 \quad a_{n+K} \geqslant 1$ $A_{J\,K-1} + 2 \leqslant a_n \leqslant A_{J-1\,K} - 2$	$a_{n+K} \geqslant 1$ $a_n \leqslant A_{n-1\,K} - 2$
$K=m$	$A_{0\,m-1} + 2 \leqslant a_n$	$a_{n-J} \geqslant 1$ $A_{J\,m-1} + 2 \leqslant a_n$	

$a_n = A_{j_1 k_1}$ and the degeneracy $\delta(a_n)$ is lower bounded by l.

We only give, in the next, the proof for $l = 2$ with $j_2 = j_1 = J$ and $k_2 = k_1 = K$. In addition to the constraints (35) and (36) of B_1 and C_1 we have the two constraints due to B_2 and C_2

$$\bar{b}_{J+1} + b_{n-J-1} \leqslant m - 2, \tag{40}$$

$$c_{K+1} + \bar{c}_{m-K-1} \leqslant n - 2. \tag{41}$$

The supertableau T being in the class T_J^K we also have the inequalities

$$b_{n-J-1} \geqslant K, \quad \bar{c}_{m-K-1} \geqslant J,$$

and the only possible solution of the constraints (35), (36), (40), and (41) is

$$b_{n-J} = b_{n-J-1} = K, \tag{42}$$

$$\bar{c}_{m-K} = \bar{c}_{m-K-1} = J. \tag{43}$$

As previously, the Kac–Dynkin parameter a_n is atypical,

$$a_n = A_{JK}, \tag{44}$$

and the equalities (42) and (43) imply the vanishing of two Kac–Dynkin parameters

$$a_{n-J-1} = 0, \quad a_{n+K+1} = 0. \tag{45}$$

We have the equality of two atypical values

$$A_{JK} = A_{J+1K+1}, \tag{46}$$

and we get $\delta(a_n) \geqslant 2$. It must be noticed that the relations (45) provide a minimal realization of the degeneracy $\delta = 2$ associated to Eq. (46). The general proof of Theorem II is given in Appendix E.

Corollary II: The supertableaux of the class Δ_l for $1 \leqslant l \leqslant L$ are atypical and the degeneracy $\delta(a_n)$ is restricted by a double inequality

$$l \leqslant \delta(a_n) \leqslant L. \tag{47}$$

In particular the atypical supertableaux of the class Δ_L have the degeneracy $\delta(a_n) = \mathrm{L}$. This result is a trivial consequence of the two theorems and of the inequality (6).

Corollary III: If a supertableau of the set S_l has a degeneracy $\delta(a_n) = l$ then it belongs to the class Δ_l and for $2 \leqslant l \leqslant L$ it is a minimal realization of the degeneracy $\delta(a_n) = L$.

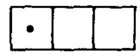
Again, Corollary III is an immediate consequence of the two theorems. In particular the supertableau T cannot be in the set S_{l+1} because in S_{l+1}, $\delta(a_n) \geqslant l + 1$ which contradicts the assumption $\delta(a_n) = l$. Being in S_l the supertableau T must belong to Δ_l.

VI. IRREDUCIBILITY OF SUPERTABLEAUX

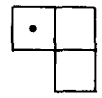
As a result (obtained in Ref. 2) the purely covariant and the purely contravariant legal supertableaux are associated to irreducible representations of SU$(n|m)$ and they will be called irreducible supertableaux. The situation is different for mixed supertableaux because of the existence of traces between the covariant part $\mathscr{T}_1$ and the contravariant part $\mathscr{T}_2$ of a mixed supertableau T. Both $\mathscr{T}_1$ and $\mathscr{T}_2$ are legal irreducible supertableaux and their tensor product is symbolically written as

$$\mathscr{T}_1 \otimes \mathscr{T}_2 = T \oplus \text{ trace terms.} \tag{48}$$

The supertableau T is irreducible if and only if the trace terms in $\mathscr{T}_1 \otimes \mathscr{T}_2$ can be isolated. If not the representation $\mathscr{T}_1 \otimes \mathscr{T}_2$ is not fully reducible. As an illustration we discuss the simple example of the supertableaux T_+ and T_- represented below:



T_+



T_-

The covariant part $\mathscr{T}_{1+}(\mathscr{T}_{1-})$ is the second rank supersymmetric (superantisymmetric) tensor and the contravariant part $\mathscr{T}_2$ is the fundamental supertableau $\bar{F}$. In the graded vector spaces of the fundamental representations F and $\bar{F}$, we define the basis

$$F{:}|Ag_A\rangle, \quad \bar{F}{:}\langle Ag_A|,$$

where g_A is the degree of the index A ($g_A = 0$ or 1 modulo 2). In the vector space $\mathscr{T}_1 \otimes \mathscr{T}_2$ the induced basis is[6]

$$(W_\pm)_{AB}^C = |Ag_A\rangle |Bg_B\rangle \langle Cg_C| \pm \epsilon(g_A, g_B) |Bg_B\rangle |Ag_A\rangle \langle Cg_C|, \tag{49}$$

where the commutation factor is defined as usual by

$$\epsilon(g_A, g_B) = (-1)^{g_A g_B}.$$

It is now straightforward to check that $\mathscr{T}_1 \otimes \mathscr{T}_2$ contains one invariant subspace X transforming like a fundamental representation F

$$(X_\pm)_A = \sum_{B,C} \delta_C^B \epsilon(g_B, g_C)(W_\pm)_{AB}^C. \tag{50}$$

The traceless part of $W_\pm$ has then the general form

$$(\bar{W}_\pm)_{AB}^C = (W_\pm)_{AB}^C - \alpha_\pm \, [\delta_B^C (X_\pm)_A \pm \delta_A^C \epsilon(g_A, g_B)(-X_\pm)_B]. \tag{51}$$

The constants $\alpha_\pm$ are such that the trace operation (50) applied on $\bar{W}$ gives zero. The result is[6]

$$(\alpha_\pm)^{-1} = S\,\mathrm{Tr}\, I + 1, \tag{52}$$

where the supertrace of the unit matrix is given for class one representations by

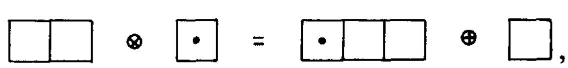
$$S\,\mathrm{Tr}\, I = \sum_A \epsilon(g_A, g_A) = n - m. \tag{53}$$

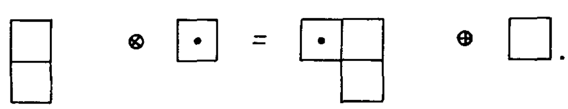
The trace term in $\mathscr{T}_1 \otimes \mathscr{T}_2$ can be separated in Eq. (48) as long as the constant α^{-1} is not vanishing. In that case the supertableau T is irreducible. We then have

T_1 irreducible, when $n - m \neq -1$,

T_- irreducible, when $n - m \neq +1$,

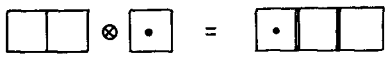
and the tensor product $\mathscr{T}_1 \otimes \mathscr{T}_2$ has the form

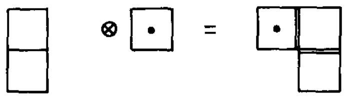




When the trace cannot be separated we obtain a nonfully reducible representation of SU$(n|m)$ and the supertableau T

is part of a generalized atypical supertableau already introduced in I:


for $n-m=-1,$


for $n-m=+1.$

This example can also be analyzed in the language of supertableaux presented in this paper. The highest weights $\Lambda_\pm$ of the supertableaux $T_\pm$ are given by

$$\Lambda_\pm \Rightarrow \{2,0,...,0|0|0,...,0,1\},$$

$$\Lambda_- \Rightarrow \{(0,1,0,...0|0|0,...,0,1\}.$$

The size $\sigma_\pm$ of the supertableaux $T_\pm$ and the degeneracies $\delta_\pm$ of the atypical value $a_n=0$ depend only on the difference $n-m$ as shown in Tables III and IV. When $\delta+\sigma=L$ the supertableaux T_+ and T_- are irreducible and they describe the irreducible atypical representations of $SU(n|m)$ of highest weights Λ_+ and Λ_-. When $\delta+\sigma=L+1$ the supertableaux T_+ and T_- have no individual meaning and they form, with the trace, a two-generalized atypical supertableau describing a nonfully reducible representation of $SU(n|m)$.

The method of separation of the trace is extremely general but its practice becomes rapidly very complicated. The result expressed in terms of size and degeneracy of the atypicity is extremely simple and we get the following theorem.

Theorem III: The supertableaux of the class Δ_l whose highest weight has a degeneracy of the atypicity $\delta(a_n)=l$ are irreducible supertableaux associated to irreducible representations of $SU(n|m)$.

The discussion of the supertableaux of the class Δ_l having $\delta(a_n)>l$ is postponed to the next section and we successively study here the three cases $l=0$, $l=1$, and $l\geqslant 2$ for $\delta=l$.

Let us first consider the typical supertableaux of the class Δ_0. They correspond to irreducible typical representations of $SU(n|m)$ but the correspondence between supertableau and representation is not one to one. In general we have an one-parameter family $F(\Lambda)$ of equivalent typical supertableaux describing the same typical representation of $SU(n|m)$ of highest weight Λ. Such a situation has been discussed in detail in I. A typical supertableau of the class T_J^K belongs to $F(\Lambda)$ if the positivity constraints of Table II are fulfilled. Then the length of its rows and columns is given by Eqs. (27) and (28) in terms of two parameters x_n and y_1 whose sum is fixed by a_n.

The supertableaux of the class Δ_1 with $\delta=1$ are atypical nondegenerate and they correspond to irreducible atypical representations of $SU(n|m)$. Conversely, to an atypical highest weight Λ with the degeneracy $\delta(a_n)=1$ we can associate one unique supertableau of Δ_1. The knowledge of the atypical value a_n determines the quantities j_1 and k_1,

$$a_n=A_{j_1k_1},$$

and the length of one row and one column,

$$b_{n-j_1}=k_1,\quad \bar{c}_{m-k_1}=j_1.$$

The other rows and columns parameters are then given by expressions analogous to (27) and (28):

$$\begin{aligned}
b_s &= k_1+\sum_{j_1+1}^{n-s} a_{n-r}, && s=1,...,n-j_1-1,\\
\bar{b}_s &= m-k_1+\sum_{s}^{j_1} a_{n-r}, && s=1,...,j_1,\\
c_t &= n-j_1+\sum_{t}^{k_1} a_{n+r}, && t=1,...,k_1,\\
\bar{c}_t &= j_1+\sum_{k_1+1}^{m-k} a_{n+r}, && t=1,...,m-k_1+1.
\end{aligned} \tag{54}$$

The irreducibility of these atypical supertableaux is discussed in Appendix F.

The supertableaux of the class Δ_l, $2\leqslant l\leqslant L$, with $\delta=l$ are atypical and they correspond to irreducible atypical representations of $SU(n|m)$ which are minimal relations of the degeneracy $\delta=l$. Conversely to an arbitrary irreducible atypical representation of $SU(n|m)$ with a degeneracy $\delta>1$ we cannot, in general, associate an irreducible atypical supertableau. The only case where such a supertableau exists and is unique is when the irreducible atypical representation is a minimal realization of the degeneracy $\delta>1$. Then the supertableau belongs to the class Δ_δ and its construction is made by using Eq. (54).

VII. GENERALIZED ATYPICAL SUPERTABLEAUX

The necessity of introducing generalized supertableaux associated to nonfully reducible representations of $SU(n|1)$ has been discussed in I for the atypical supertableaux of the class Δ_0. In the two examples of the supertableaux T_+ and T_- studied in Sec. V we have recognized that the crucial parameter is the sum $\sigma+\delta$ of the size σ of the supertableau and of the degeneracy δ of the atypicity of its highest weight. We are then lead to the following theorem.

Theorem IV: The supertableaux of the class Δ_l with $0\leqslant l\leqslant L-1$, whose highest weight has a degeneracy of atypicity $\delta>l$, have no individual existence. They participate to a ρ-generalized supertableau which is a collection of ρ atypical supertableaux describing globally a nonfully reducible representation of $SU(n|m)$ with ρ given by

$$\rho=2^{(\delta-l)}=2^{(\sigma+\delta-L)}. \tag{55}$$

TABLE III. For T_+,Λ_+: $n\geqslant 2$, $m\geqslant 2$.

$n-m$	$\leqslant -3$	-2	-1	0	$+1$	$\geqslant +2$
σ_+	0	1	1	1	1	0
δ_+	L	$L-1$	L	$L-1$	$L-1$	L

TABLE IV. For T_-,Λ_-: $n\geqslant 3$, $m\geqslant 2$.

$n-m$	$\leqslant -2$	-1	0	$+1$	$+2$	$\geqslant +3$
σ_-	0	1	1	1	1	0
δ_-	L	$L-1$	$L-1$	L	$L-1$	L

The origin of formula (55) is discussed in the Appendix G.

The case $l = 0$ has already been discussed in I. In the class Δ_0 we have a one-parameter family of equivalent generalized supertableaux describing the same nonfully reducible representation of SU($n|m$). This equivalence property disappears for $l \geqslant 1$. In particular the various atypical supertableaux of the class Δ_1, whose common highest weight Λ has a degeneracy of atypicity $\delta \geqslant 2$, are inequivalent. Their dimensions are different and they are parts of different generalized supertableaux associated to different nonfully reducible representations.

The ρ-generalized supertableau is a collection of ρ atypical supertableaux T_a whose dimension d_a can be computed—at least formally—by the determinant method of Balentekin and Bars.[1] For the dimension of the ρ-generalized supertableau we get

$$D_T = \sum_{a=1}^{a=\rho} d_a. \tag{56}$$

The nonfully reducible representation described by this ρ-generalized supertableau has N_ρ atypical components A_α. In general we simply have $N_\rho = 2^\rho$ but this number may fluctuate in particular situations. The dimension d_α of the atypical component A_α is known by counting its SU(n) $\otimes$ SU(m) $\otimes$ U(1) components and the dimension of the nonfully reducible representation is given by

$$D_R = \sum_{\alpha=1}^{\alpha=N_\rho} d_\alpha. \tag{57}$$

As a result $D_T = D_R$ but the matching between the typical supertableaux T_a and the atypical components A_α is not uniquely well-defined essentially because these supertableaux T_a have no individual existence, in general.

VIII. CONCLUDING REMARKS

The results of our previous paper I are particular cases of those obtained here. For the supergroup SU($n|1$), $L = 1$, and the set S_0 of legal supertableaux contains only two classes Δ_1 and Δ_0.

(a) The class Δ_1 of minimal size $\sigma = 0$ has $c_1 + \bar{c}_1 < n - 1$ and it contains the irreducible atypical supertableaux describing the irreducible atypical representations of SU($n|1$).

(b) The class Δ_0 of maximal size $\sigma = 1$ has $c_1 + \bar{c}_1 \geqslant n$ and it contains two types of supertableaux: (i) the typical supertableaux describing the irreducible typical representations of SU($n|1$); and (ii) the atypical supertableaux which participate to two-generalized supertableaux associated to nonfully reducible representations of SU($n|1$) with four atypical components.

TABLE V. Structure of the supertableaux of SU(3|2).

$\sigma = 0$	$\delta = 2$	Irreducible atypical supertableau Irreducible atypical representation
$\sigma = 1$	$\delta = 1$	Irreducible atypical supertableau Irreducible atypical representation
	$\delta = 2$	Two-generalized supertableau Nonfully reducible representation
$\sigma = 2$	$\delta = 0$	Irreducible typical supertableau Irreducible typical representation
	$\delta = 1$	Two-generalized supertableau Nonfully reducible representation
	$\delta = 2$	Four-generalized supertableau Nonfully reducible representation

The supergroups SU($n|2$) and SU($2|m$) offer the simplest examples of a degeneracy of atypicity $\delta(a_n) \geqslant 2$. The set S_0 of legal supertableaux has now three classes Δ_2, Δ_1, and Δ_0 with a structure represented on Table V.

The supertableaux of the supergroup SU(3|2) have been extensively studied by us as a laboratory for the general case presented here.[7]

As found in I and in our study of the supergroup SU(3|2) for a two- generalized supertableau with four atypical components a structure of the type

$$\begin{tikzcd} & A_2 \arrow[dr] & \\ A_1 \arrow[ur] \arrow[dr] & & A_4 \\ & A_2 \arrow[ur] & \end{tikzcd}$$

is certainly present for the associated nonfully reducible representation. However the general case of a ρ-generalized supertableau with N_ρ atypical components has not yet been investigated, nor the precise value of the number of components N_ρ as a function of the highest weight Λ.

ACKNOWLEDGMENT

One of the authors (M. Gourdin) wishes to thank the C.E.R.N. Theory Division for its hospitality. This work was partly done while he was visiting it.

APPENDIX A: PAIRS OF YOUNG TABLEAUX (X, Y) DESCRIBING THE HIGHEST WEIGHT Λ

We call as X the Young tableau of the U(n) irreducible representation $\{x_1, x_2, ..., x_n\}$, where the x's are algebraic integers in the order

$$x_1 \geqslant x_2 \geqslant \cdots \geqslant x_n. \tag{A1}$$

The positive (negative) parameters x measure the length of the covariant (contravariant) rows of the tableau X. The differences of two successive x's give the Dynkin parameters of the representation

$$x_s - x_{s+1} = a_s, \quad s = 1,2,...,n-1, \tag{A2}$$

and the sum of all x's is the eigenvalue of the operator Q_1 commuting with the n^2 operators of the U(n) Lie algebra. We use the normalization

$$Q_1 = \frac{1}{n} \sum_1^n x_s. \tag{A3}$$

The U(n) unequivalent irreducible representations $\{x_1 + x, x_2 + x, ..., x_n + x\}$, where x is any algebraic integer are equivalent irreducible representations of SU(n) because of the unimodular character of these transformations. However the value of Q_1 depends linearly on x,

$$Q_1(x) = Q_1(0) + x. \tag{A4}$$

We then proceed in the same way for the SU(m) part. We call as Y the Young tableau of the $U(m)$ irreducible representation $\{y_1, y_2, ..., y_m\}$, where the y's are algebraic integers in the order

$$y_1 > y_2 > \cdots > y_m. \tag{A5}$$

The differences of two successive y's determine the Dynkin parameters

$$y_t - y_{t+1} = a_{n+t}, \quad t = 1,2,\ldots,m-1, \tag{A6}$$

and the sum of all y's is the eigenvalue of the generator Q_2 commuting with the m^2 generator of the U(m) Lie algebra. We choose the normalization

$$Q_2 = \frac{1}{m}\sum_1^m y_t. \tag{A7}$$

The U(m) unequivalent irreducible representation $\{y_1 + y, y_2 + y, \ldots, y_m + y\}$ where y is any algebraic integer are equivalent irreducible representations of SU(n) and the eigenvalue of Q_2 depends linearly on y:

$$Q_2(y) = Q_2(0) + y. \tag{A8}$$

The U(1) operator Q defined in Eq. (1) is simply related to the operator Q_1 and Q_2 by

$$Q = Q_1 + Q_2. \tag{A9}$$

Its eigenvalue Q_{HW} for the highest weight Λ represented by the pair of Young tableaux (X, Y) is then given from (A3), (A7), and (A9) by

$$Q_{\text{HW}} = \frac{1}{n}\sum_1^n x_s + \frac{1}{m}\sum_1^m y_t. \tag{A10}$$

This expression allows us to compute the last Kac–Dynkin parameter a_n using Eqs. (2), (A2), (A6), and (A10) we get

$$a_n = x_n + y_1. \tag{A11}$$

Taking into account the relations (A4), (A8), and (A9) we immediately see that the unequivalent irreducible representations of U(n) $\otimes$ U(m)

$$\{x_1 + a, x_2 + a, \ldots, x_n + a\} \times \{y_1 - a, y_2 - a, \ldots, y_m - a\},$$

where a is any algebraic integer are equivalent irreducible representations of SU(n) $\otimes$ SU(m) $\otimes$ U(1). Therefore, the pairs of Young tableaux (X, Y) for which the parameters x's and y's satisfy the relations (A2), (A6), and (A11) are all associated to the same highest weight Λ and we get a one-parameter family $\mathcal{F}_\Lambda$ of equivalent pairs of Young tableaux.

The problem of determining which pair of this family is associated to a supertableau T of highest weight Λ is then solved by reducing the supertableau T with respect to the subgroup SU(n) $\otimes$ SU(m) $\otimes$ U(1). For details concerning the Young tableaux of unitary groups see Ref. 8.

APPENDIX B: LIMITS ON THE SIZE OF SUPERTABLEAUX

Let us first consider as in I the simple case of the supergroup SU($n|1$). The constraints C_0 of Eq. (14) reduce to

$$c_1 + \bar{c}_2 \leqslant n, \quad c_2 + \bar{c}_1 \leqslant n. \tag{B1}$$

We want to show that the dimension of the supertableau vanishes when the constraint (B1) are just violated and for that purpose we consider a four-column supertableau T as follows:

$$\begin{aligned} c_1 &= J_1, \quad \bar{c}_2 = n + 1 - J_1, \\ c_2 &= J_2, \quad \bar{c}_1 = n + 1 - J_2, \end{aligned} \tag{B2}$$

with $1 \leqslant J_1 \leqslant n$, $1 \leqslant J_2 \leqslant n$. The dimension of T is given by a 4×4 determinant (see Ref. 6):

$$D = \begin{vmatrix} a_{n+1-J_1} & a_{n+2-J_2} & a_{J_1-2} & a_{J_2-3} \\ a_{n-J_1} & a_{n+1-J_2} & a_{J_1-1} & a_{J_2-2} \\ a_{n-1-J_1} & a_{n-J_2} & a_{J_1} & a_{J_2-1} \\ a_{n-2-J_1} & a_{n-1-J_2} & a_{J_1+1} & a_{J_2} \end{vmatrix}, \tag{B3}$$

where a_J is the dimension of a supertableau of class one of SU($n|1$) with J boxes in one column. This dimension a_J is defined by a contour integral[6]

$$a_J = \frac{1}{2i\pi}\oint \frac{dz}{z^{J+1}}\frac{(1+z)^n}{1-z}. \tag{B4}$$

The computation of the dimensions a_J is made by using the recurrence relation

$$a_J - a_{J-1} = \frac{1}{2i\pi}\oint \frac{dz}{z^{J+1}}(1+z)^n, \tag{B5}$$

and the result of this last contour integral is

$$\begin{aligned} a_J - a_{J-1} &= C_n^J, \quad \text{for } 0 \leqslant J \leqslant n, \\ a_J - a_{J-1} &= 0, \quad \text{for } n+1 \leqslant J, \end{aligned} \tag{B6}$$

where the binomial coefficients C_n^J are given, as usual, by

$$C_n^J = n!/J!(n-J)!. \tag{B7}$$

The results for a_J are then

$$a_J = \begin{cases} 0, & \text{for } J < 0, \\ \sum_0^J C_n^k, & \text{for } 0 \leqslant J \leqslant n, \\ 2^n, & \text{for } n \leqslant J. \end{cases} \tag{B8}$$

By using the symmetry relation $C_n^k = C_n^{n-k}$ which is obvious on Eq. (B7) it is straightforward to get the interesting result

$$a_J + a_{n-J-1} = 2^n. \tag{B9}$$

It follows that all the elements of the column sum of the columns 1 and 3 of D are equal to 2^n and all the elements of the column sum of the columns 2 and 4 of D are also equal to 2^n. As a trivial consequence the determinant D vanishes.

We now extend these considerations to the supergroup SU($n|m$). We consider a supertableau T with $2(m+1)$ columns just violating the constraints C_0 of Eq. (14)

$$c_k = J_k, \quad \bar{c}_k = n + 1 - J_{m+2-k}, \quad k = 1,2,\ldots, m+1. \tag{B10}$$

The dimension of the supertableau T is computed with a $(2m+1)\times(2m+1)$ determinant D whose elements a_J are given by the contour integral[6]

$$a_j = \frac{1}{2i\pi}\oint \frac{dz}{z^{J+1}}\frac{(1+z)^n}{(1-z)^m}. \tag{B11}$$

The quantity a_J is the dimension of a supertableau of class one of SU($m|m$) with J boxes in one column and it can be computed by using a recurrence formula which now takes the form

$$\sum_0^m (-)^k C_m^k a_{J-k} = C_n^J, \quad \text{for } 0 \leqslant J \leqslant n,$$
$$\sum_0^m (-)^k C_m^k a_{J-k} = 0, \quad \text{for } n+1 \leqslant J. \tag{B12}$$

Of course, as previously, $a_J = 0$ for $J < 0$. By using again the symmetry relation $C_n^J = C_n^{n-J}$ and by combining properly the row and column elements of the determinant D as indicated in Eqs. (B12) we obtain the vanishing of the determinant D.

APPENDIX C: HIGHEST WEIGHT OF THE SUPERTABLEAUX OF SU($n|m$)

We first consider the purely covariant supertableaux. The legality condition $b_{n+1} \leqslant m$ found in Ref. 2 is simply the $\alpha = 0$ constraint B_0 and therefore these supertableaux belong to one of the classes T_0^K with $0 \leqslant K \leqslant m$. More precisely the parameter K is determined by

$$c_K \geqslant n+1, \quad c_{K+1} \leqslant n.$$

The determination of the highest weight of a purely covariant supertableau has been made by Bars, Morel, and Ruegg[2] and their result is a pair of Young tableaux (X, Y) defined by

$$x_s = b_s, \quad s = 1,2,...,n, \tag{C1}$$
$$y_t = c_t - n, \quad t = 1,2,...,K, \tag{C2}$$
$$y_t = 0, \quad t = k+1,...,m. \tag{C3}$$

The extension of this result to the mixed supertableaux of the class T_0^k is straightforward. To the n covariant rows and K covariant columns of T we may add, at most $m - K$ contravariant rows in the Young tableau Y with Eq. (C3) replaced by

$$y_t = -\bar{c}_{m+1-t}, \quad t = K+1,...,m. \tag{C4}$$

It is clear that the legality of the Young tableau Y is destroyed if we add to T a number of contravariant columns larger than $m - K$.

The case of purely contravariant supertableaux has also been treated in Ref. 2. The legality condition $\bar{c}_{m+1} \leqslant n$ is simply the $\beta = 0$ constraint C_0. Therefore $K = 0$ and these supertableaux belong to one of the classes T_J^0 with J determined by

$$\bar{b}_J \geqslant m+1, \quad \bar{b}_{J+1} \leqslant m.$$

The determination of the highest weight of a purely contravariant supertableau leads to a pair of Young tableau (X, Y) defined by[2]

$$x_s = 0, \quad s = 1,2,...,n-J, \tag{C5}$$
$$x_s = m - \bar{b}_{n-s+1}, \quad s = n-J+1,...,n, \tag{C6}$$
$$y_t = -\bar{c}_{m+1-t}, \quad t = 1,2,...,m. \tag{C7}$$

The extension of this result to the mixed supertableaux of the class T_J^0 is straightforward. To the m contravariant columns and J contravariant rows of T we may add, at most, $n - J$ covariant rows and these covariant rows will appear as covariant rows in the Young tableau X with Eq. (C1) replaced by

$$x_s = b_s, \quad s = 1,2,...,n-J. \tag{C8}$$

Again the legality of the Young tableau X is destroyed if we add to T a number of covariant rows larger than $n - J$.

Let us now study a supertableau T of the class T_n^m with n contravariant rows and m covariant columns. The contragradient supertableau $\overline{T}$ has n covariant rows and m contravariant columns and it belongs to the class T_0^0. We have the relations of contragradience

$$b_s(\overline{T}) = \bar{b}_s(T), \quad s = 1,2,...,n,$$
$$\bar{c}_t(\overline{T}) = c_t(T), \quad t = 1,2,...,m.$$

The highest weight of a supertableau of the class T_0^0 has been previously determined in the Appendix. The pair of Young tableaux $(\overline{X}, \overline{Y})$ associated to the highest weight $\overline{\Lambda}$ of $\overline{T}$ has its parameters given by Eqs. (C4) and (C8)

$$\overline{X} \Rightarrow x_s = b_s(\overline{T}), \quad s = 1,2,...,n,$$
$$\overline{T} \Rightarrow y_t = -\bar{c}_{m+1-t}(\overline{T}), \quad t = 1,2,...,m. \tag{C9}$$

The contragradient weight of the highest weight $\overline{\Lambda}$ of the supertableau $\overline{T}$ is the lowest weight Λ_0 of the supertableau T by contragradience

$$x_s \Rightarrow -x_{n+1-s}, \quad y_t \Rightarrow -y_{n+1-t},$$

and the pair of Young tableaux (X_0, Y_0) describing the lowest weight Λ_0 of T have their parameters given, from (C9), by

$$X_0 \Rightarrow x_s = -\bar{b}_{n+1-s}(T), \quad s = 1,2,...,n,$$
$$Y_0 \Rightarrow y_t = c_t(T), \quad t = 1,2,...,m. \tag{C10}$$

The highest weight Λ of the supertableau T is now constructed from its lowest weight Λ_0 by the usual procedure of transfer to boxes between the SU(n) and SU(m) parts. Both supertableaux T and $\overline{T}$ satisfy the result of Sec. III and then the transfer is maximal, e.g., the two extreme levels in Q, $Q_{\rm HW}$ and $Q_{\rm LW}$, have the same SU(n) $\otimes$ SU(m) component and the separation formula (11) holds:

$$Q_{\rm LW} = Q_{\rm HW} + n - m.$$

As a consequence the pairs of Young tableaux (X, Y) for the highest weight Λ and (X_0, Y_0) for the lowest weight Λ_0 are simply related by

$$X = X_0 \otimes \{m,m,...,m\}, \qquad \text{in U}(n),$$
$$Y = Y_0 \otimes \{-n,-n,...,-n\}, \quad \text{in U}(m), \tag{C11}$$

where the right factors in Eq. (C11) are, respectively, a SU(n) singlet and a SU(m) singlet. We then obtain the parameters of the highest weight Λ of a supertableau of class T_n^m

$$X \Rightarrow x_s = m - \bar{b}_{n+1-s}, \quad s = 1,2,...,n,$$
$$Y \Rightarrow y_t = c_t - n, \quad t = 1,2,...,m. \tag{C12}$$

Consider now the general case of a supertableau T of class T_J^K with $J \geqslant 1$, $K \geqslant 1$ as represented in Fig. 1. Using the results (C1)–(C8) we cut the covariant part of T at the row n and its contravariant part at the column m. We then obtain the two tableaux N and M respectively associated to the SU(n) and to the SU(m) parts as shown in Fig. 2. The integers j and k used in Fig. 2 are defined by

$$\bar{b}_j - m \geqslant 0, \quad \bar{b}_{j+1} - m < 0, \quad 0 \leqslant j \leqslant J,$$
$$c_k - n \geqslant 0, \quad c_{k+1} - n < 0, \quad 0 \leqslant k \leqslant K.$$

The tableau $N(M)$ is not a legal Young tableau of SU(n) [SU(m)] because it contains more than $n(m)$ rows. In order to

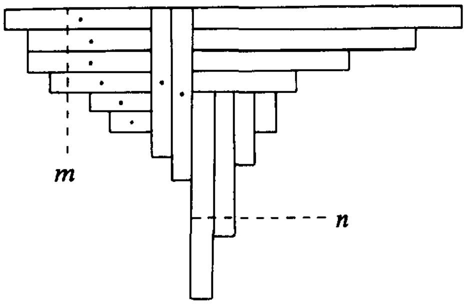


FIG. 1. General mixed supertableau T.

obtain the Young tableaux X and Y describing the highest weight of T we must modify N and M by transferring boxes and annihilating covariant and contravariant boxes. Two possibilities exist: (i) annihilate covariant and contravariant boxes in N and in M independently; and (ii) transfer boxes from $N(M)$ to $M(N)$ and annihilate covariant and contravariant boxes. We have just studied the particular case $J=n$ $K=m$ by an independent technique. The lesson contained in the final result (C12), is that the second procedure is the correct one when applied to the subtableaux A and B of N and M represented in Fig. 3. The supertableaux A and B are located within two rectangles whose sides have the lengths J and K. The transfer of boxes and the annihilation of covariant and contravariant boxes has to be made in such a way as to respect the supersymmetrization of rows and columns. More precisely the covariant box belonging to the row u and the column v is annihilated by the contravariant box belonging to the row $n+1-v$ and the column $m+1-u$ with

$$n+1-J\leqslant u\leqslant n,\quad 1\leqslant v\leqslant K,$$

and such an annihilation occurs when and only when both boxes in the subtableaux N and M exist.

The result is two subtableaux $\tilde{A}$ of A and $\tilde{B}$ of B as shown in Fig. 4, and the two Young tableaux X and Y representing the highest weight of T are given in Fig. 5. The consistency of this procedure is insured by the positivity constraints on the length of the rows and columns of T. By definition of J the J first constraints B_0 are not satisfied and therefore

$$b_s\geqslant m-\bar{b}_{n+1-s},\quad \text{for } n-J+1\leqslant s\leqslant n.$$

Equivalently, by definition of K the K first constraints C_0 are not satisfied and, as a consequence of Lemma II, the K first contraints C_1 are not satisfied and therefore

$$\bar{c}_t\geqslant n-c_{m+1-t},\quad \text{for } m-K+1\leqslant t\leqslant m.$$

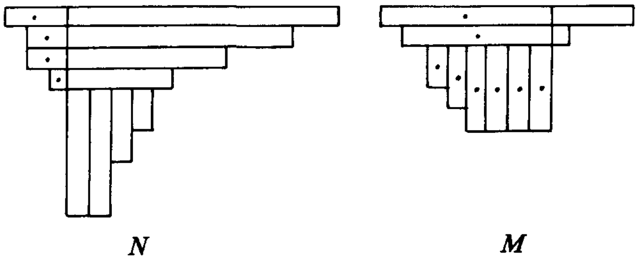


FIG. 2. Tableaux N and M.

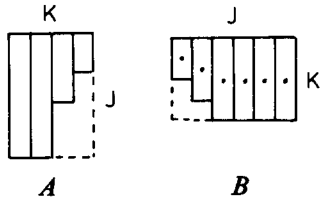


FIG. 3. Supertableaux A and B.

Finally the coordinates x_s and y_t of the Young tableaux X and Y of Fig. 9 are simply given by

$$X\Rightarrow \begin{array}{ll} x_s=b_s, & s=1,2,\dots,n-J,\\ x_s=m-\bar{b}_{m+1-s}, & s=n-J+1,\dots,n;\end{array}\tag{C13}$$

$$Y\Rightarrow \begin{array}{ll} y_t=c_t-n, & t=1,2,\dots,k,\\ y_t=-\bar{c}_{m+1-t}, & t=k+1,\dots,m,\end{array}\tag{C14}$$

It is now straightforward to check that the formulas (C13) and (C14) contain, as particular cases, those obtained earlier in this Appendix.

APPENDIX D: THE SUPERTABLEAUX OF S_1 ARE ATYPICAL

The supertableau T of the class T_J^K is assumed to belong to the set S_1 and therefore it satisfies the following inequalities of B_1 and C_1:

$$\begin{array}{l}\bar{b}_{J+1}+b_{n-J}\geqslant m,\\ \quad\vdots\\ \bar{b}_{j_1}+b_{n-j_1+1}\geqslant m,\\ \bar{b}_{j_1+1}+b_{n-j_1+1}\leqslant m-1,\end{array}\tag{D1}$$

$$\begin{array}{l}c_{K+1}+\bar{c}_{m-K+1}\geqslant n,\\ \quad\vdots\\ c_{k_1}+\bar{c}_{m-k_1+1}\geqslant n,\\ c_{k_1+1}+\bar{c}_{m-k_1}\leqslant n-1.\end{array}\tag{D2}$$

On the other hand, the supertableau T satisfies the inequalities (19) and in particular

$$\begin{array}{ll} b_{n-j_1}\geqslant b_{n-J}\geqslant K, & \text{for } j_1\geqslant J,\\ \bar{c}_{m-k_1}\geqslant\bar{c}_{m-K}\geqslant J, & \text{for } k_1\geqslant K.\end{array}\tag{D3}$$

Using the constraint $\alpha=j_1$ of B_1 we get

$$\bar{b}_{j_1+1}\leqslant m-1-K,\quad \text{or equivalently } \bar{c}_{m-K}\leqslant j_1,\tag{D4}$$

and from the constraint $\beta=k_1$ of C_1 we obtain

$$c_{k_1+1}\leqslant n-1-J,\quad \text{or equivalently } b_{n-J}\leqslant k_1.\tag{D5}$$

We then associate one equation α of (D1) with $J\leqslant\alpha\leqslant j_1-1$ and the equation $\beta=k_1$ of (D2) and we derive a sequence of inequalities:

$b_{n-\alpha}\leqslant k_1$ with α of (D1) implies

$\bar{b}_{\alpha+1}\geqslant m-k_1$ or $\bar{c}_{m-k_1}\geqslant\alpha+1$,

$\bar{c}_{m-k_1}\geqslant\alpha+1$ with β of (D2) implies

$c_{k_1+1}\leqslant n-\alpha-2$ or $b_{n-\alpha-1}\leqslant k_1$.

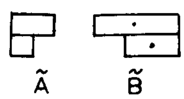


FIG. 4. Subtableaux $\tilde{A}$ and $\tilde{B}$ after the annihilation of boxes.

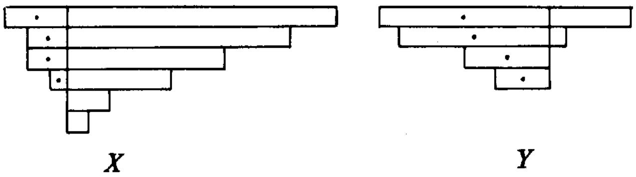


FIG. 5. Young tableaux X and Y of the highest weight Λ.

The first inequality for $\alpha = J$ is just the result (D5) and for $\alpha = j_1 - 1$ we obtain

$$\bar{c}_{m-k_1} \geqslant j_1, \quad b_{n-j_1} \leqslant k_1. \tag{D6}$$

We then proceed in an analogous way associating one equation β of (D2) $K \leqslant \beta \leqslant k_1 - 1$ and the equation $\alpha = j_1$ of (D1) and we derive a second sequence of inequalities:

$$\begin{aligned} &\bar{c}_{m-\beta} \leqslant j_1 \ \text{ with } \beta \text{ of (D2) implies} \\ &\quad c_{\beta+1} \geqslant n - j_1 \text{ or } b_{n-j_1} \geqslant \beta + 1, \\ &b_{n-j_1} \geqslant \beta + 1 \text{ with } \alpha = j_1 \text{ of (D1) implies} \\ &\quad \bar{b}_{j_1} + 1 \leqslant m - \beta - 2 \text{ or } \bar{c}_{m-\beta-1} \leqslant j_1. \end{aligned}$$

The first inequality for $\beta = K$ is just the result (D4) and for $\beta = k_1 - 1$ we obtain

$$b_{n-j_1} \geqslant k_1, \quad \bar{c}_{m-k_1} \leqslant j_1. \tag{D7}$$

Combining now the result (D6) and (D7) we get the solution

$$b_{n-j_1} = k_1, \quad \bar{c}_{m-k_1} = j_1. \tag{D8}$$

The values of the parameters x_n and y_1 are now computed from Eqs. (27) and (28),

$$x_n = k_1 - \sum_1^{j_1} a_{n-r}, \quad y_1 = -j_1 + \sum_1^{k_1} a_{n+r}, \tag{D9}$$

and the Kac-Dynkin parameter a_n takes the atypical value

$$a_n = x_n + y_1 = A_{j_1 k_1}. \tag{D10}$$

APPENDIX E: ATYPICITY OF THE SUPERTABLEAUX OF S_l

We consider a supertableau T of the class T_J^K and we assume that T belongs to the set S_l with $2 \leqslant l \leqslant L\,(J, K)$. Because of the inclusion relation (15) we have

$$T \in S_\lambda, \quad \text{for } \lambda = 0,1,\ldots,l.$$

In particular $T \in S_1$ and from Theorem I discussed in the previous Appendix T is atypical with the relations

$$b_{n-j_1} = k_1, \quad \bar{c}_{m-k_1} = j_1 \tag{E1}$$

leading to

$$a_n = A_{j_1 k_1}.$$

We first derive the consequence of the assumption $T \in S_2$. We have two sets of constraints B_2 and C_2 analogous to the sets (D1) and (D2) considered in the Appendix D:

$$\begin{aligned} &\bar{b}_{j_1+1} + b_{n-j_1-1} \geqslant m - 1, \\ &\quad \vdots \\ &\bar{b}_{j_2} + b_{n-j_2} \geqslant m - 1, \\ &\bar{b}_{j_2+1} + b_{n-j_2-1} \leqslant m - 2; \end{aligned} \tag{E2}$$

$$\begin{aligned} &c_{k_1+1} + \bar{c}_{m-k_1-1} \geqslant n - 1, \\ &\quad \vdots \\ &c_{k_2} + \bar{c}_{m-k_2} \geqslant n - 1, \\ &c_{k_2+1} + \bar{c}_{m-k_2-1} \leqslant n - 2. \end{aligned} \tag{E3}$$

We use the inequalities (19) in the form

$$\begin{aligned} &b_{n-j_2-1} \geqslant b_{n-j_1} = k_1, \quad \text{for } j_2 \geqslant j_1, \\ &\bar{c}_{m-k_2-1} \geqslant \bar{c}_{m-k_1} = j_1, \quad \text{for } k_2 \geqslant k_1. \end{aligned} \tag{E4}$$

Taking into account the constraint $\alpha = j_2$ of (E2) and the constraint $\beta = k_2$ (E3) we obtain

$$b_{n-j_1-1} \leqslant k_2, \quad \bar{c}_{m-k_1-1} \leqslant j_2. \tag{E5}$$

Now the method is the same as in Appendix D. We associate one equation α of (E2) with $j_1 \leqslant \alpha \leqslant j_2 - 1$ and the equation $\beta = k_2$ of (E3) and we derive a sequence of inequalities ending with

$$\bar{c}_{m-k_2-1} \geqslant j_2, \quad b_{n-j_2-1} \leqslant k_2. \tag{E6}$$

We then associate one equation β of (E3) with $k_1 \leqslant \beta \leqslant k_2 - 1$ and the equation $\alpha = j_2$ of (E2) and we obtain the final inequalities

$$b_{n-j_2-1} \geqslant k_2, \quad \bar{c}_{m-k_2-1} \leqslant j_2. \tag{E7}$$

The common solution of (E6) and (E7) is then

$$b_{n-j_2-1} = k_2, \quad \bar{c}_{m-k_2-1} = j_2, \tag{E8}$$

and we get a second expression for the parameter x_n and y_1,

$$x_n = k_2 - \sum_1^{j_2+1} a_{n-r}, \quad y_1 = -j_2 + \sum_1^{k_1+1} a_{n+r}, \tag{E9}$$

and the Kac–Dynkin parameter a_n,

$$a_n = A_{j_2+1 k_2+1}. \tag{E10}$$

We then have a degeneracy $\delta = 2$ for the atypical value of a_n,

$$A_{j_1 k_1} = A_{j_2+1\, k_2+1}, \tag{E11}$$

and comparing Eqs. (D9) and (E9) we obtain

$$\sum_{j_1+1}^{j_2+1} a_{n-r} = k_2 - k_1, \quad \sum_{k_1+1}^{k_2+1} a_{n+r} = j_2 - j_1. \tag{E12}$$

Equations (E12) are a minimal realization of the degeneracy $\delta = 2$ associated to Eq. (E11).

The general proof continue in the same way for $T \in S_3$ up to $T \in S_l$. At the step $T \in S_\lambda$ we freeze one row parameter and one column parameter and we can derive two relations analogous to (E12) which are again a minimal realization for a degeneracy $\delta = 2$.

The conditions

$$b_{n-j_\lambda-\lambda+1} = k_\lambda, \quad \bar{c}_{m-k_\lambda-\lambda+1} = j_\lambda \tag{E13}$$

imply for x_n and y_1 the values

$$x_n = k_\lambda - \sum_{1}^{j_\lambda + \lambda - 1} a_{n-r},$$
$$y_1 = -j_\lambda + \sum_{1}^{k_\lambda + \lambda - 1} a_{n+r}, \tag{E14}$$

and a new expression of the Kac–Dynkin parameter a_n is

$$a_n = A_{j_\lambda + \lambda - 1\, k_\lambda + \lambda - 1}. \tag{E15}$$

The degeneracy equality

$$A_{j_{\lambda-1} + \lambda - 2\, k_{\lambda-1} + \lambda - 2} = A_{j_\lambda + \lambda - 1\, k_\lambda + \lambda - 1} \tag{E16}$$

is satisfied with the minimal realization

$$\sum_{j_{\lambda-1}+\lambda-1}^{j_\lambda+\lambda-1} a_{n-r} = k_\lambda - k_{\lambda-1},$$
$$\sum_{k_{\lambda-1}+\lambda-1}^{k_\lambda+\lambda-1} a_{n+r} = j_\lambda - j_{\lambda-1}. \tag{E17}$$

APPENDIX F: IRREDUCTIBILITY OF THE ATYPICAL SUPERTABLEAUX WITH A DEGENERACY $\delta = 1$

We study the supertensor associated to a supertableau for which the highest weight Λ is atypical and nondegenerate $\delta = 1$. The problem is to recognize if this supertensor is an irreducible representation of SU$(n|m)$ or if it is a part of a nonfully reducible representation in which case a second highest weight is present in the representation. We then construct this second highest weight in order to see when it disappears.

Before proceeding let us be reminded of some useful features of the Lie superalgebra of the superunitary groups. The notations used are those of Ref. 2.

The operators of the bosonic sector of the superalgebra corresponding to simple positive and negative roots are

$$P_l^+ = E_l^{l+1}, \quad P_l^- = E_{l+1}^l,$$
$$l = 1,2,...,n-1,\ n+1,...,n+m-1.$$

The operators of the fermionic sector associated to positive and negative roots are

$$I_j^{n+k} = E_j^{n+k}, \quad I_{n+k}^j = E_{n+k}^j,$$
$$j = 1,2,...,n, \quad k = 1,2,...,m.$$

The Cartan subalgebra operators are defined by

$$H_l = E_l^l - E_{l+1}^{l+1}, \quad l = 1,2,...n-1,\ n+1,...,n+m-1,$$
$$H_n = E_n^n + E_{n+1}^{n+1}.$$

The highest weight Λ is an eigenvector of the Cartan operators with, as eigenvalues, the Kac–Dynkin parameters a_l,

$$H_l|\Lambda\rangle = a_l|\Lambda\rangle, \quad l = 1,2,...,n+m-1,$$

and it satisfies the constraints

$$P_l^+|\Lambda\rangle = 0, \quad l = 1,...,n-1,\ n+1,...,n+m-1,$$
$$I_n^{n+1}|\Lambda\rangle = 0. \tag{F1}$$

We start with the simple case of the supergroup SU$(n|1)$ and this subsection is an Appendix of I. Let us first define a set of n vectors $|\Lambda_j\rangle$ by the application of lowering fermionic operators to the highest weight Λ,

$$|\Lambda_j\rangle = I_{n+1}^{n-j} I_{n+1}^{n-j+1} \cdots I_{n+1}^{n}|\Lambda\rangle, \quad j = 0,1,...,n-1. \tag{F2}$$

We now introduce the n vectors

$$|\Omega_j\rangle = I_n^{n+1} n_{n-1}^{n+1} \cdots I_{n-j+1}^{n+1}|\Lambda_j\rangle, \quad j = 0,1,...,n-1, \tag{F3}$$

and we compute the action of the raising operators P_l^+ and I_n^{n+1} by using the commutation relations of the superalgebra and the property (F1) of the highest weight. The result is

$$P_l^+|\Omega_j\rangle = \delta_{l\,n-j}(a_n - A_j)|\Omega_{j-1}\rangle, \quad j = 1,2,...,n-1, \tag{F4}$$
$$I_n^{n+1}|\Omega_j\rangle = \delta_{j0} a_n|\Lambda\rangle,$$

where A_j is the atypical value A_{j0} defined in Eq. (3). If the representation is typical all the vectors $|\Omega_j\rangle$ are nonzero.

We suppose the representation to be atypical with $a_n = A_j$. If it is irreducible we have no second highest weight and

$$|\Omega_j\rangle = 0. \tag{F5}$$

We then introduce the vectors

$$|\Omega_j^{(p)}\rangle = I_n^{n+1} \cdots \widehat{I_p^{n+1}} \cdots I_{n-j+1}^{n+1}|\Lambda_j\rangle,$$
$$p = n-j+1,...,n, \tag{F6}$$

where the sign $\frown$ on the operator I_p^{n+1} indicates that this operator is missing in the product. The action of the raising operators P_l^+ and I_n^{n+1} is computed as previously and we find

$$P_l^+|\Omega_j^{(p)}\rangle = \delta_{lp}|\Omega_j^{(p+1)}\rangle, \quad p = n-j+1,...,n-1,$$
$$I_n^{n+1}|\Omega_j^{(p)}\rangle = \delta_{np}|\Omega_j\rangle. \tag{F7}$$

By assumption there is no other highest weight in the representation besides $|\Lambda\rangle$ and as a consequence the vectors $|\Omega_j^{(p)}\rangle$ must vanish:

$$|\Omega_j^{(p)}\rangle = 0, \quad p = n-j+1,...,n-1. \tag{F8}$$

Iterating the procedure we can prove the vanishing of all the vectors of the form

$$|\Omega_j^{(p_1 \cdots p_r)}\rangle = I_n^{n+1} \cdots \widehat{I_{p_1}^{n+1}} \cdots \widehat{I_{p_r}^{n+1}} \cdots I_{n-j+1}^{n+1}|\Omega_j\rangle = 0. \tag{F9}$$

The only possible solution to the full set of equations (F5), (F8), and (F9) is

$$|\Lambda_j\rangle = 0. \tag{F10}$$

We then obtain a necessary condition for the atypical representation of highest weight Λ and atypicity $a_n = A_j$ to be irreducible.

Let us apply condition (F10) to the atypical supertableaux of SU$(n|1)$. The supertableaux of the class Δ_1 have $c_1 + \bar{c}_1 \leqslant n-1$ and we call j the smallest value of α for which one constraint B_1 is satisfied:

$$\bar{b}_{j+1} + b_{n-j} = 0.$$

Therefore $\bar{b}_j \geqslant 1$, $\bar{c}_1 = j$, and $c_1 \leqslant n-j-1$.

The highest weight Λ of the supertableau is associated to the point P of coordinates

$$P\quad \begin{aligned} &x_s = b_s, && s = 1,...,n-j-1,\\ &x_{n-j} = 0, &&\\ &x_s = 1 - \bar{b}_{n+1-s}, && s = n-j+1,...n\\ &y_1 = -\bar{c}_1 = -j, && \end{aligned}$$

and the parameter a_n has the atypical value $a_n = A_j$.

The vector Λ_{j-1} is obtained from the highest weight after the transfer of the j controvariant boxes of the y part to the x part and the coordinates of the point P_{j-1} associated to Λ_{j-1} are simply

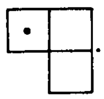
$$P_{j-1}\quad \begin{array}{ll} x_s = b_s, & s = 1,\ldots,n-j-1, \\ x_{n-j} = 0, & \\ x_s = -\bar{b}_{n+1-s}, & s = n-j+1,\ldots,n \\ y_1 = 0. & \end{array}$$

An additional transfer is not possible and we get the vanishing of $|\Lambda_j\rangle$:

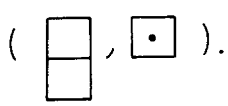
$$|\Lambda_j\rangle = I^{n-j}_{n+1}|\Lambda_{j-1}\rangle = 0.$$

For the supertableaux of the class Δ_0 which are atypical we have $c_1 + \bar{c}_1 \geqslant n+1$ as shown in I and the relation (F10) is not satisfied.

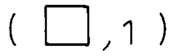
As an example consider the mixed supertableau T,



It satisifes $a_n = 0$ and the $SU(2)\otimes U(1)$ content of the first highest weight Λ is



The vector $|\Lambda_0\rangle = I^2_j|\Lambda\rangle$ represented by



is nonzero and it is a second highest weight.

We proceed in an analogous way for the general case of the $SU(n|m)$ supergroup. The algebra has been previously defined and we introduce the vectors

$$|\Lambda_{jk}\rangle = \prod_{t=n+k+1}^{t=n+1}\left(\prod_{s=n-j}^{s=n} I^s_t\right)|\Lambda\rangle \tag{F11}$$

and

$$|\Omega_{jk}\rangle = \prod_{t'=n+1}^{t'=n+k+1}\left(\prod_{s'=n}^{s'=n-j} I^{t'}_{s'}\right)|\Lambda_{jk}\rangle, \tag{F12}$$

for $j = 0,1,\ldots,n-1$ and $k = 0,1,\ldots,m-1$ and $(s',t') \neq n-j,\ n+k+1$. The action of the creation operators on the vectors $|\Omega_{jk}\rangle$ is now given by

$$\begin{aligned} P_l^+|\Omega_{jk}\rangle &= \delta^{n-j}_l \prod_{k'=0}^{k'=k}(a_n - A_{jk'})|\Omega_{j-1\,k}\rangle \\ &\quad - \delta^{n+k}_l \prod_{j'=0}^{j'=j}(a_n - A_{j'k})|\Omega_{j\,k-1}\rangle, \end{aligned} \tag{F13}$$

$$I^{n+1}_n|\Omega_{jk}\rangle = \delta_{j0}\delta_{k0}a_n|\Lambda\rangle,$$

where the A_{jk}'s are the atypical values defined in Eq. (3). If the representation is typical all the vectors $|\Omega_{jk}\rangle$ are nonzero. If the representation is atypical with $a_n = A_{jk}$ the vector $|\Omega_{jk}\rangle$ is either zero or a second highest weight. Therefore for an irreducible representation $|\Omega_{jk}\rangle = 0$.

Proceedings as in the $SU(n|1)$ case one can show that any vector deduced from $|\Omega_{jk}\rangle$ by suppressing any number of fermionic raising operators must also vanish and the final condition is

$$|\Lambda_{jk}\rangle = 0. \tag{F14}$$

APPENDIX G: GENERALIZED SUPERTABLEAUX

We first consider a supertableau T of the class Δ_0 with n rows and m columns as discussed in Sec. III. In general, T is typical and irreducible but it may accidentally become atypical when some specific relation between one row parameter b_j $(\bar{b}_j)$ and one column parameter $c_k(\bar{c}_k)$ is satisfied. Of course such a situation affects only mixed supertableaux. Consider a particular relation of atypicity

$$b_j - \bar{c}_k = n_{jk}, \tag{G1}$$

and suppose that the supertableau T of the class Δ_0 is typical and such that

$$b_j - \bar{c}_k = n_{jk} - 1.$$

In the tensor product of T by the fundamental one covariant box supertableau F we have in general two supertableaux T_1 and T_2 and two supertableaux only such that

for T_1, $b'_j = b_j + 1$ and all other parameters as in T,

for T_2, $\bar{c}'_k = \bar{c}_k - 1$ and all other parameters as in T.

It is clear that for T_1 and T_2 the constraint (G1) is fulfilled and both supertableaux are atypical. We describe this result as the pair production of atypical supertableaux.

In the particular situations for the supertableau T

$$\left|\begin{array}{l} b_{j-1} = b_j \\ \bar{c}_{k+1} = \bar{c}_k \end{array}\right| \text{ the supertableau } \left|\begin{array}{l} T_1 \\ T_2 \end{array}\right| \text{ does not exist.}$$

However it can be proven that even in these cases a second supertableau becomes atypical in the product $T\otimes F$ with a different atypicity constraint. As discussed in I the supertableaux T_1 and T_2 do not have, in general, an individual existence and only the consideration of the pair (T_1, T_2) makes sense. This pair by definition is a two-generalized supertableau and it describes a nonfully reducible representation of $SU(n|m)$.

If the supertableau T is chosen in such a way that the same procedure can be repeated for a different atypicity relation we can construct two supertableaux T_{11} and T_{12} from T_1 and two supertableaux T_{21} and T_{22} from T_2. These four supertableaux are atypical with a degeneracy of atypicity $\delta = 2$. Only the consideration of the quartet $(T_{11}, T_{12}, T_{21}, T_{22})$ makes sense and by definition this collection is a four-generalized supertableau and it describes a nonfully reducible representation of $SU(n|m)$.

Such a construction can be made for all subsets of atypical relations (G1). The full set of atypical relations has obviously the dimension L and therefore we can obtain collections of 2^p atypical supertableaux with degeneracy of atypicity $\delta = p$ as long as $p \leqslant L$. This collection is a ρ-generalized supertableau and we have

$$\rho = 2^\delta, \quad \text{with } 1 \leqslant \delta \leqslant L. \tag{G2}$$

The extension of these considerations to the atypical supertableaux of the class Δ_l for $1 \leqslant l \leqslant L-1$ is straightforward. The original supertableau is an atypical irreducible supertableau of the class Δ_l whose highest weight has a degeneracy of atypicity $\delta = l$. Therefore l atypical relations of the type (G1) have already been used and we have at our disposal only $(L - l)$ independent atypical relations. We then use the same method of construction as previously and we get ρ-generalized supertableaux of the class Δ_l with ρ now given by

$$\rho = 2^{(\delta - l)}, \quad \text{with } l < \delta \leqslant L. \tag{G3}$$

[1] A. B. Balantekin and I. Bars, J. Math. Phys. **22**, 1149,1810 (1981); **23**, 1239 (1982); A. B. Balantekin, J. Math. Phys. **23**, 486 (1982).

[2] I. Bars, B. Morel, and H. Ruegg, J. Math. Phys. **24**, 2253 (1983).

[3] V. G. Kac, Adv. Math. **26**, 1 (1977); Commun. Math. Phys. **53**, 31 (1977); Lect. Notes Math. **676**, 597 (1978).

[4] S. Hongzhou and H. Qizhi, Sci. Sin. **24**, 914 (1981); Ph. Energia Fortis **5**, 646 (1981); **6**, 317, 401; (1982); J. Thierry Mieg and B. Morel, in *Superspace and Supergravity*, edited by S. Hawking and M. Rocek (Cambridge U.P., London, 1981); J. P. Hurni and B. Morel, J. Math. Phys. **24**, 157 (1983).

[5] F. Delduc and M. Gourdin, J. Math. Phys. **25**, 1651 (1984).

[6] A. B. Balantekin and I. Bars, J. Math. Phys. **22**, 1810 (1981).

[7] M. Gourdin, Preprint Paris LPTHE 84-16, 84-31.

[8] M. Hamermesh, *Group Theory and its Applications to Physical Processes* (Addison–Wesley, Reading, MA, 1962); M. Gourdin, *Basics of Lie Groups* (Editions Frontiéres, Paris, 1982).

Quantum-mechanical representations of the group of diffeomorphisms and local current algebra describing tightly bound composite particles

Gerald A. Goldin
Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115 and Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Ralph Menikoff
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 23 January 1985; accepted for publication 5 April 1985)

A semidirect product of Schwartz' space functions on $\mathbb{R}^3$ and the group of diffeomorphisms of $\mathbb{R}^3$ can describe quantum-mechanical systems. We interpret a class of continuous unitary representations of this group, characterized by multipole moments, as describing tightly bound composite particles.

I. INTRODUCTION[1–4]

Nonrelativistic quantum theory may be described by means of unitary representations of the group of diffeomorphisms of $\mathbb{R}^3$. In such a representation, the self-adjoint generators form a Lie algebra of local currents, indexed by vector fields on $\mathbb{R}^3$. In this paper, we examine representations of the local current algebra which are inequivalent to the N-particle representations previously studied, and which describe tightly bound composite particles.

At a fixed time t, the charge density $\rho(\mathbf{x})$ and the momentum density $\mathbf{J}(\mathbf{x})$ are operator-valued distributions. That is, the averaged currents $\rho(f) = \int \rho(x) f(x) d^3x$ and $J(\mathbf{g}) = \int \mathbf{J}(\mathbf{x}) \cdot \mathbf{g}(\mathbf{x}) d^3x$ are self-adjoint operators in a Hilbert space $\mathscr{H}$, where f and the components of $\mathbf{g}$ are C^∞ functions of rapid decrease on $\mathbb{R}^3$ (Schwartz' space functions). These operators satisfy the equal-time commutation relations (with $\hbar = 1$)

$$[\rho(f_1), \rho(f_2)] = 0, \tag{1}$$

$$[\rho(f), J(\mathbf{g})] = i\rho(\mathbf{g} \cdot \nabla f), \tag{2}$$

$$[J(\mathbf{g}_1), J(\mathbf{g}_2)] = iJ(\mathbf{g}_2 \cdot \nabla \mathbf{g}_1 - \mathbf{g}_1 \cdot \nabla \mathbf{g}_2), \tag{3}$$

where $\mathbf{g}_2 \cdot \nabla \mathbf{g}_1 - \mathbf{g}_1 \cdot \nabla \mathbf{g}_2$ is the Lie bracket $[\mathbf{g}_1, \mathbf{g}_2]$ of the vector fields $\mathbf{g}_1$ and $\mathbf{g}_2$. Thus Eqs. (1)–(3) represent the Lie algebra of C^∞ scalars and vector fields on $\mathbb{R}^3$ by self-adjoint operators.

The associated one-parameter unitary groups are $U(sf) = \exp[is\rho(f)]$ and $V(\boldsymbol{\phi}_s^{\mathbf{g}}) = \exp[isJ(\mathbf{g})]$, where $s \in R$ and $\boldsymbol{\phi}_s^{\mathbf{g}}: \mathbb{R}^3 \to \mathbb{R}^3$ is the flow generated by $\mathbf{g}$; i.e., $\partial_s \boldsymbol{\phi}_s^{\mathbf{g}} = \mathbf{g} \circ \boldsymbol{\phi}_s^{\mathbf{g}}$ and $\boldsymbol{\phi}_{s=0}^{\mathbf{g}}(\mathbf{x}) = \mathbf{x}$. The operators $U(f)V(\boldsymbol{\phi})$ represent the semidirect product group $\mathscr{S} \wedge \mathscr{K}$, where $\mathscr{S}$ is Schwartz' space under addition, $\mathscr{K}$ is the group of C^∞ diffeomorphisms which rapidly tend to the identity at infinity, and the group law is given by

$$(f_1, \boldsymbol{\phi}_1)(f_2, \boldsymbol{\phi}_2) = (f_1 + f_2 \circ \boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \circ \boldsymbol{\phi}_1). \tag{4}$$

The representations of Eq. (1)–(3) that we study are obtained from continuous unitary representations of $\mathscr{S} \wedge \mathscr{K}$.

In Sec. II we present explicit representations which we interpret as describing systems of composite particles having an internal degree of freedom, the dipole moment. Section III introduces representations describing quadrupole particles and explains their interpretation as describing composite systems, concluding with a brief discussion of higher multipole representations. In Sec. IV, we elaborate further on our interpretation of these representations as describing kinematically bound particles.

II. REPRESENTATIONS DESCRIBING DIPOLE PARTICLES

First we consider a single dipole particle. Let $\mathscr{H}$ be the Hilbert space of square integrable functions $\psi(\mathbf{x},\boldsymbol{\lambda})$ with respect to Lebesgue measure on $(\mathbf{x},\boldsymbol{\lambda})$ space. We think of $\mathbf{x}$ as the position coordinate, and $\boldsymbol{\lambda}$ as the dipole moment coordinate. The group representation is defined for $(f,\boldsymbol{\phi}) \in \mathscr{S} \wedge \mathscr{K}$ by

$$(U(f)\psi)(\mathbf{x},\boldsymbol{\lambda}) = \exp[i\boldsymbol{\lambda} \cdot (\nabla f)(\mathbf{x})]\psi(\mathbf{x},\boldsymbol{\lambda}), \tag{5}$$

$$(V(\boldsymbol{\phi})\psi)(\mathbf{x},\boldsymbol{\lambda}) = \psi(\boldsymbol{\phi}(\mathbf{x}), \mathscr{J}_{\boldsymbol{\phi}}(\mathbf{x})\boldsymbol{\lambda}) \det \mathscr{J}_{\boldsymbol{\phi}}(\mathbf{x}), \tag{6}$$

where $\mathscr{J}_{\boldsymbol{\phi}}(\mathbf{x})$ is the Jacobian matrix of $\boldsymbol{\phi}$ at $\mathbf{x}$, and $\boldsymbol{\lambda}' = \mathscr{J}_{\boldsymbol{\phi}}(\mathbf{x})\boldsymbol{\lambda}$ is given by $(\lambda')^k = (\partial_j \phi^k)(\mathbf{x})\lambda^j$, summation over repeated indices being assumed. One can verify directly that Eqs. (5) and (6) provide a representation of Eq. (4). Because $(\mathbf{x},\boldsymbol{\lambda})$ space is a single orbit under $\mathscr{K}$, this representation is irreducible.

The representation of the current algebra, Eqs. (1)–(3), is easily found using $\rho(f)\psi = i^{-1}\partial_s U(sf)\psi|_{s=0}$ and $J(\mathbf{g})\psi = i^{-1}\partial_s V(\boldsymbol{\phi}_s^{\mathbf{g}})\psi|_{s=0}$. We obtain

$$(\rho(f)\psi)(\mathbf{x},\boldsymbol{\lambda}) = [\boldsymbol{\lambda} \cdot (\nabla f)(\mathbf{x})]\psi(\mathbf{x},\boldsymbol{\lambda}), \tag{7}$$

$$\begin{aligned}(J(\mathbf{g})\psi)(\mathbf{x},\boldsymbol{\lambda}) &= (1/2i)[\mathbf{g}(\mathbf{x}) \cdot \nabla + \nabla \cdot \mathbf{g}(\mathbf{x})]\psi(\mathbf{x},\boldsymbol{\lambda}) \\ &\quad + (\partial_j g^k)(\mathbf{x}) \frac{1}{2i}\left[\lambda^j \frac{\partial}{\partial \lambda^k} + \frac{\partial}{\partial \lambda^k}\lambda^j\right]\psi(\mathbf{x},\boldsymbol{\lambda}).\end{aligned} \tag{8}$$

In Eq. (8), the derivatives act on all quantities to their right.

To interpret this representation, consider two ordinary particles of equal and opposite charge q. The two-particle representation of $\mathscr{S} \wedge \mathscr{K}$ is given by[2]

$$(U(f)\Phi)(\mathbf{x}_1,\mathbf{x}_2) = \exp\{iq[f(\mathbf{x}_1) - f(\mathbf{x}_2)]\}\Phi(\mathbf{x}_1,\mathbf{x}_2), \tag{9}$$

$$\begin{aligned}(V(\boldsymbol{\phi})\Phi)(\mathbf{x}_1,\mathbf{x}_2) &= \Phi(\boldsymbol{\phi}(\mathbf{x}_1),\boldsymbol{\phi}(\mathbf{x}_2)) \\ &\quad \times [\det \mathscr{J}_{\boldsymbol{\phi}}(\mathbf{x}_1) \det \mathscr{J}_{\boldsymbol{\phi}}(\mathbf{x}_2)]^{1/2},\end{aligned} \tag{10}$$

 0022-2488/85/081880-05$02.50

where Φ is a square-integrable function of the particle coordinates $(\mathbf{x}_1,\mathbf{x}_2)$. Suppose that the particle separation is small so that $f(\mathbf{x})$ varies slowly between $\mathbf{x}_1$ and $\mathbf{x}_2$; that is, suppose that $\Phi(\mathbf{x}_1,\mathbf{x}_2) = 0$, unless $\mathbf{x}_1$ and $\mathbf{x}_2$ are close in this sense. Then from Eq. (9),

$$(\rho(f)\Phi)(\mathbf{x}_1,\mathbf{x}_2) = q[f(\mathbf{x}_1) - f(\mathbf{x}_2)]\Phi(\mathbf{x}_1,\mathbf{x}_2) \approx [\boldsymbol{\lambda}\cdot(\nabla f)(\mathbf{x})]\Phi(\mathbf{x}_1,\mathbf{x}_2), \qquad (11)$$

where $\lambda = q(\mathbf{x}_1 - \mathbf{x}_2)$ and $\mathbf{x} = \frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2)$. Thus we obtain Eq. (7) in this approximation, with $\psi(\mathbf{x},\boldsymbol{\lambda}) = \Phi(\mathbf{x}_1,\mathbf{x}_2)$. Similarly, we have for small particle separations, $\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2) \approx \mathscr{J}_\phi(\mathbf{x})[\mathbf{x}_1 - \mathbf{x}_2]$. Then from Eq. (10),

$$\begin{aligned}(V(\boldsymbol{\phi})\psi)(\mathbf{x},\boldsymbol{\lambda}) &= \psi(\tfrac{1}{2}[\boldsymbol{\phi}(\mathbf{x}_1) + \boldsymbol{\phi}(\mathbf{x}_2)],\ q[\boldsymbol{\phi}(\mathbf{x}_1) - \boldsymbol{\phi}(\mathbf{x}_2)]) \\ &\quad\cdot [\det \mathscr{J}_\phi(\mathbf{x}_1)\det \mathscr{J}_\phi(\mathbf{x}_2)]^{1/2} \\ &\approx \psi(\boldsymbol{\phi}(\mathbf{x}),\ \mathscr{J}_\phi(\mathbf{x})\boldsymbol{\lambda})\det \mathscr{J}_\phi(\mathbf{x}), \qquad (12)\end{aligned}$$

thus approximating Eq. (6). We see that for *any particular* choice $(f,\boldsymbol{\phi})$, the dipole representation $U(f)V(\boldsymbol{\phi})$ of Eqs. (5) and (6) approximates the two-particle representation for wave functions describing sufficiently tightly bound particles. However, the dipole representation and the two-particle representation of the local currents are unitarily inequivalent, and thus describe physically distinct systems.

Note that the dipole moment variable $\boldsymbol{\lambda}$ is here a continuous parameter, ranging over $\mathbb{R}^3\setminus\{0\}$. The wave function $\psi(\mathbf{x},\boldsymbol{\lambda})$ may be thought of as a probability amplitude for finding a neutral particle at $\mathbf{x}$ with dipole moment $\boldsymbol{\lambda}$. In a stationary state with no external fields, the expectation value of $\boldsymbol{\lambda}$ will be zero.

We can easily generalize Eqs. (5) and (6) to representations of $\mathscr{S}\wedge\mathscr{K}$ describing N identical dipole particles, by taking symmetric or antisymmetric tensor products. We obtain

$$\begin{aligned}[U(f)\psi](\mathbf{x}_1,\boldsymbol{\lambda}_1;\ldots;\mathbf{x}_N,\boldsymbol{\lambda}_N) &= \exp\Big[i\sum_{j=1}^{N}\boldsymbol{\lambda}_j\cdot(\nabla f)(\mathbf{x}_j)\Big]\psi(\mathbf{x}_1,\boldsymbol{\lambda}_1;\ldots;\mathbf{x}_N,\boldsymbol{\lambda}_N), \qquad (13)\\ [V(\boldsymbol{\phi})\psi](\mathbf{x}_1,\boldsymbol{\lambda}_1;\ldots;\mathbf{x}_N,\boldsymbol{\lambda}_N) &= \psi(\boldsymbol{\phi}(\mathbf{x}_1),\mathscr{J}_\phi(\mathbf{x}_1)\boldsymbol{\lambda}_1;\ldots;\boldsymbol{\phi}(\mathbf{x}_N),\mathscr{J}_\phi(\mathbf{x}_N)\boldsymbol{\lambda}_N) \\ &\quad\times\prod_{j=1}^{N}\det\mathscr{J}_\phi(\mathbf{x}_j), \qquad (14)\end{aligned}$$

where ψ is either totally symmetric or antisymmetric under exchanges $(\mathbf{x}_j,\boldsymbol{\lambda}_j)\leftrightarrow(\mathbf{x}_k,\boldsymbol{\lambda}_k)$. Both the position and dipole moment coordinates must be exchanged in defining the symmetry of ψ; thus $\boldsymbol{\lambda}$ acts like an internal degree of freedom of a composite particle.

Representations of $\mathscr{S}\wedge\mathscr{K}$ can also describe particles having a fixed net charge e as well as a variable dipole moment. One simply takes

$$(U(f)\psi)(\mathbf{x},\boldsymbol{\lambda}) = \exp[ief(\mathbf{x}) + i\boldsymbol{\lambda}\cdot(\nabla f)(\mathbf{x})]\psi(\mathbf{x},\boldsymbol{\lambda}), \qquad (15)$$

with $V(\boldsymbol{\phi})$ as in Eq. (6). Now the charge density operator becomes $\rho(f) = ef(\mathbf{x}) + \boldsymbol{\lambda}\cdot(\nabla f)(\mathbf{x})$ in the single-particle representation, with the obvious generalization to N particles.

III. QUADRUPOLE AND HIGHER MULTIPOLE PARTICLES

A quadrupole particle may be represented in the Hilbert space of functions $\psi(\mathbf{x},\boldsymbol{\lambda},Q)$, square integrable with respect to a measure ν, where $\mathbf{x}$ is the particle coordinate, $\boldsymbol{\lambda}$ is the dipole moment vector, and the 3×3 quadrupole matrix Q is a symmetric tensor. (Unlike the standard quadrupole matrix, Q is not required here to be traceless.) For a diffeomorphism $\boldsymbol{\phi}\in\mathscr{K}$, we have the transformation $\boldsymbol{\phi}$: $(\mathbf{x},\boldsymbol{\lambda},Q)\to(\mathbf{x}',\boldsymbol{\lambda}',Q')$ given by

$$\mathbf{x}' = \boldsymbol{\phi}(\mathbf{x}), \qquad (16)$$

$$(\lambda')^k = (\partial_j\phi^k)(\mathbf{x})\lambda^j + \tfrac{1}{2}Q^{mn}(\partial_m\,\partial_n\phi^k)(\mathbf{x}), \qquad (17)$$

$$(Q')^{mn} = (\partial_j\phi^m)(\mathbf{x})(\partial_k\phi^n)(\mathbf{x})Q^{jk}. \qquad (18)$$

Note that if $Q = 0$, Eqs. (16)–(18) reduce to the transformation law $\boldsymbol{\phi}$: $(\mathbf{x},\boldsymbol{\lambda})\to(\mathbf{x}',\boldsymbol{\lambda}')$ for a dipole particle.

The action of the group $\mathscr{S}\wedge\mathscr{K}$ is given by

$$(U(f)\psi)(\mathbf{x},\boldsymbol{\lambda},Q) = \exp[i\lambda^j(\partial_j f)(\mathbf{x}) + (i/2)Q^{mn}(\partial_m\,\partial_n f)(\mathbf{x})]\,\psi(\mathbf{x},\boldsymbol{\lambda},Q), \qquad (19)$$

$$(V(\boldsymbol{\phi})\psi)(\mathbf{x},\boldsymbol{\lambda},Q) = \psi(\mathbf{x}',\boldsymbol{\lambda}',Q')\Big[\frac{d\nu'}{d\nu}(\mathbf{x},\boldsymbol{\lambda},Q)\Big]^{1/2}, \qquad (20)$$

where ν' is the transformed measure, and $d\nu'/d\nu$ is the Radon–Nikodym derivative of ν' with respect to ν. Thus ν' and ν are required to have the same measure zero sets for all $\boldsymbol{\phi}\in\mathscr{K}$; i.e., ν is quasi-invariant under the action of $\mathscr{K}$. Since $(\mathbf{x},\boldsymbol{\lambda},Q)$ space is finite dimensional, ν can be chosen to be a Lebesgue measure. One can verify directly that Eqs. (19) and (20) provide a representation satisfying Eq. (4).

In general, the representation given by Eqs. (19) and (20) may have nontrivial invariant subspaces, and thus be reducible. To decompose it into irreducible representations, we need to consider the orbit structure under $\mathscr{K}$ of the space $\{\mathbf{x},\boldsymbol{\lambda},Q\}$, which is a 12-dimensional manifold. For any fixed $Q\neq0$, it is easy to see that $\mathscr{K}$ acts transitively on $\mathbf{x}$ and $\boldsymbol{\lambda}$, as follows. Given $(\mathbf{x},\boldsymbol{\lambda},Q)$ and arbitrary $\mathbf{x}'$, choose $\boldsymbol{\phi}$ so that $\mathbf{x}' = \boldsymbol{\phi}(\mathbf{x})$, while $\mathscr{J}_\phi$ is the identity matrix in a neighborhood of $\mathbf{x}$; a "local translation" accomplishes this. Thus $\boldsymbol{\phi}$: $(\mathbf{x},\boldsymbol{\lambda},Q)\to(\mathbf{x}',\boldsymbol{\lambda},Q)$. Likewise, given $(\mathbf{x},\boldsymbol{\lambda},Q)$ and arbitrary $\boldsymbol{\lambda}'$, choose a coordinate system in which Q is diagonal and $Q^{11}\neq0$; then choose $\boldsymbol{\phi}$ so that $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$, $\mathscr{J}_\phi(\mathbf{x}) = I$, and $(\partial_1^2\phi)(\mathbf{x}) = (Q^{11})^{-1}(\boldsymbol{\lambda}' - \boldsymbol{\lambda})$, with all other second derivatives vanishing at $\mathbf{x}$. Then we obtain $Q^{mn}(\partial_m\,\partial_n\phi^k)\times(\mathbf{x}) = (\lambda')^k - \lambda^k$, and $\boldsymbol{\phi}$: $(\mathbf{x},\boldsymbol{\lambda},Q)\to(\mathbf{x},\boldsymbol{\lambda}',Q)$.

Thus it remains only to describe the orbit structure of the six-dimensional manifold of Q's, which is summarized in the following theorem.

Theorem 1: The manifold $\mathscr{Q}$ of nonzero 3×3 symmetric real matrices is partitioned into mutually disjoint orbits under the action of the diffeomorphism group $\mathscr{K}$ given by Eq. (18), according to the signs of the eigenvalues of the matrices. That is, two elements of $\mathscr{Q}$ are in the same orbit if and only if they have the same number of positive, negative, and zero eigenvalues.

Proof: First we show that $\mathscr{J}_\phi(\mathbf{x})$ can always be written as a dilation followed by a rotation, or equally well as a rotation followed by a dilation. The most general dilation is of the form $R^{-1}D_+R$, where $R\in SO(3)$ and D_+ is diagonal with positive entries, i.e., R chooses a coordinate frame in which the axes are each dilated. Thus we shall show that $\mathscr{J}_\phi(\mathbf{x})$ has the form $R_1D_+R_2$ for $R_1,R_2\in SO(3)$. Letting $A=\mathscr{J}_\phi(\mathbf{x})^t \times \mathscr{J}_\phi(\mathbf{x})$, we have that $A^t=A$. Also, A is positive definite: for any $\mathbf{v}\in\mathbb{R}^3$, $(\mathbf{v},A\mathbf{v})=(\mathscr{J}_\phi(\mathbf{x})\mathbf{v},\mathscr{J}_\phi(\mathbf{x})\mathbf{v})>0$ since $\det \mathscr{J}_\phi(\mathbf{x})\neq 0$. Hence, A has the form $R_2^{-1}D_+^2R_2$ for some $R_2\in SO(3)$ and some D_+^2 diagonal and positive definite. Now let $K=\mathscr{J}_\phi(\mathbf{x})R_2^{-1}D_+^{-1}R_2$. It is straightforward to show that $K^tK=I$ using the fact that $R_2^t=R_2^{-1}$. Thus $K\in SO(3)$, and $\mathscr{J}_\phi(\mathbf{x})=KR_2^{-1}D_+R_2=R_1D_+R_2$ as desired.

Next, we show that D_+QD_+ has eigenvalues with the same signs as Q. Let $D_+(t)=I+(D_+-I)t$, for $t=[0,1]$. Let the eigenvalues of Q be $(\alpha_1,\alpha_2,\alpha_3)$, some of which may be zero. Then the eigenvalues $(\alpha_{1t},\alpha_{2t},\alpha_{3t})$ of $D_+(t)QD_+(t)$ are continuous functions of t, and equal $(\alpha_1,\alpha_2,\alpha_3)$ when $t=0$. Let η_t be the subspace of $\mathbb{R}^3$ on which $D_+(t)QD_+(t)$ vanishes, and η_0 the subspace on which Q vanishes. It is easy to see that $D_+(t)$: $\eta_t\to\eta_0$ and $D_+(t)^{-1}$: $\eta_0\to\eta_t$; thus $\dim\eta_0=\dim\eta_t$, for all t. Hence, $\operatorname{sgn}(\alpha_{jt})=\operatorname{sgn}(\alpha_j)$, for $j=1,2,3$.

Now from Eq. (18), $Q'=\mathscr{J}_\phi(\mathbf{x})Q\mathscr{J}_\phi(\mathbf{x})^t=R_1D_+R_2\times QR_2^{-1}D_+R_1^{-1}$. But the eigenvalues of Q are the same as those of $R_2QR_2^{-1}$, and therefore have the same signs as those of Q'. We thus see that Eq. (18) preserves the signs of the eigenvalues. Q.E.D.

Conversely, let Q_1 and Q_2 have eigenvalues $(\alpha_1,\alpha_2,\alpha_3)$ and $(\beta_1,\beta_2,\beta_3)$ of corresponding signs. Let $\gamma_j=\beta_j/\alpha_j$ when $\operatorname{sgn}(\alpha_j)=\operatorname{sgn}(\beta_j)\neq 0$, and $\gamma_j=1$ when $\alpha_j=\beta_j=0$. Let D_+^2 be the diagonal matrix with positive entries $(\gamma_1,\gamma_2,\gamma_3)$. We can diagonalize Q_1 and Q_2 by writing $Q_1=R_1^{-1}D_1R_1$ and $Q_2=R_2^{-1}D_2R_2$, and we define $\mathscr{J}=R_2^{-1}D_+R_1$. Then $\mathscr{J}Q_1\mathscr{J}^t=R_2^{-1}D_+D_1D_+R_2=Q_2$. Thus the matrix $\mathscr{J}$ connects Q_1 and Q_2. (Note that a rotation can exchange the positions of eigenvalues in a diagonal matrix, so that their order is arbitrary in the preceding, as long as eigenvalues of the same sign are chosen to correspond to each other.) Finally, we observe that for fixed x there exists a diffeomorphism $\phi\in\mathscr{K}$ such that $(\partial_j\phi^k)(\mathbf{x})=\mathscr{J}$ for any $\mathscr{J}$ of the form $R_2^{-1}\times D_+R_1$. Hence Q_1 and Q_2 lie on the same orbit if and only if they have the same numbers of positive, negative, and zero eigenvalues.

We thus have nine orbits in $\mathscr{Q}$, as follows. When all eigenvalues are nonzero ($\det Q\neq 0$), the orbits are six dimensional. There are four possibilities: (1) all eigenvalues positive; (2) one positive, two negative; (3) two positive, one negative; and (4) all eigenvalues negative. When one eigenvalue is zero, the orbits are five dimensional, and the possibilities are (5) two eigenvalues positive; (6) one positive, one negative; and (7) two eigenvalues negative. Finally, when two eigenvalues are zero the orbits are three dimensional, and the possibilities are (8) one eigenvalue positive, and (9) one eigenvalue negative.

Each orbit in $\mathscr{Q}$ now corresponds to a distinct irreducible quadrupole representation of the group $\mathscr{S}\wedge\mathscr{K}$, where ν in Eqs. (19) and (20) is chosen to be concentrated on that orbit. Since each orbit is a finite-dimensional manifold and ν is quasi-invariant, the measure class of ν for each orbit is unique.

In analogy with the dipole case, it is possible to interpret the quadrupole representations of $\mathscr{S}\wedge\mathscr{K}$ as "limits" of N-particle representations in which the charges become infinite and the separations tend toward zero. We shall consider one case in detail, that of an orbit in which the quadrupole matrix has just one nonzero eigenvalue. Consider a three-particle configuration in which a particle of negative charge $-q_3$ is located at $\mathbf{x}_3\in\mathbb{R}^3$, and particles of positive charge q_1 and q_2 have coordinates $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. For the composite to be neutral, we take $q_3=q_1+q_2$. Let us take moments about the point $\mathbf{x}_3$. To shrink the system and obtain a finite nonzero quadrupole moment, set $\mathbf{x}_i-\mathbf{x}_3=\delta_i\epsilon+\eta_i\epsilon^2$, $i=1,2$, and set $q_i=q_{0i}/\epsilon^2$, $i=1,2,3$, with $\boldsymbol{\delta}_i$, $\boldsymbol{\eta}_i$, and q_{0i} finite; then let $\epsilon\to 0$. The dipole moment is

$$\boldsymbol{\lambda}=\frac{1}{\epsilon}\sum_{i=1}^{2}\boldsymbol{\delta}_iq_{0i}+\sum_{i=1}^{2}\boldsymbol{\eta}_iq_{0i}.$$

For $\boldsymbol{\lambda}$ to be finite as $\epsilon\to 0$, we need $\Sigma_{i=1}^2\boldsymbol{\delta}_iq_{0i}=0$; that is, *the particles are to first order colinear*. Taking them to lie to first order on the x^1 axis, with $\mathbf{x}_3=\mathbf{0}$, we have $Q^{11}=\Sigma_{i=1}^2q_{0i}|\boldsymbol{\delta}_i|^2>0$, and $\boldsymbol{\lambda}=\Sigma_{i=1}^2q_{0i}\boldsymbol{\eta}_i$. Note that an arbitrary dipole moment can result from the *second-order* term in the position coordinates, describing a small noncolinearity. Next, the charge density is given by

$$\begin{aligned}(\rho(f)\psi)(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)&=q_1f(\mathbf{x}_1)+q_2f(\mathbf{x}_2)-q_3f(\mathbf{x}_3)\\&=\frac{q_{01}}{\epsilon^2}f(\mathbf{x}_3+\boldsymbol{\delta}_1\epsilon+\boldsymbol{\eta}_1\epsilon^2)+\frac{q_{02}}{\epsilon^2}f(\mathbf{x}_3+\boldsymbol{\delta}_2\epsilon+\boldsymbol{\eta}_2\epsilon^2)\\&\quad-[(q_{01}+q_{02})/\epsilon^2]\,f(\mathbf{x}_3).\end{aligned}$$

Keeping all terms to second order in ϵ, we have

$$\begin{aligned}&f(\mathbf{x}+\boldsymbol{\delta}\epsilon+\boldsymbol{\eta}\epsilon^2)-f(\mathbf{x})\\&\quad\approx\epsilon\delta^j(\partial_jf)(\mathbf{x})+\epsilon^2\eta^j(\partial_jf)(\mathbf{x})+\tfrac{1}{2}\epsilon^2\delta^m\delta^n(\partial_m\partial_nf)(\mathbf{x}),\end{aligned}$$

and thus

$$\begin{aligned}\rho(f)\psi\approx&\sum_{i=1}^{2}\tfrac{1}{\epsilon}q_{0i}\delta_i^j(\partial_jf)(\mathbf{x})\psi\\&+\sum_{i=1}^{2}q_{0i}\eta_i^j(\partial_jf)(\mathbf{x})\psi\\&+\tfrac{1}{2}\sum_{i=1}^{2}q_{0i}\delta_i^m\delta_i^n(\partial_m\,\partial_nf)(\mathbf{x})\psi.\end{aligned}\tag{21}$$

This is in agreement with Eq. (19), where $U(f)=\exp[i\rho(f)]$, if we impose

$$\sum_{i=1}^{2}q_{0i}\boldsymbol{\delta}_i=\mathbf{0},\quad\sum_{i=1}^{2}q_{0i}\boldsymbol{\eta}_i=\boldsymbol{\lambda},\quad\sum_{i=1}^{2}q_{0i}\delta_i^m\delta_i^n=Q^{mn}.$$

Now consider the action of a diffeomorphism $\boldsymbol{\phi}$: $(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)\to(\boldsymbol{\phi}(\mathbf{x}_1),\boldsymbol{\phi}(\mathbf{x}_2),\boldsymbol{\phi}(\mathbf{x}_3))$. We have

$$\begin{aligned}&\phi^k(\mathbf{x}+\boldsymbol{\delta}\epsilon+\boldsymbol{\eta}\epsilon^2)\\&\quad\approx\phi^k(\mathbf{x})+\epsilon\delta^j(\partial_j\phi^k)(\mathbf{x})+\epsilon^2\eta^j(\partial_j\phi^k)(\mathbf{x})\\&\qquad+\tfrac{1}{2}\epsilon^2\delta^m\delta^n(\partial_m\partial_n\phi^k)(\mathbf{x}),\end{aligned}$$

whence

$$\begin{aligned}&\mathbf{x}'=\boldsymbol{\phi}(\mathbf{x}),\quad(\delta')^k=(\partial_j\phi^k)(\mathbf{x})\delta^j,\\&(\eta')^k=(\partial_j\phi^k)(\mathbf{x})\eta^j+\tfrac{1}{2}\delta^m\delta^n(\partial_m\,\partial_n\phi^k)(\mathbf{x}).\end{aligned}$$

With

$$\lambda' = \sum_{i=1}^{2} q_{0i}\eta_i, \quad (Q')^{m,n} = \sum_{i=1}^{2} q_{0i}(\delta_i')^m(\delta_i')^n,$$

we see that Eqs. (16)–(18) are recovered.

Thus, the representation of $\mathscr{S} \wedge \mathscr{K}$ based on a quadrupole orbit with one nonzero eigenvalue can be understood as a quantum theory of tightly bound nearly colinear three-particle configurations having zero net charge. The sign of the eigenvalue is opposite to the charge of the central, differently charged particle, $\mathbf{x}$ is the coordinate of that particle, and λ and Q are the dipole and quadrupole moments about $\mathbf{x}$. The nine degrees of freedom of the three component particles agree with the nine dimensions of the orbit in $(\mathbf{x},\lambda,Q)$ space; that is, the condition of near colinearity does not restrict the number of degrees of freedom of the system.

Similarly, a representation of $\mathscr{S} \wedge \mathscr{K}$ based on a quadrupole orbit with two nonzero eigenvalues can be understood as the "limit" of shrinking nearly coplanar four-particle configurations. That is, the fourth particle lies to first order in the plane of the first three, and the first-order dipole moment is constrained to be zero. An arbitrary value for the dipole moment now results from the second-order contribution. Such a configuration has 11 degrees of freedom, as follows. Fixing the charges of the particles so that they sum to zero, the coordinates of the first three particles determine to first order the coordinates of the fourth (so that the dipole moment vanishes). Thus far, we have nine independent coordinates. However, the description of such a configuration by means of a spatial coordinate $\mathbf{x}$ and a quadrupole moment Q about $\mathbf{x}$ results in only eight degrees of freedom: one-dimensional equivalence classes of coplanar particle configurations have the same quadrupole moment. Finally, three additional degrees of freedom result from including the second-order contribution to the dipole moment, making 11 degrees of freedom. In Theorem 2, we shall see that the three quadrupole orbits with two nonzero eigenvalues correspond to the three distinct ways in which the signs of the charges can be distributed among the particles: one negative and three positive charges corresponding to two positive eigenvalues for Q, one positive and three negative charges corresponding to two negative eigenvalues for Q, or two positive and two negative charges corresponding to eigenvalues of opposite sign.

Finally, the case in which all three eigenvalues are nonzero can be obtained as a "limit" in which the separations between five charged point particles tend toward zero. Five particles with fixed charges and arbitrary coordinates in $\mathbb{R}^3$ have 15 degrees of freedom. The condition that to first order the dipole moment is zero reduces the number of degrees of freedom to 12. For a given quadrupole moment Q having three nonzero eigenvalues, the particle configuration is not unique—there is a three-parameter family of configurations having the same quadrupole moment. This brings us down to nine degrees of freedom. Thus (as the previous case of two nonzero eigenvalues), in describing a "tightly bound composite" by means of its quadrupole matrix, some information about the structure of the components is lost. Introducing the second-order terms which restore an arbitrary finite dipole moment adds three degrees of freedom, yielding the 12 degrees of freedom corresponding to the dimensionality of the orbits in $(\mathbf{x},\lambda,Q)$ space. Theorem 2 demonstrates in parallel with the earlier cases that four quadrupole orbits with three nonzero eigenvalues correspond to the four distinct ways in which the signs of the charges can be distributed among the five particles.

Theorem 2: Consider a configuration of four coplanar but not colinear particles having fixed charges, where the total charge and dipole moment are zero. Then the number of eigenvalues of the quadrupole matrix with positive (negative) sign is one less than the number of positive (negative) charges. In a configuration of five noncoplanar charged particles, with total charge and dipole moment zero, the number of positive (negative) eigenvalues of the quadrupole matrix is likewise one less than the number of positive (negative) charges.

Proof: In both cases the quadrupole moment matrix is the same about any origin, because the total charge and dipole moment are zero. Suppose all charges but one are of the same sign. Then we can choose the origin to be at the location of the particle having unlike charge, and we can choose the coordinate axes so as to diagonalize the quadrupole matrix. We thus obtain exactly one zero eigenvalue in the four-particle coplanar case, and three nonzero eigenvalues in the five-particle case, with all nonzero eigenvalues having the same sign as that of the like charges.

Now suppose in the four-particle case that there are two positive and two negative charges. Consider the two coplanar line segments connecting the pairs of particles having like charge. Because the dipole moment is zero, these line segments intersect; choose the origin at their point of intersection. Now we can perform a dilation with respect to the coordinate system defined by the angle bisector and its normal in the plane, so that the transformed line segments become perpendicular. (We saw in Theorem 1 that a dilation does not change the signs of the eigenvalues.) The resulting quadrupole matrix has one positive and one negative eigenvalue.

Finally, in the five-particle case, consider the triangular region of a plane whose vertices are at the locations of three particles having like charge. The line segment connecting the other two particles intersects this triangle because the dipole moment is zero. Now we dilate with respect to the coordinate axes defined by the bisector of the angle between the line and the plane, its normal in the plane, and their normal. Such a dilation can transform the line segment and the plane so that they are perpendicular. Now the quadrupole eigenvalue defined by the direction normal to the plane has the sign of the two like charges, and the two eigenvalues corresponding to directions within the plane have the sign of the three like charges. Q.E.D

In a manner similar to our discussion of dipole and quadrupole representations, one can construct representations of $\mathscr{S} \wedge \mathscr{K}$ on spaces of functions whose arguments are higher-rank symmetric tensors. These describe more complicated composite systems having higher multipole moments. In each case, the key step is to "lift" the action of the diffeomorphism group to the appropriate space of tensor fields. For example, "octopole composites" are obtained from the transformation law:

$$\mathbf{x}' = \boldsymbol{\phi}(\mathbf{x}), \tag{22}$$

$$(\lambda')^k = (\partial_j \phi^k)(\mathbf{x})\lambda^j + \tfrac{1}{2}Q^{mn}(\partial_m \partial_n \phi^k)(\mathbf{x}) + \tfrac{1}{6}(\partial_p \partial_q \partial_r \phi^k)(\mathbf{x})O^{pqr}, \tag{23}$$

$$(Q')^{mn} = (\partial_j \phi^m)(\mathbf{x})(\partial_k \phi^n)(\mathbf{x})Q^{jk} + \tfrac{1}{6} \sum_{\substack{\text{cyclic perms}\\(pqr)}} \left\{ (\partial_p \partial_q \phi^m)(\mathbf{x})(\partial_r \phi^n)(\mathbf{x}) + (m \leftrightarrow n) \right\} O^{pqr}, \tag{24}$$

$$(O')^{pqr} = (\partial_j \phi^p)(\mathbf{x})(\partial_k \phi^q)(\mathbf{x})(\partial_l \phi^r)(\mathbf{x})O^{jkl}, \tag{25}$$

where O is the octopole moment tensor.

The pattern is evident: all indices of the highest multipole are acted on by the Jacobian matrix, while to each lower multipole there is a contribution obtained by contracting the higher tensors with the appropriate higher derivatives of $\boldsymbol{\phi}$, and symmetrizing where necessary.

IV. DISCUSSION AND INTERPRETATION

The representations $\mathscr{S} \wedge \mathscr{K}$ discussed in this paper arise naturally in an induced representation formalism.[3,5] The single-particle and N-particle Bose representations are induced by the identity representation of the stability group. Other representations of the stability group can be used to induce additional representations of $\mathscr{S} \wedge \mathscr{K}$ based on the multipole orbits. The N-particle Fermi representations are obtained in this way, as in Ref. 3. Together with other previously obtained representations of the diffeomorphism group describing particles with spin,[6,7] our present results illustrate how internal degrees of freedom arise kinematically in a quantum theory based on local currents, i.e., from the unitary representations of $\mathscr{S} \wedge \mathscr{K}$.

The interpretation of the particles as "composites" rest on a weak correspondence. Consider the matrix elements (i.e., the outcomes of measurements) for the operators $U(f)$ and $V(\boldsymbol{\phi})$, which we assume generate a complete set of observables. Physically, this assumption implies that the system has no additional degrees of freedom beyond those described by an irreducible representation. When evaluated for N-particle states in which the particles are close together relative to distances in which f and $\boldsymbol{\phi}$ change appreciably, these matrix elements approximate the matrix elements of $U(f)$ and $V(\boldsymbol{\phi})$ in the multipole representations. Furthermore, the multipole representations are inequivalent to the N-particle representations. Thus, the particles cannot be separated and the extra variables must correspond to internal degrees of freedom. Hence, the outcomes of measurements in the multipole representations are as if the particles were tightly bound composites of more elementary "components," which are not observable individually.

The existence of such representations is a mathematical property of the group of diffeomorphisms of $\mathbb{R}^3$, much as the existence of spinor representations is a property of the group of rotations of three-dimensional space, or its covering group SU(2). Thus, the multipole particles described here are kinematical, and are compatible with any choice of Hamiltonian that can be expressed in terms of the local currents, in exactly the same sense that particles with spin are predicted by group theory without prior commitment to a particular dynamics.

From the multipole representations a model could be constructed for a system of interacting composite particles. For a given representation a Hamiltonian can be written as a differential operator in the variables, including the internal degrees of freedom ($\mathbf{x}$, $\boldsymbol{\lambda}$, Q, etc.). In such a model the composites always remain bound and thus the binding energy is not defined. Energies relative to the ground state are meaningful. This model would be a reasonable approximation to the physical reality of composite particles when the excitation energy of the internal degrees of freedom is small compared with the physically observed binding energy, and the separation between composites is large compared with their physical size. It is not the purpose of this paper to construct a dynamical model, but rather to show how these systems arise kinematically from the representations in the local current description of quantum theory.

ACKNOWLEDGMENTS

The authors are grateful to D. H. Sharp for interesting discussions. G. G. thanks the Physics Department at Princeton University for its hospitality during his 1982–83 sabbatical leave as a Visiting Fellow.

[1] R. F. Dashen and D. H. Sharp, Phys. Rev. **165**, 1857 (1968).
[2] G. A. Goldin, J. Math. Phys. **12**, 462 (1971).
[3] G. A. Goldin, R. Menikoff, and D. H. Sharp, J. Math. Phys. **21**, 650 (1980).
[4] G. A. Goldin, R. Menikoff, and D. H. Sharp, J. Phys. A: Math. Gen. **16**, 1827 (1983).
[5] G. W. Mackey, *Induced Representations of Groups and Quantum Mechanics* (Benjamin, New York, 1968).
[6] G. A. Goldin, R. Menikoff, and D. H. Sharp, Phys. Rev. Lett. **51**, 2246 (1983).
[7] G. A. Goldin and D. H. Sharp, Commun. Math. Phys. **92**, 217 (1983).

On the Wigner coefficients of the generalized Lorentz groups in the parabolic basis

G. A. Kerimov
Institute of Physics of the Academy of Sciences of the AzSSR, Baku, USSR

(Received 20 June 1984; accepted for publication 6 December 1984)

Wigner coefficients for the class 1 representations of the generalized Lorentz groups SO$(n+1,1)$ in the parabolic basis corresponding to the group reduction SO$(n+1,1)\supset E(n)\supset T(n)$ are calculated. They are in general expressible in terms of Appell's F_4 hypergeometric functions of two variables. However, in the case of $n=1,2$ they can also be expressed in terms of ordinary hypergeometric functions ${}_2F_1$.

I. INTRODUCTION

The generalized Lorentz groups SO$(n+1,1)$ are the most important noncompact Lie groups used so far in various areas of elementary particle physics. One of the problems which has arisen in the applications of these groups to elementary particle physics is the explicit construction of the Clebsch–Gordan coefficients (or, equivalently, the Wigner coefficients) defined relative to some complete set of commuting observables. In a previous paper[1] we have calculated the Clebsch–Gordan coefficients for the class 1 representations of SO$(n+1,1)$ in the canonical basis corresponding to the group reduction SO$(n+1,1)\supset$SO$(n+1)\supset\cdots\supset$SO(2).

The present paper is devoted to the computation of the related Wigner coefficients in the parabolic basis corresponding to the group reduction SO$(n+1,1)\supset E(n)\supset T(n)$. With the choice of this basis the Wigner coefficients have especially simple forms. They are in general expressible in terms of Appell's F_4 hypergeometric functions of two variables. However, in the case of $n=1,2$ they can also be expressed in terms of ordinary hypergeometric functions ${}_2F_1$.

The content of the paper is arranged as follows. In Sec. II we establish notation and review the properties of SO$(n+1,1)$ needed in the sequel. In Sec. III we obtain an integral representation for the single Wigner coefficients of SO$(n+1,1)$. In Sec. IV we carry out the explicit calculation of the Wigner coefficients of SO(2,1). In Sec. V we discuss the Wigner coefficients of SO$(n+1,1)$ for $n>1$. In Sec. VI we determine some symmetry properties of these coefficients.

II. REVIEW

In order to fix notation and terminology we start with a brief description of a class 1 representation of the SO$(n+1,1)$ groups. As is well known,[2] the class 1 representations of SO$(n+1,1)$ can be realized in the space of infinitely differentiable homogeneous functions $F(k)$ on the upper sheet of the $(n+1)$-dimensional cone $k^2=k_1^2+\cdots+k_{n+1}^2-k_{n+2}^2=0, k_{n+2}>0$, with degree j (j is an arbitrary complex number)

$$F(ak)=a^jF(k)\,,\quad a>0\,. \tag{1}$$

The representations of SO$(n+1,1)$ are given by

$$Q^j(g)F(k)=F(g^{-1}k)\,,$$

where $g\in$SO$(n+1,1)$.

Generally, we may choose a large number of different coordinate systems on the cone. It is well known that different chains of coordinate systems on the cone lead to different reductions of the SO$(n+1,1)$ group to its subgroups. The parabolic basis for SO$(n+1,1)$ is given by the decomposition according to "parabolic subgroups," SO$(n+1,1)\supset E(n)\supset T(n)$, where by $E(n)$ and $T(n)$ we denote Euclidean and translation subgroups of SO$(n+1,1)$, respectively. As a prelude to this decomposition one introduces coordinates on the cone given by

$$k_i=\omega x_i\,,\quad i=1,2,\ldots,n\,, \tag{2}$$

$$k_{n+1}=\tfrac{1}{2}\omega(1-y)\,,\quad k_{n+2}=\tfrac{1}{2}\omega(1+y)\,, \tag{3}$$

where $y=x^2$, $x^2=x_1^2+x_2^2+\cdots+x_n^2$. The parameters x_i may be taken to vary in the following regions:

$$-\infty<x_i<\infty\,,\quad \omega>0\,.$$

From (1) it follows that the homogeneous function is defined uniquely by its values on the n-dimensional Euclidean space R_n. Consequently, the class 1 representations of SO$(n+1,1)$ can be realized on the space D_j of infinitely differentiable functions $f(x)$ on R_n. In this realization the representations of SO$(n+1,1)$ are given by

$$V^j(g)f(x)=(\omega'/\omega)^jf(x')\,, \tag{4}$$

where ω' and x' are defined from the parametrization (3) of $k'=g^{-1}k$. In particular for translation subgroup $T(n)$ we have

$$V^j(a)f(x)=f(x-a)\,, \tag{5}$$

$$a=\begin{pmatrix}1_n & a^t & a^t\\ -a & 1-\tfrac{1}{2}a^2 & -\tfrac{1}{2}a^2\\ a & \tfrac{1}{2}a^2 & 1+\tfrac{1}{2}a^2\end{pmatrix}\in T(n)\,, \tag{6}$$

where a represents both an n-dimensional vector and the corresponding matrix.

The representations (4), so defined, can be extended (by an appropriate completion of D_j) to unitary (irreducible) representations of SO$(n+1,1)$ for the following values of j.

(a) $j=-\tfrac{1}{2}+i\rho$, $0<\rho<\infty$ (principal series).

(b) j lies in the range $-n<j<0$ (supplementary series).

(c) $j=-n-1$, where 1 is a positive integer (discrete series).

Let us now give the explicit expression for the trilinear invariant functionals and their kernels that are related with

 0022-2488/85/081885-04$02.50

the problem of the tensor product decomposition of two representations (see, e.g., Chap. IV of Ref. 3)

$$\int_{R_n}\int_{R_n}\int_{R_n} K(x_1j_1,x_2j_2,x_3j_3)f_1(x_1)f_2(x_2) \times f_3(x_3)d^nx_1\,d^nx_2\,d^nx_3 = \text{inv.}, \tag{7}$$

where $f_i \in D_{j_i}$. If all three j_i belong to the principal series, K is, up to a constant, uniquely determined and is given by

$$K(x_1j_1,x_2j_2,x_3j_3) = N|x_2-x_3|^{-2b_1}|x_1-x_3|^{-2b_2}|x_1-x_2|^{-2b_3}, \tag{8}$$

where the parameters b_i are linear combinations of the j_i

$$2b_i = J - 2j_i + n\,, \quad J = j_1+j_2+j_3\,. \tag{9}$$

The verification of its invariance is based on the relations

$$d^nx' = (\omega'/\omega)^{-n}\,d^nx \tag{10}$$

and

$$(x_i'-x_m')^2 = (\omega_i'/\omega_i)^{-1}(\omega_m'/\omega_m)^{-1}(x_i-x_m)^2\,. \tag{11}$$

The last equation is obviously a consequence of the relation $k_i'\cdot k_m' = k_i\cdot k_m$. The module of constant N will be fixed by the orthogonality relations of the following form:

$$\int_{R_n}\int_{R_n} K(x_1j_1,x_2j_2,x_3j_3)\,\overline{K(x_1j_1,x_2j_2,x_3'j_3')}\,d^nx_1\,d^nx_2 = \frac{2(2\pi)^{n/2+1}}{\Gamma(n/2)}\left|\frac{\Gamma(n/2+j_3)}{\Gamma(-j_3)}\right|^2\delta(\rho_3-\rho_3')\delta(x_3-x_3')\,. \tag{12}$$

We shall choose a phase of N as

$$N = 2^{J+3n/2}\left[\frac{\Gamma(b_1)(b_2)\Gamma(b_3)\Gamma(-n/2+b_1+b_2+b_3)}{\Gamma(n/2-b_1)\Gamma(n/2-b_2)\Gamma(n/2-b_3)\Gamma(n-b_1-b_2-b_3)}\right]^{1/2}. \tag{13}$$

Trilinear invariant functionals for three arbitrary representations can be obtained by analytic continuation in the three j's from the trilinear invariant functional for three representations of the principal series.

We are interested in examining the Wigner coefficients of SO$(n+1,1)$ in a basis in which the operators P_i, $i=1,2,\ldots,n$ corresponding to the generators of the subgroup $T(n)$ are diagonal. These vectors are denoted in the usual fashion by the kets $|j;p\rangle$, $p=(p_1,p_2,\ldots,p_n)$,

$$P_i|j;p\rangle = p_i|j;p\rangle\,, \quad \langle j;q|j;p\rangle = \delta(q-p)\,. \tag{14}$$

In what follows, we restrict our discussion to the case when all three representations belong to the principal series. Wigner coefficients for the other cases can be obtained by the analytic continuation of those from the principal series (cf. Ref. 4).

III. AN INTEGRAL REPRESENTATION FOR THE WIGNER COEFFICIENTS

By carrying out a transformation

$$f(x) = (2\pi)^{-n/2}\int_{R_n} d^np\,|p|^{-n/2-j}e^{ipx}|j;p\rangle\,, \tag{15}$$

we pass to the parabolic basis. Indeed, it follows from (5) that

$$U^j(a)|j;p\rangle = e^{ipa}|j;p\rangle\,, \quad a\in T(n)\,. \tag{16}$$

It is worthwhile to note that the matrix elements of the construction representation may be calculated from the formula

$$\langle j;q|U^j(g)|j;p\rangle = (2\pi)^{-n}\left(\frac{|q|}{|p|}\right)^{-n/2-j}\int_{R_n} d^nx\left(\frac{\omega'}{\omega}\right)^j \times\exp(iqx'-ipx)\,. \tag{17}$$

This has been partially investigated by Vilenkin,[2] who has given the matrix elements in this basis for SO(2,1).

For the explicit calculation of the Wigner coefficients in the parabolic basis we rewrite (7) in the following form:

$$\int_{R_n}\int_{R_n}\int_{R_n}\begin{pmatrix} j_1 & j_2 & j_3\\ p_1 & p_2 & p_3\end{pmatrix}|j_1;p_1\rangle|j_2;p_2\rangle \times|j_3;p_3\rangle\,d^np_1\,d^np_2\,d^np_3 = \text{inv.} \tag{18}$$

This then gives the integral representations of the Wigner coefficients

$$\begin{pmatrix} j_1 & j_2 & j_3\\ p_1 & p_2 & p_3\end{pmatrix} = (2\pi)^{3n/2}\kappa\int_{R_n} d^nx_1\int_{R_n} d^nx_2 \times\int_{R_n} d^nx_3\,K(x_1j_1,x_2j_2,x_3j_3) \times\exp(ip_1x_1+ip_2x_2+ip_3x_3)\,, \tag{19}$$

where

$$\kappa = |p_1|^{-j_1-n/2}|p_2|^{-j_2-n/2}|p_3|^{-j_3-n/2}\,.$$

Using the integral formula (see, e.g., Chap. II of Ref. 5)

$$|x|^{-2c} = \frac{\pi^{-n/2}2^{-2c}\Gamma(-c+n/2)}{\Gamma(c)}\int_{R_n} d^nq\,|q|^{2c-n-iqx}\,, \tag{20}$$

we can rewrite (19) as

$$\begin{pmatrix} j_1 & j_2 & j_3\\ p_1 & p_2 & p_3\end{pmatrix} = \delta(p_1+p_2+p_3)2^{-J-3n/2}\kappa N \times\frac{\Gamma(n/2-b_1)\Gamma(n/2-b_2)\Gamma(n/2-b_3)}{\Gamma(b_1)\Gamma(b_2)\Phi(b_3)} \times\int_{R_n} d^nq\,|q+p_2|^{2b_1-n} \times|q-p_1|^{2b_2-n}|q|^{2b_3-n}\,. \tag{21}$$

IV. WIGNER COEFFICIENTS FOR SO(2,1)

In this section we carry out the integration appearing in the formula (21) when $n=1$. Suppose, for example, that $p_1,p_2>0$. Performing the integration (see, e.g., Eqs. 3.197.2 and 3.197.8 of Ref. 6), we obtain

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = \delta(p_1+p_2+p_3)2^{-J-3/2}\kappa N \frac{\Gamma(\frac{1}{2}-b_1)\Gamma(\frac{1}{2}-b_2)\Gamma(\frac{1}{2}-b_3)}{\Gamma(b_1)\Gamma(b_2)\Gamma(b_3)} \{ p_2^{2b_1-1} p_1^{2b_2+2b_3-1}$$

$$\times B(2b_2,2b_3)\,{}_2F_1(1-2b_1,2b_3,2b_2+2b_3;-p_2/p_1) + p^{2b_1+2b_2+2b_3-2}B(2-2b_1-2b_2-2b_3,2b_2)$$

$$\times {}_2F_1(1-2b_1,2-2b_1-2b_2-2b_3;2-2b_1-2b_3;-p_2/p_1) + (b_1 \leftrightarrow b_2, p_1 \leftrightarrow p_2)\}, \tag{22}$$

where B is the Euler beta function. Using the formula (9.132) of Ref. 6 one can express the Wigner coefficients of SO(2,1) in terms of two hypergeometric functions ${}_2F_1$. The result is

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = C\delta(p_1+p_2+p_3)\{A(p_2/p_1)^{2b_3}\,{}_2F_1(2b_3,1-2b_2;2b_1+2b_3;-p_2/p_1) + (j_2 \leftrightarrow -j_2-1)\}, \tag{23}$$

where

$$C = 2^{-J-3/2}\pi^{1/2} p_1^{2b_2+2b_3-1} p_2^{2b_1-1}\kappa N\Gamma(\tfrac{1}{2}-b_3)/\Gamma(b_3),$$

$$A = [\Gamma(1/2-b_2)\Gamma(b_3)\Gamma(1/2-b_1-b_3)]/[\Gamma(1/2-b_3)\Gamma(b_2)\Gamma(b_1+b_3)].$$

Expression (23) is valid for $p_1,p_2>0$. Wigner coefficients of SO(2,1) for different values of sgn p_i can be obtained from (23) by employing the symmetry relations (30) (see Sec. VI).

V. WIGNER COEFFICIENTS OF SO($n+1$,1), $n>1$

Using the integral formula (see, e.g., Eq. 8.312 of Ref. 6)

$$|q|^{2c} = \frac{1}{\Gamma(-c)}\int_0^\infty ds\, s^{-c-1}e^{-q^2 s}, \tag{24}$$

we can rewrite (21) as (see, e.g., Eq. 3.323 of Ref. 6)

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = \delta(p_1+p_2+\pi_3)\pi^{n/2}2^{-J-3n/2}\kappa N \frac{1}{\Gamma(b_1)\Gamma(b_2)\Gamma(b_3)}\int_0^\infty ds \int_0^\infty dt \int_0^\infty du$$

$$\times s^{-1-b_1+n/2}\, t^{-1-b_2+n/2} u^{-1-b_3+n/2}(s+t+u)^{-n/2}\exp\left(-\frac{tup_1^2+sup_2^2+stp_3^2}{s+t+u}\right). \tag{25}$$

It is convenient now to introduce new variables v, w such that

$$s = uv, \quad t = uw.$$

Integration over u yields

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = \delta(p_1+p_2+p_3)\pi^{n/2}2^{-J-3n/2}\kappa N \frac{\Gamma(n-b_1-b_2-b_3)}{\Gamma(b_1)\Gamma(b_2)\Gamma(b_3)}\int_0^\infty dv\int_0^\infty dw$$

$$\times v^{-1+b_2+b_3-n/2}\, w^{-1+b_1+b_3-n/2}(1+v+w)^{-b_1-b_2-b_3+n/2}(p_1^2/v+p_2^2/w+p_3^2)^{b_1+b_2+b_3-n}. \tag{26}$$

The most symmetrical integral representations for the Wigner coefficients can be obtained by noting that (26) can be rewritten as

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = \delta(p_1+p_2+p_3)\kappa D\int_0^\infty dr\int_0^\infty ds\int_0^\infty dt\int_0^\infty du\, s^{-1+b_2+b_3-n/2}$$

$$\times t^{-1+b_1+b_3-n/2}\, u^{-1+b_1+b_2-n/2}\exp\{-(s+t+u)r-p_1^2/s-p_2^2/t-p_3^2/u\},$$

where

$$D = \pi^{n/2}2^{-J-3n/2}N\,[\Gamma(b_1)\Gamma(b_2)\Gamma(b_3)\Gamma(b_1+b_2+b_3-n/2)]^{-1}.$$

Here, we have used (24) again. Using further the integral representation of the Bessel K function (see, e.g., Eq. 8.432.7 of Ref. 6), we obtain the result

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = D\delta(p_1+p_2+p_3)\int_0^\infty dr\, r^{-1+n/2}K_{j_1+n/2}(2|p_1|r)K_{j_2+n/2}(2|p_2|r)K_{j_3+n/2}(2|p_3|r). \tag{27}$$

Finally, the r integration can also be performed with the help of the formulas (6.578.2) and (8.485) of Ref. 6. One obtains

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = \delta(p_1+p_2+p_3)D\,|p_3|^{-n/2}\left\{\left[\frac{|p_1|}{|p_3|}\right]^{j_1+n/2}\left[\frac{|p_2|}{|p_3|}\right]^{j_2+n/2}\right.$$

$$\times RF_4\left(\frac{j_1+j_2+j_3+n}{2},\frac{j_1+j_2+j_3+2n}{2},1+j_1+\frac{n}{2},1+j_2+\frac{n}{2};\frac{p_1^2}{p_3^2},\frac{p_2^2}{p_3^2}\right)$$

$$\left.+(j_1\leftrightarrow -n-j_1)+(j_2\leftrightarrow -n-j_2)+(j_1\leftrightarrow -n-j_1, j_2\leftrightarrow -n-j_2)\right\}, \tag{28}$$

where

$$R=\Gamma(-j_1-n/2)\Gamma(-j_2-n/2)\Gamma((j_1+j_2+j_3+n)/2)\Gamma((j_1+j_2+j_3+2n)/2).$$

Here F_4 is the Appell function.

When $n=2$, the Appell functions in (28) can be expressed as a product of ordinary hypergeometric functions (see Ref. 7, p. 100). Consequently, the Wigner coefficients of SO(3,1) can be written as

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = \delta(p_1+p_2+p_3)D\,|p_3|^{-1}\left\{\left[\frac{|p_1|}{|p_3|}\right]^{1+j_1}\left[\frac{|p_2|}{|p_3|}\right]^{1+j_2}\frac{\Gamma(-1-j_1)}{\Gamma(2+j_1)}\right.$$
$$\times\Gamma\left(\frac{j_1+j_2-j_3+2}{2}\right)\Gamma\left(\frac{j_1+j_2+j_3+4}{2}\right)\Gamma\left(\frac{j_1-j_2+j_3}{2}\right)\Gamma\left(\frac{j_1-j_2-j_3}{2}\right)$$
$$\times{}_2F_1\left(\frac{j_1+j_2-j_3+2}{2},\frac{j_1+j_2+j_3+4}{2};2+j_1;z\right){}_2F_1\left(\frac{j_1+j_2-j_3+2}{2},\right.$$
$$\left.\left.\times\frac{j_1+j_2+j_3+4}{2};2+j_1;Z\right)+(j_1\leftrightarrow -2-j_1)\right\}, \tag{29}$$

where

$$zZ=p_1^2/p_3^2\,,\quad (1-z)(1-Z)=p_2^2/p_3^2\,.$$

VI. SYMMETRY RELATIONS FOR THE WIGNER COEFFICIENTS

From the integral representations (27) one can derive the following symmetry relations for the Wigner coefficients:

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = \begin{pmatrix} j_i & j_k & j_m \\ p_i & p_k & p_m \end{pmatrix} = \begin{pmatrix} j_1 & j_2 & j_3 \\ -p_1 & -p_2 & -p_3 \end{pmatrix} \tag{30}$$
$$= \begin{pmatrix} -n-j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix} = \begin{pmatrix} j_1 & -n-j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix}$$
$$= \begin{pmatrix} j_1 & j_2 & -n-j_3 \\ p_1 & p_2 & p_3 \end{pmatrix}. \tag{31}$$

It is also worth noting that the Wigner coefficients of SO($n+1,1$), so defined, are real

$$\overline{\begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix}} = \begin{pmatrix} j_1 & j_2 & j_3 \\ p_1 & p_2 & p_3 \end{pmatrix}. \tag{32}$$

VII. CONCLUSION

We have considered the problem of the Wigner coefficients for the SO($n+1,1$) groups in the parabolic basis and have obtained a general expression for these coefficients for class 1 representations. The merit of our derivation is that it generalizes to other representations and other semisimple Lie groups. We intend in the near future to discuss the Wigner coefficients for the SL(n,C) groups.

[1] G. A. Kerimov and Yi. A. Verdiev, Rep. Math. Phys. **20**, 117 (1981).

[2] N. J. Vilenkin, *Special Functions and the Theory of Group Representations* (Amer. Math. Soc., Providence, RI, 1968).

[3] V. K. Dobrev, G. Mack, V. B. Petkova, S. G.Petrova, and I. T. Todorov, "Harmonic analysis on the n-dimensional Lorentz group and its application to conformal quantum field theory," in *Lecture Notes in Physics, Vol. 63* (Springer, New York, 1977).

[4] G. A. Kerimov and Yi. A. Verdiev, Rep. Math. Phys. **13**, 315 (1978).

[5] I. M. Gelfand and G. E. Shilov, *Generalized Functions* (Academic, New York, 1964), Vol. 1.

[6] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products* (Academic, New York, 1965).

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

Generalized FS$\times$W versus Dyson's classification of irreducible representations

Claudio Garola and Luigi Solombrino
Dipartimento di Fisica dell'Università di Lecce, 73100 Lecce, Italy and Istituto Nazionale di Fisica Nucleare, Sez. di Bari, 70126, Bari, Italy

(Received 4 December 1984; accepted for publication 1 March 1985)

We study the characterization of the 13 cases obtained in the classification of the irreducible linear–antilinear representations of semigroups in a finite-dimensional vector space X over an algebraically closed field K with a conjugation j (generalized Frobenius–Schur–Wigner, or FS$\times$W, classification). It has already been shown that each case can be characterized by various equivalent properties, some of which can be endowed with a physical interpretation. We show here that, whenever K is the complex field C, each case can be characterized by the structure (in the sense specified by the Weyl theorem on the structure of the matrix algebras and their commutators) of a pair of suitable operator algebras over the real field R. This characterization coincides with the one given by Dyson for each case of his classification of symmetry groups. Thus, the latter classification is recovered under more general assumptions and in a generalized framework, and its one-to-one correspondence with the generalized FS$\times$W classification is shown. In the process, we obtain a classification of the algebras over R generated by complex semigroup representations of the aforesaid kind, and inquire into the connections between some properties of the representations in X over C and the structure (in the Weyl sense) of the representations in the space X_R over R obtained by decomplexification of the space X.

I. INTRODUCTION

The classical threefold classification of the unitary, finite-dimensional complex group representations by Frobenius and Schur (FS classification) is well known and still widely used.[1–4]

Also well known is the threefold Wigner classification[5] (W classification) of the irreducible finite-dimensional complex group corepresentations (i.e., representations by unitary and antiunitary operators).

The FS classification and the W classification cannot be combined in their original form; indeed the W classification becomes trivial whenever the antiunitary part of the representation is void.

In the early 1960's, Dyson[6] proposed a classification (D classification) in 13 cases (ten cases only occur if the antiunitary part of the representation is assumed to be nonvoid) which applies to the same class of representations as the W classification. In the Dyson approach, each case is characterized by the (matrix) form of a pair of group algebras over the real field R; all the matrices have elements in R and represent linear operators on the vector space X_R over R obtained by decomplexification of the original vector space X over the complex field C to which the quantum mechanical vectors representing (pure) states belong. The D classification is connected to the above classifications through two equivalence theorems; in the sense established by these theorems, we can say that each case in the D classification is obtained by classifying the representation according to the W classification and its linear part according to a suitable refinement of the FS classification which, incidentally, Dyson uses in the form reported by Wigner (that refers to irreducible representations only).

Dyson's choice of X_R in place of X is motivated by some relevant arguments. First, the Weyl theorem on the structure of matrix algebras and their commutators,[7] which is the fundamental tool in Dyson's treatment, refers to linear representations only; shifting from X to X_R allows any antiunitary operator to be represented by a linear (orthogonal) mapping of X_R (then the classical Frobenius theorem can be applied, which restricts any division algebra over R to be isomorphic either to R or to C or to the real quaternion field Q). Second, Dyson explicitly declares his belief that "the appropriate ground field for much of quantum mechanics is real rather than complex" and his purpose in making "the use of the real ground field in quantum mechanics official and undisguised."[8]

Concerning the latter argument, one can object that the set of all the linear operators on X_R contains R-linear operators, which do not correspond to linear or antilinear operators on X and do not have any apparent physical interpretation; hence, the choice of R as ground field makes the distinction between the physical and the "unphysical" part of the operator algebras under examination rather intriguing in some cases. Moreover, Dyson's characterization of the cases does not admit any transparent physical interpretation.

On the other hand, a direct combination of the FS classification with the W classification becomes possible if the former is generalized so as to apply to (irreducible or not) corepresentations; of course, the space X_R has no role in this procedure. This generalization has been made by ourselves, together with Ascoli and Teppati, in a previous paper[9]; in our classification the vector space need not be finite dimensional, the ground field is any field K with an involutory automorphism j (conjugation) and the representation U is any (not necessarily irreducible, nor completely reducible) operator representation of some set $\mathscr{S}$, which is not assumed to be unitary–antiunitary (since no scalar product is assumed

 0022-2488/85/081889-13$02.50

in X over K) nor linear–antilinear. In this classification, every case is characterized by sets of equivalent properties of $\mathscr{U} = U(\mathscr{S})$, or of the group $\mathscr{U}^c$ (commutant group) of all the nonzero linear and antilinear (i.e., j-semilinear) mappings of X which commute with $\mathscr{U}$. This characterization seems particularly interesting, since $\mathscr{U}^c$ admits a physical interpretation; indeed, whenever $\mathscr{U}$ is a given group of symmetries of some physical system, $\mathscr{U}^c$ may represent symmetries which are considered as "internal" with respect to the group. In a second paper, our classification has been refined further[10] so as to obtain a sixfold classification (generalized FS classification) instead of the original threefold one, which applies, in particular, to the same class of representations as the W classification. As a further step, we have also generalized the W classification[11] to irreducible linear–antilinear semigroup representations in a finite-dimensional vector space X (on which no scalar product is assumed) over an algebraically closed field K with a conjugation; then, a new classification (the generalized FS $\times$ W classification) has been obtained for the last class of representations by combining the generalized FS classification with the generalized W classification.

In the generalized FS $\times$ W classification the 13 cases which may occur are characterized by equivalent properties of the group $\mathscr{U}^c$ and by the explicit form of its linear part. Besides, in each case further properties can be attributed to $\mathscr{U}$ by making use of the equivalent properties that characterize each case in the FS generalized classification.

It seems at first sight rather casual that the number of the possible alternatives is the same as in the D classification; indeed the generalized FS $\times$ W classification does not make any reference to the explicit form of $\mathscr{U}$ in X nor in X_R (except for distinguishing the cases in which $\mathscr{U}$ contains antilinear mappings or not), nor it is established by making use of the Weyl and Frobenius theorems. Thus, one may wonder about the relations between the two classifications in that subclass of representations to which the D classification also applies. Rather surprisingly, we shall presently show that the correspondence is one-to-one.

More precisely, in the present paper (Theorem 1) we first study the structure in X_R (in the sense specified by the Weyl theorem) of the (linear–antilinear graded) algebra $\mathfrak{A}$ over R generated by any semigroup $\mathscr{U}$ of linear and antilinear mappings of a finite-dimensional vector space X over C. By making use of the same tools as Dyson (i.e., the Weyl and the Frobenius theorems), we obtain a new classification of $\mathfrak{A}$ in eight cases.

Then, we inquire further (Theorem 2) into the conditions that characterize each case in the generalized FS classification (see the rows in Table II) in the present framework ($\mathscr{U}$ irreducible, $K = C$) and study their implications on the structure in X_R of $\mathfrak{A}$ and of its C-linear part $\mathfrak{A}^l$.

Finally, we study (Theorem 3) the relation between the generalized FS $\times$ W classification (particularized to C) and the classification obtained in Theorem 1 (the former refines the latter); moreover, we show that each case of the generalized FS $\times$ W classification can be characterized by the structure in X_R of the pair $(\mathfrak{D}, \mathfrak{A}^l)$ of auxiliary algebras (where $\mathfrak{D}$ is the algebra over R generated by $\mathscr{U} \cup \{iE\}$, i being the imaginary unit and E the identity mapping in X). This characterization coincides, whenever the matrix representations with respect to a basis of the operators in X_R are considered, with the characterization of each case in the D classification (which is, however, obtained under more restrictive assumptions: unitary–antiunitary representations over a vector space X endowed with a scalar product and group instead of semigroup representations). Thus, the generalized FS $\times$ W classification particularized to C generalizes the Dyson classification, and the statement of our Theorem 3 shows the one-to-one correspondence between the two classifications. As a relevant consequence, each case in the Dyson classification is characterized by suitable properties of $\mathscr{U}^c$, which can be endowed of a physical interpretation (internal symmetries) as we have outlined above.

All the aforesaid results are concentrated in Sec. IV of the present paper. We devote Sec. II to basic definitions and to a synthesis of previous results and mathematical tools, and Sec. III to a restatement in terms of operators rather than matrix algebras of the fundamental (though, surprisingly, unfrequently quoted) Weyl theorem. In Sec. V we give some simple examples which illustrate the correspondence between the generalized FS $\times$ W classification and the D classification, with special attention to the "factorizable groups,"[12] whose physical interest has been already commented upon by Dyson.

II. THE GENERALIZED FS $\times$ W CLASSIFICATION

As we have mentioned in the Introduction, in this section we assemble some definitions and results that we have already introduced in various papers and that are needed in order to develop our main argument in Sec. IV.

Definition 1: We call any division ring K endowed with a nonidentical involutory automorphism $j:\alpha \to \bar{\alpha}$ a "division ring with a conjugation." We call the division ring Λ of K which consists of the self-conjugated elements of K the "j-invariant subring of K."

Definition 2: Let X be a vector space over a division ring K with a conjugation j and let Λ be the j-invariant subring of K. Then we call any additive mapping of X which is semilinear with respect to j an "antilinear mapping." We denote the vector space obtained from X by restriction to Λ of the scalar field by X_Λ.[13,14] Let $\mathscr{E}$ be a basis in X. We call the antilinear involutory mapping $J_{\mathscr{E}}$ that leaves the elements of $\mathscr{E}$ invariant, "conjugation in X associated with the basis $\mathscr{E}$." For any mapping M of X and any basis $\mathscr{E}$, we define the "conjugate mapping" $\overline{M}_{\mathscr{E}} = J_{\mathscr{E}} M J_{\mathscr{E}}$ of M with respect to the basis $\mathscr{E}$.

Definition 3: Let X be a vector space over a division ring K with a conjugation j and let Λ be the j-invariant subring of K. Let $\mathscr{U}$ be any set of mappings of X. We denote the subsets of all the linear, antilinear, Λ-linear mappings of $\mathscr{U}$ by $\mathscr{U}^l$, $\mathscr{U}^a$, $\mathscr{U}^\Lambda$, respectively. Let H be any commutative subfield of K and let $\mathfrak{A}$ be any algebra over H of mappings of X. We say that $\mathfrak{A}$ is a "linear–antilinear algebra" if $\mathfrak{A} = \mathfrak{A}^l \oplus \mathfrak{A}^a$ (the sum is necessarily direct); a linear–antilinear algebra will be said to be "graded" whenever $\mathfrak{A}^l$ and $\mathfrak{A}^a$ are vector spaces over H.

Definition 4: Let X be a vector space over a division ring K with a conjugation j and let Λ be the j-invariant subring of K. Let $\mathscr{U}$ be any set of mappings of X. We denote the set of mappings of X that commute with all the mappings of $\mathscr{U}$ by $\mathscr{U}'$. Hence $\mathscr{U}'^{l}$ (equivalently, $\mathscr{U}'^{K}$ whenever the field to which linearity refers needs to be stressed), $\mathscr{U}'^{a}$, $\mathscr{U}'^{\Lambda}$, respectively, denote the "linear commutant," the "antilinear commutant" and the "Λ-linear commutant" of $\mathscr{U}$, i.e., the subsets of the linear, antilinear, and Λ-linear mappings of $\mathscr{U}'$, respectively.

We call the group $\mathscr{U}^{c}$ of the invertible linear and antilinear mappings of $\mathscr{U}'$, the "linear–antilinear centralizer" (or "commutant group") of $\mathscr{U}$. Hence $\mathscr{U}^{cl}$ and $\mathscr{U}^{ca}$, respectively, denote the linear and antilinear part of $\mathscr{U}^{c}$. Furthermore, we say that $\mathscr{U}$ is "potentially real" whenever an involutory mapping [hence, a conjugation with respect to some basis $\mathscr{E}$ (see Ref. 15)] $J_{\mathscr{E}} \in \mathscr{U}^{ca}$ exists; we say that $\mathscr{U}$ is "pseudoreal" whenever $\mathscr{U}^{ca}$ is nonvoid but no $A \in \mathscr{U}^{ca}$ is involutory; we say that $\mathscr{U}$ is "complex" whenever $\mathscr{U}^{ca}$ is void.

Remark 1: We have proved in a previous paper[16] that the cases (i) potentially real, (ii) pseudoreal, and (iii) complex can be characterized by a number of equivalent properties; in particular: (i) occurs iff a basis $\mathscr{E}$ of X exists such that $\overline{M}_{\mathscr{E}} = M$ for any $M \in \mathscr{U}$; (ii) occurs iff for every basis $\mathscr{E}$ the "conjugate set" $\overline{\mathscr{U}}_{\mathscr{E}} = J_{\mathscr{E}} \mathscr{U} J_{\mathscr{E}}$ is equivalent to $\mathscr{U}$ but does not coincide with it; and (iii) occurs iff $\overline{\mathscr{U}}_{\mathscr{E}}$ is inequivalent to $\mathscr{U}$ in every basis.

Definition 5: Let X be a vector space over a division ring K with a conjugation j and let Λ be the j-invariant subring of K. Let $\mathscr{S}$ be any semigroup; we call any semigroup homomorphism U from $\mathscr{S}$ into the multiplicative semigroup of all the linear and antilinear mappings of X a "linear–antilinear representation" of $\mathscr{S}$ in X. Let $\mathscr{A}$ be an algebra over a subfield H of the center of Λ; we also call any H-algebra homomorphism U from $\mathscr{A}$ into the linear–antilinear algebra over H of all the Λ-linear mappings of X a "linear–antilinear representation" of $\mathscr{A}$ in X. In both cases, we denote by U^{l} the "linear part" of U, i.e., the restriction of U to $U^{-1}(U(\mathscr{S})^{l})$ or to $U^{-1}(U(\mathscr{A})^{l})$, respectively.

A list of the not-yet-defined symbols that will be needed can be found in Table I. With these definitions and symbols, the following proposition holds, which summarizes our results in a previous paper[17] (generalized FS×W classification). As we have already observed in the Introduction, it is worth noting that each of the mutually exclusive cases corresponding to the squares in Table II is characterized by the explicit form of the linear commutant $\mathscr{U}'^{l}$ of $\mathscr{U}$ together with suitable properties of the linear–antilinear centralizer.

TABLE I. List of symbols.

θ:	the mapping from $\mathscr{U}^{c}$ into the set of the mappings of $\mathscr{U}^{cl}$ into itself such that, for any $A \in \mathscr{U}^{c}$, $\theta(A): L \in \mathscr{U}^{cl} \to A L A^{-1} \in \mathscr{U}^{cl}$.
$\otimes_{\varphi}$:	the semidirect product associated with the morphism φ (the symbol φ may be omitted whenever the morphism need not to be stressed).
$\times$:	direct product.
G_2	the abstract two elements group.
E:	the identity mapping
$\varnothing$:	the empty set.
K_{*}:	for any field K, the multiplicative subgroup $K \setminus \{0\}$.
$\sim$:	denotes equivalence of representations.
$\approx$:	denotes group or algebra isomorphism.

In each case, further characterizing properties of $\mathscr{U}$ or $\mathscr{U}^{c}$ can be deduced from our results in some previous papers.[9,10]

FS×W Theorem: Let X be a finite-dimensional vector space over an algebrically closed field K with a conjugation j. With reference to Definitions 1–5, let $\mathscr{S}$ be a semigroup and let U be an irreducible linear–antilinear representation of $\mathscr{S}$ in X such that the antilinear part $\mathscr{U}^{a}$ of $\mathscr{U} = U(\mathscr{S})$ either is void or contains an invertible mapping A together with its inverse. Then, the linear part U^{l} of U either is irreducible or splits into two equidimensional irreducible subrepresentations U_1^{l}, U_2^{l}, which may be equivalent or not, and the corresponding matrix representation[18] of A is off-diagonal,

$$A = \begin{pmatrix} 0 & A_{12} \\ A_{21} & 0 \end{pmatrix}.$$

Whenever $U_1^{l} \sim U_2^{l}$, for any T_{12} such that $U_1^{l} = T_{12} \times U_2^{l} T_{12}^{-1}$ a unique element $\eta \in K$ exists such that, referring to the matrix representation considered above, $\eta A_{21} = T_{12} A_{12} T_{12}$. Furthermore $\mathscr{U}$ can be classified according to Table II. (In the table, J is any antilinear involutory mapping, Ω a division algebra of rank 4 over Λ, which is isomorphic to the real quaternions whenever $K = C$, E_1 and E_2 the identity mappings in the U^{l}-invariant subspaces whenever U^{l} is reducible.)

III. THE WEYL THEOREM

The Weyl theorem, together with the Frobenius theorem about the division algebras over the real field, is the basic mathematical tool for the classification introduced by Dyson[6] and for the rest of our paper as well. It can be derived in the framework of the theory of semisimple modules.[19] We restate it here by considering operators rather than matrix algebras, taking care to introduce minimal changes with respect to the original statement by Weyl[7] and also to preserve, as far as possible, Weyl's notations (which are in some case most suitable for physicist's use), so as to make immediate the comparison of our results with Dyson's. We premise the following definition.

Definition 6: Let X be a vector space over a field K. We denote by $\mathscr{L}_K(X)$ the algebra over K of all the (K-) linear mappings of X.

Let $X = \oplus_{j=1}^{t} X_j$ be any decomposition of X into a direct sum of subspaces. For any $A \in \mathscr{L}_K(X)$ we denote by A_{jk} the linear mapping in the jth row, kth column of the matrix representation of A with respect to this decomposition. Furthermore, we write $X = tY$ whenever each of the X_j is isomorphic to a vector space Y over K. Let $\mathfrak{A} \subset \mathscr{L}_K(X)$ be any algebra over K of linear mappings of X.

Then, whenever $\mathfrak{A}$ decomposes into s equivalent components, we write $\mathfrak{A} = s\mathscr{A}$, where $\mathscr{A}$ denotes, up to isomorphisms, any one of the components (hence $\mathscr{A}$ is an algebra over K isomorphic to $\mathfrak{A}$).

Furthermore, let a decomposition $X = \oplus_{j=1}^{t} X_j = tY$ of the space X exist such that, for every $j, k \in \{1,2,\ldots,t\}$, the vector space $\mathfrak{A}_{jk} = \{A_{jk} : A \in \mathfrak{A}\}$ is endowed of a structure of algebra (through the identification $X_j = Y = X_k$) isomorphic to a given algebra $\mathscr{A} \subset \mathscr{L}_K(Y)$ of linear mappings of Y, and let $\mathfrak{A}$ be the direct sum of the $\mathfrak{A}_{jk}$; then we write $\mathfrak{A} = \mathscr{A}_t$.

TABLE II. Generalized FS$\times$W classification of the irreducible linear–antilinear representations by means of the commutant group.

	$\mathscr{U}^{a}=\{0\}$	$\mathscr{U}^{a}\neq\{0\}$		
	$\mathscr{U}^{1}$irreducible		$\mathscr{U}^{1}$ reducible	
			$U_1^1 \not\sim U_2^1$	$U_1^1 \sim U_2^1$
	$\mathscr{U}'^{1}=KE$	$\mathscr{U}'^{1}=\Lambda E$	$\mathscr{U}'^{1}=\left\{\begin{pmatrix}\alpha E_1 & 0\\ 0 & \bar{\alpha}E_2\end{pmatrix}:\alpha\in K\right\}=K$	$\mathscr{U}'^{1}=\left\{\begin{pmatrix}\alpha E_1 & \beta T_{12}\\ \bar{\beta}\eta T_{12}^{-1} & \bar{\alpha}E_2\end{pmatrix}:\alpha,\beta\in K\right\}\approx\Omega$
$\mathscr{U}$ is potentially real		$\mathscr{U}^{c}=\Lambda_{*}\times\{E,J\}$ $\approx\Lambda_{*}\times G_2$ ($\mathscr{U}\cup\mathscr{U}^{c}$ is potentially real)	$\mathscr{U}^{c}=\mathscr{U}^{cl}\times\{E,J\}$ $\approx K_{*}\times G_2$ ($\mathscr{U}\cup\mathscr{U}^{c}$ is potentially real)	$\mathscr{U}^{c}=\mathscr{U}^{cl}\times\{E,J\}$ $\approx\Omega_{*}\times G_2$ ($\mathscr{U}\cup\mathscr{U}^{c}$ is potentially real)
	$\mathscr{U}^{c}=K_{*}\otimes_{\theta}\{E,J\}$ $\approx K_{*}\otimes G_2\not\simeq K_{*}\times G_2$ ($\mathscr{U}\cup\mathscr{U}^{c}$ is complex)		$\mathscr{U}^{c}=\mathscr{U}^{cl}\otimes_{\theta}\{E,J\}$ $\approx K_{*}\otimes G_2\not\simeq K_{*}\times G_2$ ($\mathscr{U}\cup\mathscr{U}^{c}$ is complex)	$\mathscr{U}^{c}=\mathscr{U}^{cl}\otimes_{\theta}\{E,J\}$ $\approx\Omega_{*}\otimes G_2\not\simeq\Omega_{*}\times G_2$ ($\mathscr{U}\cup\mathscr{U}^{c}$ is pseudoreal)
$\mathscr{U}$ is pseudoreal	For any $A\in\mathscr{U}^{ca}$, $\mathscr{U}^{c}=K_{*}E\cup K_{*}A$ $\not\simeq K_{*}\otimes G_2$ [$\theta(A)$ is not identical, $\mathscr{U}\cup\mathscr{U}^{c}$ is complex]	For any $A\in\mathscr{U}^{ca}$, $\mathscr{U}^{c}=\Lambda_{*}E\cup\Lambda_{*}A$ $\not\simeq\Lambda_{*}\otimes G_2$ [$\theta(A)$ is identical, $\mathscr{U}\cup\mathscr{U}^{c}$ is pseudoreal]	For any $A\in\mathscr{U}^{ca}$, $\mathscr{U}^{c}=\mathscr{U}^{cl}\cup\mathscr{U}^{cl}A$ $\not\simeq K_{*}\otimes G_2$ [$\theta(A)$ is not identical, $\mathscr{U}\cup\mathscr{U}^{c}$ is complex]	
$\mathscr{U}$ is complex	$\mathscr{U}^{c}=K_{*}E$	$\mathscr{U}^{c}=\Lambda_{*}E$	$\mathscr{U}^{c}=\mathscr{U}^{cl}\approx K_{*}$	$\mathscr{U}^{c}=\mathscr{U}^{cl}\approx\Omega_{*}$

Weyl's theorem: Let X be a vector space of finite dimension n over a field K. With reference to Definition 6, let $\mathfrak{A}$ be any completely reducible subalgebra (over K) of $\mathscr{L}_K(X)$. Then, a direct sum decomposition $X=\oplus_{j=1}^{p}s_jX_j=\oplus_{j=1}^{p}s_j(t_jY_j)$ exists such that the corresponding matrix representation of $\mathfrak{A}$ takes the canonical form

$$\mathfrak{A}=\bigoplus_{j=1}^{p}s_j\mathscr{A}^{j}_{t_j},$$

where $\mathscr{A}^{j}$ is an irreducible division algebra over K of linear mappings of Y_j (hence the order h_j of $\mathscr{A}^{j}$ coincides with the dimension of Y_j, so that $\Sigma_j h_js_jt_j=n$). Furthermore, a relabeling of the subspaces exists such that the corresponding matrix representation of the (K-) linear commutant $\mathfrak{A}'^{K}$ of $\mathfrak{A}$ takes the form

$$\mathfrak{A}'^{K}=\bigoplus_{j=1}^{p}t_j\,{}^{0}\mathscr{A}^{j}_{s_j},$$

where ${}^{0}\mathscr{A}^{j}=(\mathscr{A}^{j})'^{K}$ is an irreducible division algebra over K (hence it is isomorphic to the opposite algebra[20] of $\mathscr{A}^{j}$ and its order h_j^0 coincides with h_j).

Remark 2: Following Weyl, we notice that $(\mathfrak{A}'^{K})'^{K}=\mathfrak{A}$. Most important for our purposes, we observe that it follows from the proof of the Weyl theorem[21] that in the decomposition of X in which $\mathfrak{A}$ takes the canonical form, the (K-) linear commutant takes the form

$$\mathfrak{A}'^{K}=\bigoplus_{j=1}^{p}(t_j\,{}^{0}\mathscr{A}^{j})_{s_j}.$$

The classical Frobenius theorem is well known. For the sake of completeness, we report it here.

Frobenius' theorem: The real field R and the complex field C are the only finite-dimensional associative-commutative algebras over R without divisors of zero. The division ring of quaternions Q is the only finite-dimensional associative but not commutative algebra over R without divisors of zero.

Remark 3: Whenever in the Weyl theorem the field K is assumed to coincide with the real field, then the algebras $\mathscr{A}^{j}$ and ${}^{0}\mathscr{A}^{j}$ must be isomorphic either to R or to C or to Q because of the Frobenius theorem; in these three alternatives the dimension of the space Y_j either is 1 or 2 or 4, respectively, because of the statements about h_j and h_j^0 in the Weyl theorem itself. It follows that $\mathscr{A}^{j}$ coincides with ${}^{0}\mathscr{A}^{j}$ whenever they are isomorphic to R or to C, while $\mathscr{A}^{j}$ is isomorphic to ${}^{0}\mathscr{A}^{j}$ but does not coincide with it whenever it is isomorphic to Q.

IV. GENERALIZED FS$\times$W VERSUS DYSON CLASSIFICATION

As we mentioned in the Introduction, in this section we discuss the relations between the generalized FS$\times$W classification reported in Sec. II and the classifications that can be obtained by making suitable use of the Weyl and Frobenius theorems. In the sequel, except for Lemma 1, the basic field is always the complex field. Then, we make use of all the definitions and symbols introduced in Sec. II, with C in place of K, R in place of Λ, j the usual complex conjugation, while i always denotes the imaginary unit. Furthermore, we make use of the Weyl theorem with R in place of K and X_R in place of X. Then, making reference to remark 3, we directly write in the formulas R, or C, or Q in place of $\mathscr{A}^{j}$; furthermore, whenever $\mathscr{A}^{j}$, hence ${}^{0}\mathscr{A}^{j}$, is isomorphic to Q, we write ${}^{0}\mathscr{A}^{j}=Q^{0}$ (rather than Q) so as to remind the reader that the algebras $\mathscr{A}^{j}$ and ${}^{0}\mathscr{A}^{j}$ have different elements.

Lemma 1: Let X be a finite-dimensional vector space over a field K, $\mathscr{A}$ any algebra over K, and ξ a (linear) representation of $\mathscr{A}$ in X. Let us put $\mathfrak{A}=\xi(\mathscr{A})$ and let S be any additive mapping of X such that the following properties hold:

(i) $S^{-1}\mathfrak{A}S\subset\mathfrak{A}$,
(ii) $S^{2}\in\mathfrak{A}$,
(iii) $\mathscr{V}=\mathfrak{A}\cup\{S\}$ is irreducible.

Then, either ξ is irreducible or it reduces into two equidimensional irreducible subrepresentations, ξ_1 and ξ_2 (hence ξ is completely reducible), such that the algebras (over K) $\mathfrak{A}_1 = \xi_1(\mathscr{A})$ and $\mathfrak{A}_2 = \xi_2(\mathscr{A})$ are isomorphic and $\mathfrak{A}_2 X = S\mathfrak{A}_1 X$.

Proof: Whenever ξ is reducible, let X_1 be an irreducible $\mathfrak{A}$-invariant subspace of X, and let us put $X_2 = SX_1$. Then, X_2 is also an $\mathfrak{A}$-invariant subspace of X, since $\mathfrak{A}(SX_1) = S(S^{-1}\mathfrak{A}SX_1) \subset SX_1$ because of (i). Furthermore, the sum $X_1 + X_2$ is a $\mathscr{V}$-invariant subspace of X, since $S(X_1 + X_2) \subset X_1 + X_2$ because of (ii), so that it coincides with X because of (iii). Thus, $X_1 \neq X_2$, and $X_2 \cap X_2 = \{0\}$ since $X_1 \cap X_2$ is a proper $\mathfrak{A}$-invariant subspace of X_1. It follows $X = X_1 \oplus X_2$, which proves the first statement in the lemma.[22]

In the above decomposition of X, every $L \in \mathfrak{A}$ takes the matrix form

$$L = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix},$$

with $L_1 \in \mathfrak{A}_1 = \xi_1(\mathscr{A}), L_2 \in \mathfrak{A}_2 = \xi_2(\mathscr{A})$, while

$$S = \begin{pmatrix} 0 & S_{12} \\ S_{21} & 0 \end{pmatrix}.$$

Since $S^{-1}\mathfrak{A}S \subset \mathfrak{A}$, we get $S_{21}^{-1} L_2 S_{21} \in \mathfrak{A}_1$ and $S_{12}^{-1} L_1 S_{12} \in \mathfrak{A}_2$, that is, $\mathfrak{A}_1$ and $\mathfrak{A}_2$ are isomorphic to subalgebras of $\mathfrak{A}_2$ and $\mathfrak{A}_1$, respectively, hence $\mathfrak{A}_1$ and $\mathfrak{A}_2$ are isomorphic. Furthermore, $\mathfrak{A}_2 X = X_2 = SX_1 = S\mathfrak{A}_1 X$. ■

Theorem 1: Let X be a vector space of finite dimension n over the complex field C. With reference to Definition 3, let $\mathfrak{A}$ be any irreducible linear–antilinear (graded) algebra with unit element (whose order we denote by h) over the real field R of mappings of X, and $\mathfrak{A}'^R$, coherently with Definition 4, be the linear–antilinear graded algebra (whose order we denote by h') over R of all the R-linear mappings of X which commute with $\mathfrak{A}$ (see Ref. 23). Then, with reference to Definition 6, $\mathfrak{A}$ and $\mathfrak{A}'^R$ simultaneously take one of the following (mutually exclusive) "canonical" forms (hence $\mathfrak{A}$ and $\mathfrak{A}'^R$ are completely reducible in X_R):

(1)	$\mathfrak{A} = R_{2n}$	$(h = 4n^2)$;
	$\mathfrak{A}'^R = 2nR$	$(h' = 1)$;
(2)	$\mathfrak{A} = C_n$	$(h = 2n^2)$;
	$\mathfrak{A}'^R = nC$	$(h' = 2)$;
(3)	$\mathfrak{A} = Q_{n/2}$	$(h = n^2)$;
	$\mathfrak{A}'^R = (n/2)Q^0$	$(h' = 4)$;
(4)	$\mathfrak{A} = 2R_n$	$(h = n^2)$;
	$\mathfrak{A}'^R = (nR)_2$	$(h' = 4)$;
(5)	$\mathfrak{A} = 2C_{n/2}$	$(h = n^2/2)$;
	$\mathfrak{A}'^R = [(n/2)C]_2$	$(h' = 8)$;
(6)	$\mathfrak{A} = R_n \oplus R_n$	$(h = 2n^2)$;
	$\mathfrak{A}'^R = nR \oplus nR$	$(h' = 2)$;
(7)	$\mathfrak{A} = C_{n/2} \oplus C_{n/2}$	$(h = n^2)$;
	$\mathfrak{A}'^R = (n/2)C \oplus (n/2)C$	$(h' = 4)$;
(8)	$\mathfrak{A} = Q_{n/4} \oplus Q_{n/4}$	$(h = n^2/2)$;
	$\mathfrak{A}'^R = (n/4)Q^0 \oplus (n/4)Q^0$	$(h' = 8)$.

Proof: In each case the orders h and h' immediately follow from the explicit form of $\mathfrak{A}$ and $\mathfrak{A}'^R$. Furthermore, the forms of $\mathfrak{A}'^R$ follow from the forms of $\mathfrak{A}$ via Weyl's theorem (the parentheses being positioned coherently with the equation in remark 2).

Since $\mathfrak{A}$ is irreducible in X, two cases may occur in X_R:

(a) $\mathfrak{A}$ is irreducible in X_R,

(b) $\mathfrak{A}$ is reducible in X_R.

Whenever a occurs, we get in X_R, by making use again of the Weyl theorem, $\mathfrak{A} = \mathscr{A}_t$, with $\mathscr{A}$ a division algebra over R; hence, $\mathscr{A}$ is isomorphic either to R or to C or to Q because of the Frobenius theorem, and $\mathfrak{A}$ takes one of the forms listed in the cases 1–3.

Whenever b occurs, let us put $S = iE$ and $\mathscr{V} = \mathfrak{A} \cup \{S\}$. This is irreducible in X_R since $\mathfrak{A}$ is irreducible in X. Furthermore, $S^2 = -E \in \mathfrak{A}$ (since $\mathfrak{A}$ has a unit) and $S^{-1}\mathfrak{A}S \subset \mathfrak{A}$ (since $\mathfrak{A}$ is a linear-antilinear algebra). Thus, Lemma 1 applies with R and X_R in place of K and X, respectively. Then, by making use again of the Weyl and Frobenius theorems, $\mathfrak{A}$ must take one of the forms $2R_n$, $2C_{n/2}$, $2Q_{n/4}$, $R_n \oplus R_n$, $C_{n/2} \oplus C_{n/2}$, $Q_{n/4} \oplus Q_{n/4}$. The possibility $\mathfrak{A} = 2Q_{n/4}$ cannot occur, since the order h or $\mathfrak{A}$ in this case would be $n^2/4$, while a lower bound $n^2/2$ exists for h. (Indeed, let $\mathfrak{D}$ be the algebra over R generated by $\mathscr{V}$. Then, $\mathfrak{D}$ is irreducible in X_R, like $\mathscr{V}$, and its order g either is $4n^2$, or $2n^2$ or n^2, as follows from the cases 1–3 discussed above with $\mathfrak{D}$ in the place of $\mathfrak{A}$. Since g is in any case equal or twice the order of $\mathfrak{A}$, this cannot be less than $n^2/2$.) The five remaining cases are listed in the theorem as cases 4–8, respectively. ■

Theorem 2: Let X be a finite-dimensional vector space over the complex field C. With reference to Definitions 1–5, let ξ be any irreducible linear–antilinear representation in X of some algebra $\mathscr{A}$ over R such that $\mathfrak{A} = \xi(\mathscr{A})$ is a linear–antilinear graded algebra (over R) with unit element, whose antilinear part either is void or contains an antilinear mapping A together with its inverse. Then, $\mathfrak{A}$ is potentially real, pseudoreal or complex iff $\mathfrak{A}'$ is potentially real, pseudoreal, or complex, respectively. Moreover, $\mathfrak{A}$ is potentially real iff ξ is reducible in X_R (i.e., one of the cases 4–8 of Theorem 1 occurs); in this case, a subspace $X_1 \subset X_R$ is $\mathfrak{A}$ invariant iff X_1 is the subspace $R\langle\mathscr{B}\rangle$ of X_R generated by some basis $\mathscr{B}$ in X such that $\overline{M}_{\mathscr{B}} = M$ for every $M \in \mathfrak{A}$ (see Ref. 16) (canonical basis), and the direct sum decompositions of $\mathfrak{A}$ listed in Theorem 1 refer (in the sense specified by the Weyl theorem) to any decomposition $X_R = R\langle\mathscr{B}\rangle \oplus iR\langle\mathscr{B}\rangle$ (canonical decomposition), with $\mathscr{B}$ a canonical basis.

Furthermore, whenever $\mathfrak{A}$ is potentially real, the linear parts ξ_1^1 and ξ_1^1 of the (irreducible) representations ξ_1 and ξ_2 of $\mathscr{A}$ induced by ξ in $R\langle\mathscr{B}\rangle$ and $iR\langle\mathscr{B}\rangle$, respectively, are equivalent, and the following conditions are equivalent.

(i) $\mathfrak{A} = \xi_1(\mathscr{A}) \oplus \xi_2(\mathscr{A})$ (i.e., one of the cases 6–8 of Theorem 1 occurs).

(ii) $\mathfrak{A} \cup \mathfrak{A}'^R$ is reducible in X_R.

(iii) $\mathfrak{A} \cup \mathfrak{A}'^R$ is potentially real.

(iv) $\mathfrak{A}^a \neq \{0\}$ and ξ_1^1 (hence ξ_2^1) is irreducible in X_R.

Proof: First, we note that $\mathfrak{A}$, which is a multiplicative semigroup, can be classified according to Table II in Sec. II. From now on, we refer to this classification and to the notation in Table II. Let us come now to the first statement in the Theorem. The part of the statement that regards the complex case is obvious. Indeed, $\mathfrak{A}^1$ complex implies $\mathfrak{A}$ complex. Conversely, whenever $\mathfrak{A}$ is complex, i.e., $\mathfrak{A}'^a$ is void, then $iE \in (\mathfrak{A}'^R)'^R = \mathfrak{A}$ (see remark 2, Sec. III), hence $iE \in \mathfrak{A}^1$, so that $(\mathfrak{A}^1)'^a$ is void, i.e., $\mathfrak{A}^1$ is complex.

It is also straightforward that $\mathfrak{A}$ potentially real implies $\mathfrak{A}^1$ potentially real; the proof of our statement will be completed by showing that this last implication can be reversed. To this end, let us assume that $\mathfrak{A}^1$ is potentially real and let us consider the four possibilities which characterize the columns in Table II. Whenever $\mathfrak{A}^a = \{0\}$ (column 1), then $\mathfrak{A}^1 = \mathfrak{A}$, and our statement is trivial; whenever $\mathfrak{A}^1$ is reducible and $U_1^1 \sim U_2^1$ (column 4), it is straightforward. In order to discuss the remaining cases, let us observe that a conjugation J exists which commutes with $\mathfrak{A}^1$, since $\mathfrak{A}^1$ is potentially real. Coherently with Definition 2, for every mapping P of X, let us put $\overline{P} = JPJ$. Then, the following properties hold for the antilinear invertible mapping A which has been assumed to belong to $\mathfrak{A}$: (a) $\overline{A}^2 = A^2$ and (b) $A\,\overline{A}^{-1} \in (\mathfrak{A}^1)'^1$. Indeed, for every $L \in \mathfrak{A}^1$, $\overline{L} = L$; it follows $\overline{A}^2 = A^2$ (since $A^2 \in \mathfrak{A}^1$), hence, trivially, statement (a), and $\overline{A^{-1}LA} = A^{-1}LA$ (since $A^{-1}LA \in \mathfrak{A}^1$), hence $A\,\overline{A}^{-1}L = LA\,\overline{A}^{-1}$, which proves (b).

Furthermore, let us briefly put $M = A\,\overline{A}^{-1}$. Then, trivially, $\overline{M}M = E$, and $A\,M A^{-1}M = A^2\overline{A}^{-2}$; from the latter, by making use of a, we get $A\,MA^{-1}M = E$.

Now, let us assume that $\mathfrak{A}^1$ is irreducible (column 2) or reducible and such that $U_1^1 \not\sim U_2^1$ (column 3). Let us show that in both cases a mapping $N \in (\mathfrak{A}^1)'^1$ exists such that $N^2 = M$, $\overline{N}N = E$, $A\,NA^{-1}N = E$. To this end, let us recall[24] that whenever $\mathfrak{A}^1$ is irreducible, $(\mathfrak{A}^1)'^1 = CE$, while in our alternative case X decomposes into the direct sum of two $\mathfrak{A}^1$-invariant subspaces, $X = Y_1 \oplus Y_2$, with $Y_2 = A\,Y_1$; making reference to this decomposition we can write

$$(\mathfrak{A}^1)'^1 = \left\{ \begin{pmatrix} \alpha E_1 & 0 \\ 0 & \delta E_2 \end{pmatrix} : \alpha, \delta \in C \right\}.$$

Hence, by making use of statement (b) above, we get that either some (nonzero) $\alpha \in C$ exists such that $M = \alpha E$ or some (nonzero) α, $\delta \in C$ exists, with $\alpha \neq \delta$, such that

$$M = \begin{pmatrix} \alpha E_1 & 0 \\ 0 & \delta E_2 \end{pmatrix}.$$

In the former case, the mapping $N = \alpha^{1/2}E$, with the square roots of α and $\bar{\alpha}$ chosen in such a way that $\bar{\alpha}^{1,2} = \bar{\alpha}^{1/2}$, has the desired properties (the equation $\overline{N}N = E$ follows from $\bar{\alpha}\alpha = 1$, which follows in turn from $\overline{M}M = E$). In the latter case, let us recall first, making reference to the above decomposition of X, that the mapping A takes off-diagonal form (see the FS$\times$W Theorem in Sec. II); by substituting in the equation $A\,MA^{-1}M = E$ we obtain $\delta = \bar{\alpha}^{-1}$. Since $J^2 = E$, while $\alpha \neq \delta$ in the present case, it follows, by making use of the equation $\overline{M}M = E$ and with some simple calculations that we do not report here, that J also is off-diagonal in the above decomposition of X. Then, the mapping

$$N = \begin{pmatrix} \alpha^{1/2}E_1 & 0 \\ 0 & \bar{\alpha}^{-1/2}E_2 \end{pmatrix},$$

again with the square roots of α and $\bar{\alpha}$ chosen in such a way that $\overline{\alpha^{1/2}} = \bar{\alpha}^{1/2}$, has the required properties.

Let us consider now the mapping $J' = NJ$. This is involutory, since $J'^2 = \overline{N}N = E$, and commutes with $\mathfrak{A}^1$, since N and J belongs to $(\mathfrak{A}^1)'^R$. Furthermore, $J'A = NJA = N\overline{A}J = NM^{-1}A\quad J = N^{-1}A\,J = A\,NJ = AJ'$. Hence, $J' \in \mathfrak{A}'^a$, so that $\mathfrak{A}$ is potentially real.

This completes the proof of the first statement in the theorem.[25]

Before coming to the proof of the other statements we make the following remarks.

(i) For any basis $\mathscr{B}$ in X and for every M in the set $\mathscr{L}_R(X)$ of all the R-linear mappings of X, the identity

$$M = \tfrac{1}{2}(M + \overline{M}_{\mathscr{B}}) + \tfrac{1}{2}(M - \overline{M}_{\mathscr{B}})$$

holds. It turns out at once that the $\frac{1}{2}(M + \overline{M}_{\mathscr{B}})$ map into themselves the (proper) subspaces $R\langle\mathscr{B}\rangle$ and $iR\langle\mathscr{B}\rangle$, while $\frac{1}{2}(M - \overline{M}_{\mathscr{B}})$ map (bijectively) $R\langle\mathscr{B}\rangle$ onto $iR\langle\mathscr{B}\rangle$.

(ii) Whenever $\mathfrak{A}$ is reducible in X_R, let X_1 be any $\mathfrak{A}$-invariant subspace of X_R, $X_2 = iX_1$. Then, a positive integer p and two subspaces $Y_1 \subset X_1$ and $Y_2 \subset X_2$ exist such that $X_1 = pY_1$, $X_2 = pY_2$ and that the canonical forms of $\mathfrak{A}$ (see Theorem 1) occur in the decomposition $X = pY_1 \oplus pY_2$.

The proof of the remark (i) is straightforward. As for remark (ii), it can easily be deduced from the arguments in the proofs of Lemma 1 and Theorem 1.

Now, let $\mathfrak{A}$ be potentially real. Then, some basis $\mathscr{B}$ in X exists such that for any $M \in \mathfrak{A}$, $\overline{M}_{\mathscr{B}} = M$ (see Ref. 16). Because of remark (i), with $M - \overline{M}_{\mathscr{B}} = 0$, $X_1 = R\langle\mathscr{B}\rangle$ is an $\mathfrak{A}$-invariant (proper) subspace of X_R, hence $\mathfrak{A}$ is reducible in X_R. Moreover, because of remark (ii), the direct sum decompositions of $\mathfrak{A}$ listed in Theorem 1 refer to the decomposition $X_R = X_1 \oplus iX_1$ (see Ref. 26).

Conversely, let $\mathfrak{A}$ be reducible in X_R, let X_1 be any $\mathfrak{A}$-invariant subspace of X_R and let $\mathscr{B}$ be any basis in X_1. Because of remark (ii) the direct sum decompositions of $\mathfrak{A}$ listed in Theorem 1 refer to the decomposition $X_R = X_1 \oplus X_2 = R\langle\mathscr{B}\rangle \oplus iR\langle\mathscr{B}\rangle$. Furthermore, $\mathscr{B}$ is a basis in X, and in the above decomposition of X_R the conjugation $J_{\mathscr{B}}$ associated in X with $\mathscr{B}$ takes the matrix form

$$J_{\mathscr{B}} = \begin{pmatrix} I_1 & 0 \\ 0 & -I_2 \end{pmatrix},$$

with I_1 and I_2 the identity mappings in X_1 and X_2 respectively; thus, for any $M \in \mathfrak{A}$, $J_{\mathscr{B}}M = MJ_{\mathscr{B}}$ (since in the above decomposition also M is diagonal). Hence, $\mathfrak{A}$ is potentially real; furthermore, $\mathscr{B}$ is a canonical basis.

Thus, we have proved the second statement in the Theorem.[27] In order to prove the remaining statements, let us assume from now on that $\mathfrak{A}$ is potentially real and let us refer to the canonical decomposition of X_R when writing any mapping in matrix form. Furthermore, let us establish that $\mathscr{B}$ always denotes a canonical basis of X and that I_1 and I_2, X_1, X_2 have the same meaning as above.

Let us now prove the statement regarding ξ_1^1 and ξ_2^1 in the theorem. To this end, observe that the representations ξ_1 and ξ_2 turn out to be irreducible in the cases 4–8 of Theorem 1 by inspection of the canonical forms of $\mathfrak{A} = \xi(\mathscr{A})$. Furthermore,

$$iE = \begin{pmatrix} 0 & E_{12} \\ -E_{12}^{-1} & 0 \end{pmatrix},$$

with E_{12} a vector space isomorphism of iX_1 onto X_1. Let

$$L = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix} \in \mathfrak{A}^1 \subset \mathfrak{A};$$

since $iL = Li$, we get at once $L_2 = E_{12}^{-1} L_1 E_{12}$, which proves our statement.

Let us come now to the proof of the equivalence of the conditions (i)–(iv).

(i)$\Longrightarrow$(ii). Straightforward, by inspection of the canonical forms of $\mathfrak{A}$ and $\mathfrak{A}'^R$ listed in the cases 6–8 of Theorem 1.

(ii)$\Longrightarrow$(iii). Straightforward, because of the second statement in the theorem, with $\mathfrak{A}\cup\mathfrak{A}'^R$ in place of $\mathfrak{A}$.

(iii)$\Longrightarrow$(iv). Whenever $\mathfrak{A}\cup\mathfrak{A}'^R$ is potentially real, the conjugation $J_{\mathscr{C}}$ can be assumed to belong to $(\mathfrak{A}\cup\mathfrak{A}'^R)'^R = \mathfrak{A}'^R \cap \mathfrak{A}$ (remark 2), hence $J_{\mathscr{C}}$ belongs to $\mathfrak{A}^a$, so that $\{0\} \neq \mathfrak{A}^a = J_{\mathscr{C}} \mathfrak{A}^l$ (see Ref. 28) and $\mathfrak{A} = \mathfrak{A}^l \oplus J_{\mathscr{C}} \mathfrak{A}^l$. Should ξ_1^1 (hence ξ_2^1) be reducible, also ξ_1 and ξ_2 would be reducible, since

$$J_{\mathscr{C}} = \begin{pmatrix} I_1 & 0 \\ 0 & -I_2 \end{pmatrix}.$$

(iv)$\Longrightarrow$(i). Whenever ξ_1^1 is irreducible, we get from Theorem 1, with X_1 in place of X and $\xi_1(\mathscr{A})$ in place of $\mathfrak{A}$, that $\xi_1^1(\mathscr{A})$ is either R_n, or $C_{n/2}$ or $Q_{n/4}$. Since ξ_1^1 and ξ_2^1 are equivalent, it follows that either $\mathfrak{A}^l = 2R_n$ or $\mathfrak{A}^l = 2C_{n/2}$ or $\mathfrak{A}^l = 2Q_{n/4}$. Hence, by making use of the Weyl theorem and by recalling the last observation in remark 2, we obtain that the commutant (linear–antilinear graded) algebra $(\mathfrak{A}^l)'^R$ of $\mathfrak{A}^l$ simultaneously takes the forms $(nR)_2$, $[(n/2)C]_2$, $[(n/4)Q^0]_2$ respectively. Our proof will now be shortened by choosing a (canonical) basis $\mathscr{C} = \{e_j\}_{j=1,\ldots,n}$ in X, hence a basis $\{e_j\}_{j=1,\ldots,n} \cup \{ie_j\}_{j=1,\ldots,n} = \mathscr{C} \cup i\mathscr{C}$ in X_R, and by considering the matrix $\mathscr{M}(S)$ that realizes any R-linear mapping S with respect to this basis. In particular, with this choice, we get

$$\mathscr{M}(iE) = \begin{pmatrix} 0 & -\mathscr{I}_n \\ \mathscr{I}_n & 0 \end{pmatrix}$$

(with $\mathscr{I}_n$ the identity in the set of the $n \times n$ matrices with elements in R). Furthermore, for every

$$L = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix} \in \mathfrak{A}^l,$$

the matrices $\mathscr{M}_1(L_1)$ and $\mathscr{M}_2(L_2)$ that realize L_1 in $X_1 = R\langle\mathscr{C}\rangle$ and L_2 in $X_2 = iR\langle\mathscr{C}\rangle$, respectively, must be equal [since $\mathscr{M}(L)$ commutes with $\mathscr{M}(iE)$]. Let us assume that the basis $\mathscr{C}$ has been chosen so that the algebra $\mathscr{M}_1(\xi_1^1(\mathscr{A}))$ takes canonical form, i.e., so that it coincides with either $\mathscr{R}_n$, or $\mathscr{C}_{n/2}$ or $\mathscr{Q}_{n/4}$; here, $\mathscr{R} = R$, $\mathscr{C}$, and $\mathscr{Q}$ are some regular matrix representations of R, C, and Q, respectively.[29] Hence, in the basis $\mathscr{C} \cup i\mathscr{C}$ the forms of the algebras $\mathscr{M}(\mathfrak{A}^l)$ and $\mathscr{M}((\mathfrak{A}^l)'^R)$ follow easily from the forms of $\mathfrak{A}^l$ and $(\mathfrak{A}^l)'^R$ written above; furthermore [since $(\mathfrak{A}^l)'^R$ commutes with iE] we get

$$\mathscr{M}((\mathfrak{A}^l)'^l) = \left\{ \begin{pmatrix} r\mathscr{I}_p & s\mathscr{I}_p \\ -s\mathscr{I}_p & r\mathscr{I}_p \end{pmatrix} : r,s \in \mathscr{K} \right\},$$

where p can take the values $n, n/2, n/4$, and $\mathscr{K}$ denotes a regular representation of R, C, Q^0.

Let $G \in (\mathfrak{A}^l)'^l$,

$$\mathscr{M}(G) = \begin{pmatrix} r\mathscr{I}_p & s\mathscr{I}_p \\ -s\mathscr{I}_p & r\mathscr{I}_p \end{pmatrix}.$$

Since we have hypothesized $\mathfrak{A}^a \neq \{0\}$, we have $G \in \mathfrak{A}'^l \subset (\mathfrak{A}^l)'^l$ iff $GN = NG$ for some $N \in \mathfrak{A}^a$ [equivalently, for all $N \in \mathfrak{A}^a$ (see Ref. 28)]. In the canonical decomposition

$$N = \begin{pmatrix} N_1 & 0 \\ 0 & N_2 \end{pmatrix},$$

so that, since $Ni = -iN$,

$$\mathscr{M}(N) = \begin{pmatrix} \mathscr{N} & 0 \\ 0 & \mathscr{N} \end{pmatrix},$$

with $\mathscr{N}$ an $n \times n$ matrix with elements in R. Therefore, $G \in \mathfrak{A}'^l$ iff (a) $(r\mathscr{I}_p)\mathscr{N} = \mathscr{N}(r\mathscr{I}_p)$ and (b) $(s\mathscr{I}_p)\mathscr{N} = -\mathscr{N}(s\mathscr{I}_p)$. Let us denote by $\mathscr{L}$ and $\mathscr{H}$ the subsets of the elements of $\mathscr{K}$ that satisfy (a) and (b) respectively. Then, it follows from (a) and (b) that $\mathscr{H} \cap \mathscr{L} = \{0\}$. This property, together with the fact that $\mathfrak{A}'^l$ is a division subalgebra of $(\mathfrak{A}^l)'^l$ since $\mathfrak{A}$ is irreducible in X (see Ref. 30) implies, after some simple calculations that we do not report here, that the off-diagonal elements of $\mathscr{M}(G)$ vanish (hence, $\mathscr{L} = \{0\}$). Therefore, also the off-diagonal elements of every $P \in \mathfrak{A}'^a$ vanish; indeed, $J_{\mathscr{C}} \in \mathfrak{A}'^a$, so that $\mathfrak{A}'^a = J_{\mathscr{C}} \mathfrak{A}'^l$ (see Ref. 31) while, as we have previously proved,

$$J_{\mathscr{C}} = \begin{pmatrix} I_1 & 0 \\ 0 & -I_2 \end{pmatrix}.$$

By inspection of the cases 4–8 in Theorem 1 we see that in the canonical decomposition the off-diagonal elements of $\mathfrak{A}'^R = \mathfrak{A}'^l \oplus \mathfrak{A}'^a$ vanish iff $\mathfrak{A}$ is the direct sum of $\xi_1(\mathscr{A})$ and $\xi_2(\mathscr{A})$. ■

By making use of Theorems 1 and 2 and of Lemma 1, we can now prove a new theorem which, as we have anticipated in the Introduction, recovers Dyson's classification under more general assumptions and in a generalized framework, and shows its one-to-one correspondence with the generalized FS$\times$W classification particularized to C.

Theorem 3: Let X be a finite-dimensional vector space over the complex field C. With reference to Definitions 1–5, let $\mathscr{S}$ be a semigroup, let U be an irreducible linear–antilinear representation of $\mathscr{S}$ in X such that the antilinear part of $\mathscr{U} = U(\mathscr{S})$ is either void or contains an invertible mapping A together with its inverse, and let $\mathfrak{A}$ be the algebra over R generated by $\mathscr{U}$. Then, $\mathfrak{A}$ is a linear–antilinear (graded) algebra, and in each of the 13 mutually exclusive cases which appear in the generalized FS$\times$W classification $\mathfrak{A}$ takes one and only one of the eight possible forms listed in Theorem 1, according to the correspondence exhibited in Table III. Moreover, let $\mathfrak{D}$ be the (linear–antilinear graded) algebra over R generated by $\mathfrak{A} \cup \{iE\}$. Then, each case of the generalized FS$\times$W classification is characterized by the structure (in the sense of Theorem 1) of the pair $(\mathfrak{D}, \mathfrak{A}^l)$ in the space X_R, as shown in Table III. (In Table III, the sysmbols U_1^1, U_2^1, E_1, E_2, T_{12} have the same meaning as in Table II of Sec. II; the explanation of the other symbols used in Table III is given in the definitions and in Table I.)

Proof. The statement that $\mathfrak{A}$ and $\mathfrak{D}$ are linear–antilinear graded algebras is straightforward.

The characterization of the columns in Table III immediately follows from the characterization in Table II whenever $K = C$ (the equation $\mathfrak{A}'^l = \mathscr{U}'^l$ being obvious); the char-

TABLE III. One-to-one correspondence between the generalized FS×W classification and the Dyson classification of the irreducible linear–antilinear representations.

	$\mathscr{U}^a=\{0\}$	$\mathscr{U}^a\neq\{0\}$		
	$\mathscr{U}^1$ irreducible (in X)		$\mathscr{U}^1$ reducible (in X)	
			$U_1^1 \not\sim U_2^1$	$U_1^1 \sim U_2^1$
	$\mathfrak{A}'^1=\mathscr{U}'^1=CE$	$\mathfrak{A}'^1=\mathscr{U}'^1=RE$	$\mathfrak{A}'^1=\mathscr{U}'^1=\left\{\begin{pmatrix}\alpha E_1 & 0\\ 0 & \bar{\alpha}E_2\end{pmatrix}:\alpha\in C\right\}\approx C$	$\mathfrak{A}'^1=\mathscr{U}'^1=\left\{\begin{pmatrix}\alpha E_1 & \beta T_{12}\\ \bar{\beta}\eta T_{12}^{-1} & \bar{\alpha}E_2\end{pmatrix}:\alpha,\beta\in C\right\}\approx Q$
		$\mathfrak{A}=R_n\oplus R_n$ $\begin{cases}\mathfrak{D}=R_{2n}\\ \mathfrak{A}^1=2R_n\end{cases}$	$(\mathscr{U}^c\approx C_*\times G_2)$ $\mathfrak{A}=C_{n/2}\oplus C_{n/2}$ $\begin{cases}\mathfrak{D}=C_n\\ \mathfrak{A}^1=2C_{n/2}\end{cases}$	$(\mathscr{U}^c\approx Q_*\times G_2)$ $\mathfrak{A}=Q_{n/4}\oplus Q_{n/4}$ $\begin{cases}\mathfrak{D}=Q_{n/2}\\ \mathfrak{A}^1=2Q_{n/4}\end{cases}$
$\mathscr{U}$ is potentially real	$\mathfrak{A}=2R_n$ $\begin{cases}\mathfrak{D}=C_n\\ \mathfrak{A}^1=2R_n\end{cases}$		$(\mathscr{U}^c\not\approx C_*\times G_2)$ $\mathfrak{A}=2R_n$ $\begin{cases}\mathfrak{D}=C_n\\ \mathfrak{A}^1=2(R_{n/2}\oplus R_{n/2})\end{cases}$	$(\mathscr{U}^c\not\approx Q_*\times G_2)$ $\mathfrak{A}=2C_{n/2}$ $\begin{cases}\mathfrak{D}=Q_{n/2}\\ \mathfrak{A}^1=4R_{n/2}\end{cases}$
$\mathscr{U}$ is pseudoreal	$\mathfrak{A}=Q_{n/2}$ $\begin{cases}\mathfrak{D}=C_n\\ \mathfrak{A}^1=Q_{n/2}\end{cases}$	$\mathfrak{A}=C_n$ $\begin{cases}\mathfrak{D}=R_{2n}\\ \mathfrak{A}^1=Q_{n/2}\end{cases}$	$\mathfrak{A}=Q_{n/2}$ $\begin{cases}\mathfrak{D}=C_n\\ \mathfrak{A}^1=Q_{n/4}\oplus Q_{n/4}\end{cases}$	
$\mathscr{U}$ is complex	$\mathfrak{A}=C_n$ $\begin{cases}\mathfrak{D}=C_n\\ \mathfrak{A}^1=C_n\end{cases}$	$\mathfrak{A}=R_{2n}$ $\begin{cases}\mathfrak{D}=R_{2n}\\ \mathfrak{A}^1=C_n\end{cases}$	$\mathfrak{A}=C_n$ $\begin{cases}\mathfrak{D}=C_n\\ \mathfrak{A}^1=C_{n/2}\oplus C_{n/2}\end{cases}$	$\mathfrak{A}=Q_{n/2}$ $\begin{cases}\mathfrak{D}=Q_{n/2}\\ \mathfrak{A}^1=2C_{n/2}\end{cases}$

acterization of the rows coincides with the one in Table II (in the sequel, the words "column" and "row" will always refer to Table III).

Before coming to the proof of the statements in the squares of Table III we establish that $\mathfrak{A}$ is considered as the range of a representation ξ of some abstract algebra $\mathscr{A}$ in the vector space X_R; whenever $\mathfrak{A}$ is reducible, the representations ξ_1 and ξ_2 will then be defined as in Theorem 2. Furthermore, we premise the following remark.

Remark (α): Let l, h, g be the orders of the algebras $\mathfrak{A}^1$, $\mathfrak{A}$, $\mathfrak{D}$, respectively. Then, by making reference to the generalized FS×W classification in Table II, we have the following relations between the orders:

column 1, rows 1, 2:	$l=h=\frac{1}{2}g$
column 1, row 3:	$l=h=g$
columns 2,3, 4, rows 1, 2:	$l,=\frac{1}{2}h=\frac{1}{4}g$
columns 2, 3, 4, row 3:	$l=\frac{1}{2}h=\frac{1}{2}g$.

Indeed, $l=h$ whenever $\mathscr{U}^a=\{0\}$, since $\mathfrak{A}=\mathfrak{A}^1$ in this case, while, trivially,[28] $l=\frac{1}{2}h$ whenever $\mathscr{U}^a\neq\{0\}$. Furthermore, $h=g$ whenever $\mathscr{U}$ is complex, since $iE\in\mathfrak{A}^1$ in this case, as we have already observed in the proof of Theorem 2, so that $\mathfrak{A}=\mathfrak{D}$. Finally, $h=\frac{1}{2}g$ in the other cases, since $iE\notin\mathfrak{A}$.

Now, let us consider the algebra $\mathfrak{D}$. We have already observed in the proof of Theorem 1 that the set $\mathfrak{A}\cup\{iE\}$ is irreducible in X_R, since $\mathfrak{A}$ is irreducible in X. Therefore, $\mathfrak{D}$ is irreducible both in X and in X_R, so that, by making use of Theorem 1 with $\mathfrak{D}$ in place of $\mathfrak{A}$, we get that either $\mathfrak{D}=R_{2n}$ or $\mathfrak{D}=C_n$ or $\mathfrak{D}=Q_{n/2}$; hence, either $\mathfrak{D}'^R=2nR$, or $\mathfrak{D}'^R=nC$ or $\mathfrak{D}'^R=(n/2)Q^0$, respectively. Since, trivially, $\mathfrak{D}'^R=\mathfrak{A}'^1$, it follows (by comparison with the forms of $\mathfrak{A}'^1$ which characterize the columns) that $\mathfrak{D}=C_n$, $\mathfrak{D}=R_{2n}$, $\mathfrak{D}=C_n$, $\mathfrak{D}=Q_{n/2}$, respectively, in columns 1, 2, 3, 4 (with $g=2n^2$, $g=4n^2$, $g=2n^2$, and $g=n^2$, respectively).

Let us consider now the algebra $\mathfrak{A}$. Let $\mathscr{U}$ be complex (row 3). Then, as we have already seen in the proof of remark (α), $\mathfrak{A}=\mathfrak{D}$. Let $\mathscr{U}$ be pseudoreal (row 2). In this case, $\mathfrak{A}$ is irreducible in X_R because of the second statement in Theorem 2 and hence one of the cases 1–3 of Theorem 1 occurs. Furthermore, it follows from remark (α) that $h=\frac{1}{2}g$; since $\mathfrak{D}$ is known, we get $h=n^2$, $h=2n^2$, $h=n^2$, and therefore $\mathfrak{A}=Q_{n/2}$, $\mathfrak{A}=C_n$, $\mathfrak{A}=Q_{n/2}$ in columns 1, 2, 3, respectively. Let $\mathscr{U}$ be potentially real (row 1). Then, $\mathfrak{A}$ is reducible in X_R because of the second statement in Theorem 2 and hence one of the cases 4–8 of Theorem 1 occurs.

Whenever $\mathscr{U}'^1=CE$ (column 1), $\mathfrak{A}=\mathfrak{A}^1$ (since $\mathscr{U}^a=\{0\}$); then, since ξ_1^1 and ξ_2^1 are equivalent because of a statement in Theorem 2, one of the cases 4–5 of Theorem 1 occurs; from remark (α), since $\mathfrak{D}=C_n$, we get $h=n^2$, hence $\mathfrak{A}=2R_n$. In order to find $\mathfrak{A}$ in the other cases of row 1, let us recall that $\mathscr{U}^c\approx\mathscr{U}^{c1}\times G_2$ iff $\mathscr{U}\cup\mathscr{U}^c$, equivalently $\mathfrak{A}\cup\mathfrak{A}'^R$, is potentially real,[32] that is, because of the equivalence between the conditions (i) and (iii) in Theorem 2, iff one of the cases 6–8 of Theorem 1 occurs. Therefore, whenever $\mathscr{U}^c$ is not a direct product, one of the cases 4–5 of Theorem 1 occurs. Since from remark (α) we get $h=2n$, n^2, and $n^2/2$ in columns 2, 3, and 4, respectively, we conclude that $\mathfrak{A}=R_n\oplus R_n$ in column 2, while $\mathfrak{A}=C_{n/2}\oplus C_{n/2}$ or $\mathfrak{A}=2R_n$ in column 3 and $\mathfrak{A}=Q_{n/4}\oplus Q_{n/4}$ or $\mathfrak{A}=2C_{n/2}$ in column 4, depending whether $\mathscr{U}^c$ is a direct product or not.

Let us consider now the algebra $\mathfrak{A}^1$. Let $\mathscr{U}^a=\{0\}$ (column 1). Here, $\mathfrak{A}^1=\mathfrak{A}$ in every row. Let $\mathscr{U}^a\neq\{0\}$, $\mathscr{U}^1$ irreducible (column 2). By considering $\mathfrak{A}^1$ in place of $\mathscr{U}$ and applying the results in column 1, we get that either $\mathfrak{A}^1=2R_n$, or $\mathfrak{A}^1=Q_{n/2}$, or $\mathfrak{A}^1=C_n$; furthermore, we get that these forms of $\mathfrak{A}^1$ occur in rows 1, 2, and 3, respectively by making use of the first statement in Theorem 1.

Let $\mathscr{U}^1$ be reducible in X (columns 3 and 4). First, let us consider the cases of $\mathscr{U}$ potentially real, and of $\mathscr{U}^c$ isomorphic to a direct product. In these cases, as we have already noted above, $\mathfrak{A}\cup\mathfrak{A}'^R$ is potentially real, hence ξ^1_1 and ξ^1_2 are irreducible because of the equivalence between (iii) and (iv) in Theorem 2. Hence [as we have already seen in the proof of (iv)$\Longrightarrow$(i) in Theorem 2], $\mathfrak{A}^1$ takes one of the forms $2R_n$, $2C_{n/2}$, $2Q_{n/4}$. Since $l = n^2/2$ and $l = n^2/4$ in columns 3 and 4, respectively [remark (α)], it follows $\mathfrak{A}^1 = 2C_{n/2}$ in the former case, $\mathfrak{A}^1 = 2Q_{n/4}$ in the latter.

Second, let us consider the cases of $\mathscr{U}$ potentially real, and of $\mathscr{U}^c$ not isomorphic to a direct product. Let $\mathscr{E}$ be a canonical basis in X. Then, in the canonical decomposition $X_R = X_1 \oplus X_2$, with $X_1 = R\langle\mathscr{E}\rangle$, $X_2 = iR\langle\mathscr{E}\rangle$, every $M\in\mathfrak{A}$ takes diagonal form (Theorem 2), so that, in particular, the invertible mapping $A\in\mathscr{U}^a$ can be written in the form

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}.$$

Let us put $\mathfrak{A}^1_1 = \xi^1_1(\mathscr{A})$. Then, $A_1^{-1}\mathfrak{A}^1_1 A_1 \subset \mathfrak{A}^1_1$, $A_1^2 \in \mathfrak{A}^1_1$; furthermore, $\mathfrak{A}^1_1$ is a reducible algebra over R because of the equivalence between the conditions (iii) and (iv) in Theorem 2, while the set $\mathfrak{A}^1_1\cup\{A_1\}$ is irreducible since it is a set of generators for $\xi_1(\mathscr{A})$, which is irreducible (see again Theorem 2). Hence, Lemma 1 applies with the substitution $K\to R$, $X\to X_1$, $\mathfrak{A}\to\mathfrak{A}^1_1$, $\xi\to\xi^1_1$, $S\to A_1$, so that $X_1 = Z\oplus A_1 Z$ (hence, $X_1 = Z\oplus AZ$), with Z and $A_1 Z$ $\mathfrak{A}^1_1$-invariant subspaces, and ξ^1_1 reduces into two equidimensional irreducible subrepresentations, ξ^1_{11} and ξ^1_{12}, while the algebras $\mathfrak{A}^1_{11} = \xi^1_{11}(\mathscr{A})$ and $\mathfrak{A}^1_{12} = \xi^1_{12}(\mathscr{A})$ are isomorphic. Then, by making use of Theorem 1, we obtain that $\mathfrak{A}^1_{11}$ and $\mathfrak{A}^1_{12}$ must take one of the forms $R_{n/2}$, $C_{n/4}$, $Q_{n/8}$, hence $\mathfrak{A}^1_1$ must take one of the forms $2R_{n/2}$, $2C_{n/4}$, $2Q_{n/8}$, $R_{n/2}\oplus R_{n/2}$, $C_{n/4}\oplus C_{n/4}$, $Q_{n/8}\oplus Q_{n/8}$; the corresponding forms of $\mathfrak{A}^1$ are obtained by doubling the latter ones, since ξ^1_1 and ξ^1_2 are equivalent (Theorem 2). Now, we get $l = n^2/2$ and $l = n^2/4$ in columns 3 and 4, respectively [remark (α)]. Hence, $\mathfrak{A}^1 = 2(R_{n/2}\oplus R_{n/2})$ in column 3. In column 4, the order limits the possibilities for $\mathfrak{A}^1$ to $4R_{n/2}$ and $2(C_{n/4}\oplus C_{n/4})$. In order to eliminate the latter possibility, let us observe that, since $X_R = X_1 \oplus iX_1$, then $X_R = Z\oplus AZ\oplus iZ\oplus iAZ$. Let us put $Y_1 = Z\oplus iZ$ and $Y_2 = AY_1$. Then, Y_1 and Y_2 are vector spaces over C which are $\mathfrak{A}^1$-invariant subspaces of X, hence the representations U^1_1 and U^1_2 of $\mathfrak{A}^1$ in Y_1 and Y_2 respectively must be equivalent because of the assumptions which characterize the cases in column 4. Should $\mathfrak{A}^1$ take the form $2(C_{n/2}\oplus C_{n/2})$, U^1_1 and U^1_2 would easily turn out to be inequivalent. Hence, $\mathfrak{A}^1 = 4R_{n/2}$.

Third, let us consider the cases $\mathscr{U}$ pseudoreal and $\mathscr{U}$ complex. In these cases $\mathfrak{A}$ is irreducible in X_R, while $\mathfrak{A}^1$ is reducible. Furthermore, $A^{-1}\mathfrak{A}^1 A\subset\mathfrak{A}^1$, $A^2\in\mathfrak{A}^1$ and, trivially, $\mathfrak{A}^1\cup\{A\}$ is irreducible in X_R. Thus, Lemma 1 applies with the substitutions $K\to R$, $X\to X_R$, $\xi\to\xi^1$, $\mathfrak{A}\to\mathfrak{A}^1$, $S\to A$, hence ξ^1 reduces into two equidimensional irreducible subrepresentations ξ^1_1 and ξ^1_2, while the algebras (over R) $\mathfrak{A}^1_1 = \xi^1_1(\mathscr{A})$ and $\mathfrak{A}^1_2 = \xi^1_2(\mathscr{A})$ are isomorphic. Then, six forms must be taken into account for $\mathfrak{A}^1$, precisely $2R_n$, $2C_{n/2}$, $2Q_{n/4}$, $R_n\oplus R_n$, $C_{n/2}\oplus C_{n/2}$, $Q_{n/4}\oplus Q_{n/4}$. Since $Y_1 = \mathfrak{A}^1_1 X_R$ is any $\mathfrak{A}^1$-invariant subspace of X_R (see the proof of Lemma 1), Y_1, hence $Y_2 = \mathfrak{A}^1_2 X_R = AY_1$, can be assumed to be iE invariant (since $\mathscr{U}^1$ is reducible in X) so that Y_1 and Y_2 are vector spaces over C. Therefore, in column 4, row 3, $\mathfrak{A}^1_1$ and $\mathfrak{A}^1_2$ must be equivalent in X_R as a consequence of the assumption $U^1_1 \sim U^1_2$ which characterizes column 4; moreover we get $l = n^2/2$ [remark (α)] in this case, so that $\mathfrak{A}^1 = 2C_{n/2}$. In order to discuss the remaining cases in column 3, let us observe that, with reference to the decomposition $X_R = Y_1\oplus Y_2$, we can write

$$iE = \begin{pmatrix} iE_1 & 0 \\ 0 & iE_2 \end{pmatrix},$$

with E_1, E_2 the identity mappings in Y_1, Y_2, respectively; furthermore, $\mathfrak{A}^1_1$ and $\mathfrak{A}^1_2$ are equivalent (in X_R) iff a (involutory) mapping

$$T = \begin{pmatrix} 0 & T_{12} \\ T_{12}^{-1} & 0 \end{pmatrix}$$

exists which commutes with $\mathfrak{A}^1$. Now, let $\mathscr{U}$ be complex (row 3). Since $iE\in\mathfrak{A}^1$ in this case [see the proof of remark (α)], should $\mathfrak{A}^1_1$ be equivalent to $\mathfrak{A}^1_2$ in X_R, the mapping T_{12} would satisfy the equation $T_{12}(iE_2) = (iE_1)T_{12}$, i.e., T_{12} would be C linear, hence $U^1_1 \sim U^1_2$, contrary to the assumption $U^1_1 \nsim U^1_2$ which characterizes column 3. Since $l = n^2$ [remark (α)], it follows $\mathfrak{A}^1 = C_{n/2}\oplus C_{n/2}$.

Let $\mathscr{U}$ be pseudoreal (row 2). Then, $l = n^2/2$ [remark (α)]; hence, $\mathfrak{A}^1$ must take one of the forms $2C_{n/2}$, $Q_{n/4}\oplus Q_{n/4}$. Let us assume that $\mathfrak{A}^1 = 2C_{n/2}$. Then, $(\mathfrak{A}^1)'^R = [(n/2)C]_2$ (see remark 2 in Sec. III). From now on, let us choose suitable bases $\mathscr{E}_1$, $\mathscr{E}_2$ in Y_1, Y_2 respectively, hence a basis $\mathscr{E}_1\cup\mathscr{E}_2$ in X_R, and let $\mathscr{M}(S)$ be the matrix that realizes any R-linear mapping S with respect to this basis. Let us denote with $\mathscr{I}_p$ the identity in the set of the $p\times p$ matrices with elements in R. Then, since $iE\in(\mathfrak{A}^1)'^R$ and $(iE)^2 = -E$, while iE takes diagonal form, we get

$$\mathscr{M}(iE) = \begin{pmatrix} \alpha\mathscr{I}_{n/2} & 0 \\ 0 & \beta\mathscr{I}_{n/2} \end{pmatrix},$$

with α,β belonging to some regular representation of C, and $\alpha^2 = \beta^2 = -\mathscr{I}_2$. Therefore, $\alpha = -\beta$, since $iE\notin\mathfrak{A}^1 = 2C_{n/2}$. Let us consider the involutory mapping $T\in(\mathfrak{A}^1)'^R$ whose matrix is given by

$$\mathscr{M}(T) = \begin{pmatrix} 0 & \gamma\mathscr{I}_{n/2} \\ \gamma^{-1}\mathscr{I}_{n/2} & 0 \end{pmatrix},$$

with γ belonging to the aforesaid regular representation of C. Then,

$$\mathscr{M}(Ti) = \begin{pmatrix} 0 & -\gamma\alpha\mathscr{I}_{n/2} \\ \gamma^{-1}\alpha\mathscr{I}_{n/2} & 0 \end{pmatrix} = -\mathscr{M}(iT),$$

that is, T is antilinear, hence $\mathfrak{A}^1$ is potentially real. Because of the first statement in Theorem 2, this implies $\mathfrak{A}$ potentially real, contrary to the assumption $\mathfrak{A}$ pseudoreal in row 2. Thus, $\mathfrak{A}^1 = 2C_{n/2}$ is impossible; hence $\mathfrak{A}^1 = Q_{n/4}\oplus Q_{n/4}$. ■

Remark 4 (on the Dyson classification of matrix algebras): As we have anticipated in the Introduction, the characterization of each case in Table III by means of the pair $(\mathfrak{D}, \mathfrak{A}^1)$ coincides by inspection with the characterization of each case in Dyson's classification.[6] However, the Dyson classification is obtained under the following restrictive assumptions: (a) a scalar product is assumed on X and the matrices of $\mathscr{U}$ are supposed to be either unitary or antiunitary, and (b) $\mathscr{U}$ is a group. Thus, our classification also generalizes Dyson's.

In particular, Dyson's cases (40)–(42) in Ref. 6, respectively, coincide with the cases in column 1 of Table III, rows 1–3, while Dyson's cases (61)–(70) are in one-to-one correspondence with the cases in columns 2–4 of Table III.[33]

In order to underline some further connections between the present work and Dyson's paper, let us say that an algebra (over R) of R-linear mappings of X_R is of "type R," "type C," or "type Q" whenever its canonical form is $\oplus_{j=1}^{p} s_j\, R_{t_j}$, $\oplus_{j=1}^{p} s_j\, C_{t_j}$ or $\oplus_{j=1}^{p} s_j\, Q_{t_j}$, respectively. Then, it follows at once from the first statement in Theorem 2 and from Theorem 3 that, whenever $\mathscr{U}^1$ is irreducible (columns 1 and 2 in Table III), then $\mathfrak{A}^1$ is potentially real, pseudoreal, or complex iff it is of type R, Q, or C, respectively, This result clearly rephrases in our generalized framework Dyson's equivalence Theorem I (it must be noted that the above equivalences do not hold whenever $\mathscr{U}^1$ is reducible, as shown by the cases in columns 3, 4, row 1, $\mathscr{U}^c$ isomorphic to a direct product, and by the examples in the next section).

Finally, we notice that Dyson's equivalence Theorem II also follows at once from Table III in our generalized framework.

Remark 5: Let R_+, R_- respectively, denote the positive and negative reals; then, it follows in particular from Theorem 3 that, whenever $\mathscr{U}'^1$ is isomorphic to Q (column 4 in Table III), the statements

(a) $\mathscr{U}\cup\mathscr{U}^c$ is potentially real,
(b) $\mathscr{U}\cup\mathscr{U}^c$ is pseudoreal,

can respectively be characterized as follows: (a′) $\mathscr{U}^{ca}$ [equivalently, $(\mathscr{U}\cup\mathscr{U}^c)^{ca}$] is nonvoid and for any (equivalently, for some) $T\in(\mathscr{U}\cup\mathscr{U}^c)^{ca}$, $T^2\in R_+E$; and (b′) $\mathscr{U}^{ca}$ [equivalently, $(\mathscr{U}\cup\mathscr{U}^c)^{ca}$] is nonvoid and for any (equivalently, for some) $T\in(\mathscr{U}\cup\mathscr{U}^c)^{ca}$,] $T^2\in R_-E$ (see Ref. 34). In order to prove this characterization, let us preliminarily observe that, whenever $(\mathscr{U}\cup\mathscr{U}^c)^{ca}\neq\varnothing$, then $T^2\in\mathfrak{D}\cap\mathfrak{D}'^R$ for any $T\in(\mathscr{U}\cup\mathscr{U}^c)^{ca}$. Indeed, $T^2\in(\mathscr{U}\cup\mathscr{U}^c)^{cl}=(\mathscr{U}^{cc}\cap\mathscr{U}^c)^l$. Now, $\mathscr{U}^{cl}\subset\mathfrak{A}'^l=\mathfrak{D}'^R$. Furthermore, as we have already observed in the proof of Theorem 2, $\mathfrak{A}'^l$ is a division algebra over R, since $\mathfrak{A}$ is irreducible in X, while $\mathfrak{A}'^R=\mathfrak{A}'^l\oplus\mathfrak{A}'^a=\mathfrak{A}'^l\oplus T\mathfrak{A}'^l$ (see Refs. 23 and 28), hence, $\mathscr{U}^c=(\mathfrak{A}'^l\cup\mathfrak{A}'^a)\setminus\{0\}$, and therefore, $(\mathscr{U}^{cc})^l\subset\mathscr{U}^{cc}\subset(\mathfrak{A}'^R)'^R=\mathfrak{A}\subset\mathfrak{D}$ (see remark 2). Thus, $(\mathscr{U}^{cc}\cap\mathscr{U}^c)^l\subset\mathfrak{D}\cap\mathfrak{D}'^R$, hence $T^2\in\mathfrak{D}\cap\mathfrak{D}'^R$, as stated. Then let us come to the following properties, which hold whenever $\mathscr{U}'^l\approx Q$.

First, by making use of the above result, we note that in this case $T^2\in RE$ for any $T\in(\mathscr{U}\cup\mathscr{U}^c)^{ca}$; indeed, by inspection of Table III we get $\mathfrak{D}\cap\mathfrak{D}'^R=Q_{n/2}\cap(n/2)Q^0=2nR$. Second, it follows at once by inspection of Table II that $\mathscr{U}^{ca}$ is nonvoid iff $(\mathscr{U}\cup\mathscr{U}^c)^{ca}$ is nonvoid. Third, let $T, T'\in(\mathscr{U}\cup\mathscr{U}^c)^{ca}\neq\varnothing$. Then, $T'=TL$ (see Ref. 28), with $L\in(\mathscr{U}\cup\mathscr{U}^c)^{cl}\subset\mathfrak{D}\cap\mathfrak{D}'^R=2nR$, so that some $l\in R$ exists such that $T'=lT$, hence $T'^2=l^2T^2$, i.e., by setting $T^2=rE$ with $r\in R$, $T'^2=rl^2E$.

Now, the proof of the equivalences in (a′), (b′) and the proof that (a) implies (a′) are straightforward. Furthermore, whenever $T\in(\mathscr{U}\cup\mathscr{U}^c)^{ca}$ and $T^2=rE$, with $r\in R_+$, then the mapping $|r^{-1/2}|T$ is involutory and belongs to $(\mathscr{U}\cup\mathscr{U}^c)^{ca}$, hence (a′) implies (a). This suffices to prove our characterization above, the possibilities (a) and (b), so as (a′) and (b′), being mutually exclusive.

Remark 6: With reference to the generalized FS×W classification we notice that a particular case which is physically important[35] occurs whenever $\mathscr{U}=U(\mathscr{S})$ is a factorizable group (we recall that $\mathscr{U}$ is said to be factorizable whenever, for some—equivalently, for every[36]—$S\in\mathscr{U}^a$, the mapping θ: $L\in\mathscr{U}^1\to S^{-1}L\,S\in\mathscr{U}^1$ is an inner automorphism of $\mathscr{U}^1$). In order to discuss this case, let us observe first that, whenever $\mathscr{U}$ is a group, the following conditions are equivalent: (i) $\mathscr{U}$ is factorizable, and (ii) $(\mathscr{U}\cap\mathscr{U}^c)^a$ is nonvoid. For, it is immediate to show that $\mathscr{U}$ is factorizable iff for every $S\in\mathscr{U}^a$ a mapping $M_S\in\mathscr{U}^1$ exists such that $S\,M_S^{-1}\in\mathscr{U}^{ca}$. Since $S\,M_S^{-1}\in\mathscr{U}^a$, $\mathscr{U}$ factorizable implies $(\mathscr{U}\cap\mathscr{U}^c)^a\neq\varnothing$. Conversely, let $T\in(\mathscr{U}\cap\mathscr{U}^c)^a$; then, $\mathscr{U}^a=T\mathscr{U}^1$ (see Ref. 28), hence for every $S\in\mathscr{U}^a$ an $L_S\in\mathscr{U}^1$ exists such that $S=TL_S$. Thus, since $T\,L=L\,T$ for every $L\in\mathscr{U}^1$, we get $S^{-1}L\,S=L_S^{-1}T^{-1}L\,TL_S=L_S^{-1}LL_S$ for every $L\in\mathscr{U}^1$, and therefore $\mathscr{U}$ is factorizable.

Second, observe that $\mathscr{U}\cap\mathscr{U}^c\subset\mathscr{U}^{cc}\cap\mathscr{U}^c=(\mathscr{U}\cup\mathscr{U}^c)^c$.

Now, let $\mathscr{U}$ be factorizable; then, $(\mathscr{U}\cup\mathscr{U}^c)^{ca}$ is nonvoid, that is $\mathscr{U}\cup\mathscr{U}^c$ cannot be complex. Hence, eight cases in the generalized FS×W classification cannot occur; the five remaining cases actually occur, as it can be easily shown by means of examples.[37] Whenever $K=C$, by comparison between Table II and Table III we obtain that these cases are characterized by $\mathfrak{A}^1=2R_n$ and $\mathfrak{A}^1=Q_{n/4}$ in column 2, $\mathfrak{A}^1=2C_{n/2}$ in column 3, $\mathfrak{A}^1=2Q_{n/4}$ and $\mathfrak{A}^1=4R_{n/2}$ in column 4 of Table III (analogous results, for the subclass of representations to which his classification applies, are obtained by Dyson). In addition, let us observe that the results in remark 5 and in the remark quoted and restated in footnote 34 can be assembled together so that the following statement holds. Let $\mathscr{U}$ be a factorizable group, i.e., $(\mathscr{U}\cap\mathscr{U}^c)^a\neq\varnothing$, and let $\mathfrak{A}^1\neq 2C_{n/2}$. Then, the cases (a) $\mathscr{U}\cup\mathscr{U}^c$ is potentially real (i.e., $\mathfrak{A}^1=2R_n$ or $\mathfrak{A}^1=2Q_{n/4}$), (b) $\mathscr{U}\cup\mathscr{U}^c$ is pseudoreal (i.e., $\mathfrak{A}^1=Q_{n/2}$ or $\mathfrak{A}^1=4R_{n/2}$), respectively, can be characterized by

(a′) for any $T\in(\mathscr{U}\cap\mathscr{U}^c)^a$, $T^2\in R_+E$,

(b′) for any $T\in(\mathscr{U}\cap\mathscr{U}^c)^a$, $T^2\in R_-E$.

Furthermore, whenever $\mathfrak{A}^1=2C_{n/2}$, then for any $T\in(\mathscr{U}\cap\mathscr{U}^c)^a$, $T^2\in nC$ (indeed, as we have seen in remark 5, $T^2\in\mathfrak{D}\cap\mathfrak{D}'^R$, while $\mathfrak{D}\cap\mathfrak{D}'^R=C_n\cap nC=nC$ in the present case).

The above statements clearly reformulate Dyson's theorem III (see Ref. 35) in a generalized framework (in particular, since the operators of $\mathscr{U}$ are not supposed to be unitary or antiunitary, Dyson's statement "$[M(T)]=\pm 1$" is substituted by the statement $T^2\in R_\pm E$).

V. EXAMPLES

In this section we give some examples which illustrate, in some cases of the generalized FS×W classification, the properties collected in the corresponding squares of Table III. For the sake of brevity we limit ourselves here to considering matrix algebras that can be generated by factorizable groups (remark 6); thus, five cases only occur (we do not give explicitly the factorizable group which can be chosen as a set of generators for the algebra in each case, since it is easily recognizable by looking at the explicit form of the algebra). In each case we give a set $\mathscr{U}$ of linear and antilinear mappings (which coincides with the set introduced in Ref. 11, Sec. 5, and, except for minor changes and with the exception of case 4, with the set introduced by Dyson in this examples, so that it can be endowed of the same physical interpretation discussed by Dyson) by specifying the set $\mathscr{M}_{\mathscr{F}}(\mathscr{U})$ of the matrices that realize the operators of $\mathscr{U}$ with respect to some basis $\mathscr{F}=\{f_j\}_{j=1,\ldots,n}$ in the vector space X over C (see Ref. 18). Whenever $\mathscr{U}$ is potentially real, we write the operator S which maps $\mathscr{F}$ onto the canonical basis $\mathscr{C}=\{Sf_j\}_{j=1,\ldots,n}$ in X (see Ref. 38); then, we give the matrix set $\mathscr{M}_{\mathscr{C}}(\mathscr{U})$ and the set $\mathscr{M}_{\mathscr{C}\cup i\mathscr{C}}(\mathfrak{A})$ of the matrices that realize the operators of the algebra $\mathfrak{A}$ (over R) generated by $\mathscr{U}$ with respect to the basis $\mathscr{C}\cup i\mathscr{C}$ of X_R (see Ref. 39) (see Theorem 2). Whenever $\mathscr{U}$ is not potentially real, we directly write the matrix set $\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(\mathfrak{A})$ of the matrices that represent $\mathfrak{A}$ with respect to the basis $\mathscr{F}\cup i\mathscr{F}$ of X_R. Then, in each case we give the canonical forms of $\mathfrak{A}$, $\mathfrak{D}$, $\mathfrak{A}^1$, together with the new basis of X_R (if it does not coincide with the old one) to which these canonical forms refer.

In what follows we make use of all the definitions and symbols introduced in the rest of the present paper. In particular, $\mathscr{I}_p$ denotes the identity in the set of the $p\times p$ matrices with elements in R, $\mathscr{C}$, and $\mathscr{Q}$ denote regular representations of C and Q, respectively.[29]

For the sake of brevity, we also put

$$e_0=\begin{pmatrix}1&0\\0&1\end{pmatrix};\quad e_1=\begin{pmatrix}0&1\\1&0\end{pmatrix};$$
$$e_2=\begin{pmatrix}0&-1\\1&0\end{pmatrix};\quad e_3=\begin{pmatrix}1&0\\0&-1\end{pmatrix}.$$

(1) $\mathscr{U}'^1=RE$, $\mathscr{U}$ *potentially real* (*column* 2, *row* 1): X one-dimensional, basis $\mathscr{F}=\{f\}$, $\mathscr{M}_{\mathscr{F}}(\mathscr{U}^1)=R$, $\mathscr{M}_{\mathscr{F}}(\mathscr{U}^a)=\alpha R$, with $\alpha=\rho\, e^{i\varphi}\in C$, $\mathscr{C}=\{Sf\}$, with $S=e^{i\varphi/2}$, and $\mathscr{M}_{\mathscr{C}}(\mathscr{U}^1)=R=\mathscr{M}_{\mathscr{C}}(\mathscr{U}^a)$. The matrix representation of $\mathfrak{A}$ with respect to the basis $\mathscr{C}\cup i\mathscr{C}$ takes the canonical form

$$\mathscr{M}_{\mathscr{C}\cup i\mathscr{C}}(\mathfrak{A})=\left\{\begin{pmatrix}r&0\\0&s\end{pmatrix}:r,s\in R\right\}=R\oplus R.$$

Also the matrix representations of $\mathfrak{D}$ and $\mathfrak{A}^1$ with respect to the same basis take the canonical form

$$\mathscr{M}_{\mathscr{C}\cup i\mathscr{C}}(\mathfrak{D})=R_2,$$
$$\mathscr{M}_{\mathscr{C}\cup i\mathscr{C}}(\mathfrak{A}^1)=2R.$$

(2) $\mathscr{U}'^1=RE$, $\mathscr{U}$ *pseudoreal* (*column*, *row* 2): X two-dimensional, basis $\mathscr{F}=\{f_j\}_{j=1,2}$,

$$\mathscr{M}_{\mathscr{F}}(\mathscr{U}^1)=\left\{\begin{pmatrix}\alpha&\beta\\-\bar\beta&\bar\alpha\end{pmatrix}:\alpha,\beta\in C\right\}=\mathscr{M}_{\mathscr{F}}(\mathscr{U}^a);$$

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(\mathfrak{A})=\left\{\begin{pmatrix}ae_0+be_2&ce_1+de_3\\ee_1+fe_3&ge_0+he_2\end{pmatrix}:a,b,c,d,e,f,g,h\in R\right\}.$$

By choosing the basis $S\,(\mathscr{F}\cup i\mathscr{F})=\mathscr{F}\cup i\mathscr{F}'$ in X_R, with

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(S)=\begin{pmatrix}e_0&0\\0&e_3\end{pmatrix},$$

we get

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}'}(\mathfrak{A})=\left\{\begin{pmatrix}ae_0+be_2&ce_0+de_2\\ee_0+fe_2&ge_0+he_2\end{pmatrix}:a,b,c,d,e,f,g,h\in R\right\}=\mathscr{C}_2.$$

Also the matrix representations of $\mathfrak{D}$ and $\mathfrak{A}^1$ with respect to the same basis take the canonical form

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}'}(\mathfrak{D})=R_4,\quad \mathscr{M}_{\mathscr{F}\cup i\mathscr{F}'}(\mathfrak{A}^1)=\mathscr{Q}.$$

(3) $\mathscr{U}^1\approx C$, $\mathscr{U}$ *potentially real*, $\mathscr{U}^c\approx C_*\times G_2$ (*column* 3, *row* 1, *upper case*): X two-dimensional, basis $\mathscr{F}=\{f_j\}_{j=1,2}$,

$$\mathscr{M}_{\mathscr{F}}(\mathscr{U}^1)=\left\{\begin{pmatrix}\alpha&0\\0&\bar\alpha\end{pmatrix}:\alpha\in C\right\},$$
$$\mathscr{M}_{\mathscr{F}}(\mathscr{U}^a)=\left\{\begin{pmatrix}0&\alpha\\\bar\alpha&0\end{pmatrix}:\alpha\in C\right\},$$
$$\mathscr{C}=S\mathscr{F}=\{Sf_i\}_{i=1,2},$$
$$\text{with } \mathscr{M}_{\mathscr{F}}(S)=\gamma e_0+\bar\gamma e_1,\quad \gamma\neq\bar\gamma\in C.$$

By choosing $\gamma=1+i$,

$$\mathscr{M}_{\mathscr{C}}(\mathscr{U}^1)=\{ae_0+be_2:a,b\in R\}=\mathscr{M}_{\mathscr{C}}(\mathscr{U}^a).$$

The matrix representation of $\mathfrak{A}$ with respect to the basis $\mathscr{C}\cup i\mathscr{C}$ takes the canonical form

$$\mathscr{M}_{\mathscr{C}\cup i\mathscr{C}}(\mathfrak{A})=\left\{\begin{pmatrix}ae_0+be_2&0\\0&ce_0+de_2\end{pmatrix}:a,b,c,d\in R\right\}=\mathscr{C}\oplus\mathscr{C}.$$

Also the matrix representations of $\mathfrak{D}$ and $\mathfrak{A}^1$ with respect to the same basis take the canonical form

$$\mathscr{M}_{\mathscr{C}\cup i\mathscr{C}}(\mathfrak{D})=\mathscr{C}_2,\quad \mathscr{M}_{\mathscr{C}\cup i\mathscr{C}}(\mathfrak{A}^1)=2\mathscr{C}.$$

(4) $\mathscr{U}'^1\approx Q$, $\mathscr{U}$ *potentially real*, $\mathscr{U}^c\approx Q_*\times G_2$ (*column* 4, *row* 1, *upper case*): X four-dimensional, basis $\mathscr{F}=\{f_j\}_{j=1,\ldots,4}$,

$$\mathscr{M}_{\mathscr{F}}(\mathscr{U}^1)=\left\{\begin{pmatrix}\alpha&\beta&0&0\\-\bar\beta&\bar\alpha&0&0\\0&0&\bar\alpha&-\bar\beta\\0&0&\beta&\alpha\end{pmatrix}:\alpha,\beta\in C\right\},$$

$$\mathscr{M}_{\mathscr{F}}(\mathscr{U}^a)=\left\{\begin{pmatrix}0&0&\alpha&-\beta\\0&0&-\bar\beta&-\bar\alpha\\\bar\alpha&\bar\beta&0&0\\\beta&-\alpha&0&0\end{pmatrix}:\alpha,\beta\in C\right\},$$
$$=\mathscr{M}_{\mathscr{F}}(\mathscr{U}^1)\begin{pmatrix}0&e_3\\e_3&0\end{pmatrix}.$$

$$\mathscr{C}=S\mathscr{F}=\{Sf_j\}_{j=1,\ldots,4},$$
$$\text{with } \mathscr{M}_{\mathscr{F}}(S)=\begin{pmatrix}\gamma e_0&\bar\gamma e_3\\\bar\gamma e_3&\gamma e_0\end{pmatrix},\quad \gamma\neq\bar\gamma\in C.$$

By choosing $\gamma=1+i$,

$$\mathscr{M}_{\mathscr{C}}(\mathscr{U}^1)=\left\{\begin{pmatrix}ae_0-be_2&ce_0-de_2\\-ce_0-de_2&ae_0+be_2\end{pmatrix}:a,b,c,d\in R\right\}.$$
$$=\mathscr{M}_{\mathscr{F}}(\mathscr{U}^a).$$

The matrix representation of $\mathfrak{A}$ with respect to the basis $\mathscr{E}\cup i\mathscr{E}$ takes the canonical form

$$\mathscr{M}_{\mathscr{E}\cup i\mathscr{E}}(\mathfrak{A}) = \left\{ \begin{pmatrix} ae_0 - be_2 & ce_0 - de_2 & 0 & 0 \\ -ce_0 - de_2 & ae_0 + be_2 & 0 & 0 \\ 0 & 0 & ee_0 - fe_2 & ge_0 - he_2 \\ 0 & 0 & -ge_0 - he_2 & ee_0 + fe_2 \end{pmatrix} : \begin{matrix} a,b,c,d, \\ e,f,g,h \in R \end{matrix} \right\} = \mathscr{D} \oplus \mathscr{D}.$$

Also the matrix representations of $\mathfrak{D}$ and $\mathfrak{A}^{1}$ with respect to the same basis take the canonical form

$$\mathscr{M}_{\mathscr{E}\cup i\mathscr{E}}(\mathfrak{D}) = \mathscr{D}_2, \quad \mathscr{M}_{\mathscr{E}\cup i\mathscr{E}}(\mathfrak{A}^{1}) = 2\mathscr{D}.$$

(5) $\mathscr{U}'^{1} \approx Q$, *$\mathscr{U}$ potentially real*, $\mathscr{U}^{c} \not\approx Q_{*} \times G_2$ *(column* 4, *row* 1, *lower case)*: X two-dimensional, basis $\mathscr{F} = \{f_j\}_{j=1,2}$, $\mathscr{M}_{\mathscr{F}}(\mathscr{U}^{1}) = \{re_0 : r\in R\}$, $\mathscr{M}_{\mathscr{F}}(\mathscr{U}^{a}) = \{re_2 : r\in R\}$, $\mathscr{E} = \mathscr{F}$.

The matrix representation of $\mathfrak{A}$ with respect to the basis $\mathscr{F}\cup i\mathscr{F}$ takes the form

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(\mathfrak{A}) = \left\{ \begin{pmatrix} re_0 - se_2 & 0 \\ 0 & re_0 + se_2 \end{pmatrix} : r,s\in R \right\}.$$

By choosing the basis $S(\mathscr{F}\cup i\mathscr{F}) = \mathscr{F}\cup i\mathscr{F}'$, with

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(S) = \begin{pmatrix} e_0 & 0 \\ 0 & e_3 \end{pmatrix},$$

we get

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}'}(\mathfrak{A}) = \left\{ \begin{pmatrix} re_0 + se_2 & 0 \\ 0 & re_0 + se_2 \end{pmatrix} : r,s\in R \right\} = 2\mathscr{C}.$$

Also the matrix representations of $\mathfrak{D}$ and $\mathfrak{A}^{1}$ with respect to the same basis take the canonical form

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}'}(\mathfrak{D}) = \mathscr{D}, \quad \mathscr{M}_{\mathscr{F}\cup i\mathscr{F}'}(\mathfrak{A}^{1}) = 4R.$$

ACKNOWLEDGMENTS

We wish to thank Professor Renato Ascoli and Professor Pietro Rotelli for useful discussions and suggestions.

This research was sponsored by Consiglio Nazionale delle Ricerche (CNR) and Istituto Nazionale di Fisica Nucleare (INFN) (Italy).

[1]G. Frobenius and I. Schur, "Über die reellen Darstellungen der endlichen Gruppen," Berl. Berichte, 186 (1906).

[2]E. P. Wigner, *Group Theory* (Academic, New York, 1959), Sec. 24.

[3]L. C. Biedenharn and J. D. Louck, "The Racah-Wigner Algebra in Quantum Theory," in *Encyclopedia of Mathematics and its Applications*, edited by G. C. Rota (Addison-Wesley, Reading, MA, 1981), Chap. 5, Secs. 6 and 10.

[4]R. Ascoli, C. Garola, L. Solombrino, and G. Teppati, "Real versus Complex Representations and Linear-Antilinear Commutant," in *Physical Reality and Mathematical Description*, edited by C. P. Enz and J. Mehra (Reidel, Dordrecht, 1974), Sec. 1. pp. 239–251.

[5]See Ref. 2, Sec. 26.

[6]F. J. Dyson., "The Threefold Way. Algebraic Structure of Symmetry Groups and Ensembles in Quantum Mechanics," J. Math. Phys. **3**, 1199 (1962), Secs. 3–5; see also Ref. 3, Chap. 5, Sec. 10.

[7]H. Weyl, *The Classical Groups, Their Invariants and Representations* (Princeton U.P., Princeton, NJ, 1939), Chap. 3, Sec. 5, Theorem (3.5.B); see also Ref. 6, Sec. 2.

[8]See Ref. 6, Sec. 1.

[9]See Ref. 4, Sec. 3, Table.

[10]C. Garola and L. Solombrino, "Commutation in Vector Spaces over Division Rings with a Conjugation," Linear Algebra Appl. **36**, 41 (1981), Sec. 1, Table 1.

[11]C. Garola and L. Solombrino, "Irreducible Linear-Antilinear Representations and Internal Symmetries," J. Math. Phys. **22**, 1350 (1981), Sec. 3.

[12]See Ref. 6, Sec. 4.

[13]N. Bourbaki, *Éléments de Mathématique, Algèbre, Chap. 2: Algèbre lineaire* (Hermann, Paris, 1962), Sec. 8.

[14]We note that, whenever X has finite dimension n, X_A has dimension $2n$, see R. Ascoli, C. Garola, L. Solombrino, and G. Teppati, "Vector Spaces over Fields with a Conjugation and Linear-Antilinear Commutants," Rend. Mat. Ser. VI **10**(1), 129 (1977), Sec. 2, Theorem 1.

[15]See Ref. 4, Sec. 3, Lemma 1.

[16]See Ref. 4, Sec. 3, Theorem 3.

[17]See Ref. 11, Secs. 3 and 4, Propositions 1 and 2.

[18]See Ref. 4, Sec. 4.

[19]N. Bourbaki, *Éléments de Mathématique, Algèbre* (Hermann, Paris, 1973), Chap. 8; see in particular Sec. 4, No. 4, Propositions 4 and 5.

[20]N. Bourbaki, *Éléments de Mathématique, Algèbre* (Hermann, Paris, 1971), Chap. 3, Sec. 1.

[21]See Ref. 7, Chap. 3, Sec. 4.

[22]Whenever K is any field with a conjugation and S any antilinear mapping the first statement in Lemma 1 can be deduced from Proposition 1 in Ref. 11, Sec. 3.

[23]See Ref. 14, Sec. 5, Theorem 4.

[24]See Ref. 11, Sec. 3, proof of Proposition 1.

[25]We observe that the first statement in the theorem holds invariant whenever C is generalized to be any algebraically closed field K with a conjugation; in the proof, no change is needed besides the substitution $C \rightarrow K$.

[26]Observe that the decomposition $X_R = R\langle\mathscr{E}\rangle + iR\langle\mathscr{E}\rangle$ does not coincide with the one induced in X_R by the decomposition $X = Y_1 \oplus Y_2$ of X introduced in the first part of the present proof.

[27]An immediate proof of the first part of the second statement in the theorem can also be given as follows. Because of a statement proved by the authors in a previous paper (see Ref. 10, Sec. 3, Proposition 6) $\mathfrak{A}$ is potentially real iff $\mathfrak{A}'^{R}$ has divisors of zero. By inspection of the possible forms of $\mathfrak{A}'^{R}$ listed in Theorem 1, we see that this case occurs iff $\mathfrak{A}$ is reducible in X_R.

[28]See Ref. 10, Sec. 2, first remark in the proof of Proposition 1.

[29]See Ref. 7, Chap. 3, Sec. 1 and Ref. 6, Sec. 2, Remark 4.

[30]See Ref. 10, Sec. 3, Proposition 5.

[31]See Ref. 10, Sec. 2. Proposition 1.

[32]See Ref. 10, Sec. 2, Proposition 4.

[33]It follows in particular that Dyson's cases CC1 and CC2 correspond to the cases in row 3, column 3 and to the upper subcase in row 1, column 3, of Table III, respectively; the distinction between the two cases is not irrelevant, contrary to our last sentence in Ref. 11, footnote 15.

[34]These results complete, in some sense, the result in the remark 5 in Ref. 11, Sec. 5, which refer to columns 1, 2, 3 of Table III only. However, this remark has been stated improperly (its proof being correct) since no reference to the case $\mathscr{U}$ complex should have been made. The following constitutes a proper (and simplified) statement. Whenever $\mathscr{U}^{1}$ is isomorphic to R or to C (columns 2 and 1,3, respectively, in Table II) and $\mathscr{U}^{c} \not\approx C_{*} \times G_2$, then the cases (a) $\mathscr{U}$ potentially real and (b) $\mathscr{U}$ pseudoreal, can be characterized by (a') $\mathscr{U}^{ca} \neq \varnothing$ and for any $T\in\mathscr{U}^{ca}$, $T^2 \in R_{+}E$; (b') $\mathscr{U}^{ca} \neq \varnothing$ and for any $T\in\mathscr{U}^{ca}$, $T^2\in R_{-}E$, respectively.

[35]See Ref. 6, last part of Sec. 4.

[36]The equivalence between the two definitions is an immediate consequence of the equation $\mathscr{U}^{a} = S\mathscr{U}^{1}$, which holds for any $S\in\mathscr{U}^{a}$ (see footnote 28).

[37]See next section and Ref. 6, Sec. 5.

[38]It is worthy of note that such a basis $\mathscr{E}$ can easily be constructed whenever an involutory $J\in\mathscr{U}^{ca}$ is known. Indeed, let us consider the mapping $S = \gamma E + \bar{\gamma} J J_{\mathscr{F}}$. Then, some $\gamma\in C$ exists such that S is invertible. For, should this not be the case, for every $\gamma\in C$ some nonzero $x_\gamma \in X$ would exist such that $Sx_\gamma = 0$, that is, $JJ_{\mathscr{F}} x_\gamma = -\gamma\bar{\gamma}^{-1} x_\gamma$, which is impossible because of basic statements about the eigenvalues of any linear mapping of X. Whenever S is invertible, we get $JS = \bar{\gamma} J + \gamma J_{\mathscr{F}} = SJ_{\mathscr{F}}$, hence $J = \bar{S}\bar{S}_{\mathscr{F}}^{-1} J_{\mathscr{F}}$. Now, let us put $\mathscr{E} = \{Sf_j\}_{j=1,\dots,n}$. Then, easily, $J_{\mathscr{E}} = \bar{S}\bar{S}_{\mathscr{F}}^{-1} J_{\mathscr{F}}$, i.e., $J = J_{\mathscr{E}}$. Since $J\in\mathscr{U}^{ca}$, it follows $\bar{M}_{\mathscr{E}} = M$ for every $M\in\mathscr{U}$, which proves that $\mathscr{E}$ has the desired property.

[39]We notice that, whenever $\mathscr{M} = \mathscr{M}_{\mathscr{F}}(M)$ is the matrix realization with respect to any basis $\mathscr{F}$ in X of some $M \in \mathscr{U}$, the matrix realization of M with respect to the basis $\mathscr{F} \cup i\mathscr{F}$ in X_R is given by

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(M) = \begin{pmatrix} \mathrm{Re}\,\mathscr{M} & -\mathrm{Im}\,\mathscr{M} \\ \mathrm{Im}\,\mathscr{M} & \mathrm{Re}\,\mathscr{M} \end{pmatrix},$$

if M is a linear, and by

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(M) = \begin{pmatrix} \mathrm{Re}\,\mathscr{M} & \mathrm{Im}\,\mathscr{M} \\ \mathrm{Im}\,\mathscr{M} & -\mathrm{Re}\,\mathscr{M} \end{pmatrix},$$

if M is antilinear. In particular, we obtain

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(iE) = \begin{pmatrix} 0 & -\mathscr{I}_n \\ \mathscr{I}_n & 0 \end{pmatrix}$$

and

$$\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(J_{\mathscr{F}}) = \begin{pmatrix} \mathscr{I}_n & 0 \\ 0 & -\mathscr{I}_n \end{pmatrix}$$

[see also the proof of (iv)$\Rightarrow$(i) in Theorem 2]. We observe explicitly that, whenever M is antilinear, it is realized in X by the pair $[\mathscr{M}_{\mathscr{F}}(M), j]$ (see footnote 18), while the matrix $\mathscr{M}_{\mathscr{F}\cup i\mathscr{F}}(M)$ alone realizes M in X_R.

Explicit form of the Haar measure of U(n) and differential operators

Takayoshi Maekawa
Department of Physics, Kumamoto University, Kumamoto, Japan

(Received 6 November 1984; accepted for publication 29 March 1985)

The Haar measure of the group U(n) is explicitly introduced using the Euler-like parameters, and the differential operators of the first- and second-parameter groups are given. The D-matrix elements are defined through the Gel'fand and Tsetlin basis and the orthogonality and the completeness relations for the d-matrix elements are given explicitly.

I. INTRODUCTION

The Euler-like parameters and the Haar measure are useful to analyze the representation matrix elements of SO(n) and SO(n,1).[1] The Euler-like parameters of U(n) are also known[2] and the representation matrix elements are studied by introducing the matrix elements of U(n) as the parameters.[3] The latter procedure seems useful when we discuss the representation matrix elements by using the Young tableaux. However, we know that the representation matrix elements of SO(n) and SO(n,1) are nicely treated with the Euler-like parameters.[1]

In this paper, we introduce the Haar measure in terms of the Euler-like parameters for U(n) and the differential operators of the first- and the second-parameter groups. Though the orthogonality and the completeness relations for the D-matrix elements of the compact groups[4] are well known, we give those to the d-matrix elements of SU(n).

In Sec. II, the Haar measure invariant under the right and the left shifts is given in terms of the Euler-like parameters. In Sec. III, the D-matrix elements are defined and then the orthogonality and the completeness relations are given to the d-matrix elements as well as the D-matrix elements. In Sec. IV, the differential operators of the first- and the second-parameter groups[5] are constructed in terms of the parameters and their differential operators.

II. HAAR MEASURE OF SU(n)

In this section, we first introduce the Euler-like parameters[2] in order to explain the notations used in the following and then give the explicit form to the Haar measure.

As is well known, any elements U(n), whose members make $|z_1|^2 + \cdots + |z_n|^2$ invariant, can be uniquely written as a matrix of $U(1)$ and SU(n), i.e.,

$$U = e^{i\phi} U_0,$$

with $\exp(i\phi) \in U(1)$ and $U_0 \in \mathrm{SU}(n)$. Therefore, it is sufficient for us to introduce $n^2 - 1$ parameters of SU(n). The generators e_{ij} $(= e_{ji}^\dagger)$ of U(n), which have matrix elements $(e_{ij})_{kl} = \delta_{ik}\delta_{jl}$, satisfy the commutation relations

$$[e_{ij}, e_{kl}] = \delta_{jk} e_{il} - \delta_{il} e_{kj}. \tag{2.1}$$

From these generators, we construct the generators of SU(n) as follows:

$$\begin{aligned} E_{ij} &= \tfrac{1}{2}(e_{ij} + e_{ji}) = E_{ji}, \quad \text{for } i \neq j, \\ E_{(ij)} &= (i/2)(e_{ji} - e_{ij}) = -E_{(ji)}, \quad \text{for } i<j, \\ E_{jj} &= \frac{1}{\sqrt{2j(j+1)}} \left(\sum_k^j e_{kk} - j e_{j+1\, j+1} \right), \\ &\quad \text{for } j = 1,2,\ldots,n-1. \end{aligned} \tag{2.2}$$

Of course, we obtain the generators of U(n) if we add the unit generator $E = \Sigma e_{jj}$ to the above $n^2 - 1$ generators of SU(n). The $n^2 - 1$ generators E_{ij} may be arranged from 1 to $n^2 - 1$ as follows:

$$\begin{aligned} &\left.\begin{aligned} E_{(j-1)^2+2i-2} &= E_{ij} \\ E_{(j-1)^2+2i-1} &= E_{(ij)} \end{aligned}\right\} (i = 1,2,\ldots,j-1;\ j = 2,3,\ldots,n), \\ &E_{j^2-1} = E_{j-1\, j-1} \quad (j = 2,3,\ldots,n). \end{aligned} \tag{2.3}$$

We may write the commutation relations for the E_j's as follows:

$$[E_j, E_k] = i f_{jkl} E_l, \tag{2.4}$$

where the f_{ijk}'s are the structure constants which are totally antisymmetric because of the normalization of $\mathrm{tr}(E_j E_k) = \delta_{jk}/2$.

We can parametrize the elements of SU(n) in a similar way as in Ref. 2 as follows:

$$\begin{aligned} g^{(n)} &= g^{(n-1)} S_n^{(n)}, \\ S_n^{(n)} &= \left(\prod_{j=n}^{2} e^{i\phi_{n\,n-j+1}\Gamma_{j-1}} e^{i\theta_{n\,n-j+1}E_{j^2-2}} \right) \\ &\quad \times e^{i\Psi_{n\,n-1}\Gamma_1}, \\ g^{(2)} &= e^{i\phi_{21}\Gamma_1} e^{i\theta_{21}E_2} e^{i\Psi_{21}\Gamma_1}, \end{aligned} \tag{2.5}$$

where

$$\begin{aligned} \Gamma_j &= (e_{jj} - e_{j+1\,j+1})/2 \\ &= (-\sqrt{j-1}E_{j^2-1} + \sqrt{j+1}E_{(j+1)^2-1})/\sqrt{2j}. \end{aligned}$$

The ranges of the parameters are given as follows:

$$0 \leqslant \theta_{jk} \leqslant \pi, \quad 0 \leqslant \phi_{jk} < 4\pi, \quad 0 \leqslant \psi_{jj-1} < 2\pi. \tag{2.6}$$

The meaning of (2.5) is clarified by writing them explicitly, for instance,

$$\begin{aligned} (R)_{j-1\,j-1} &= \cos(\theta/2) e^{(i/2)\phi}, \\ (R)_{jj} &= \cos(\theta/2) e^{-(i/2)\phi}, \\ (R)_{j-1\,j} &= \sin(\theta/2) e^{(i/2)\phi}, \\ (R)_{jj-1} &= -\sin(\theta/2) e^{-(i/2)\phi}, \\ (R)_{lk} &= \delta_{lk}, \quad \text{for } l \neq j,\ j-1, \end{aligned} \tag{2.7}$$

where

$$R \equiv e^{i\phi\Gamma_{j-1}} e^{i\theta E_{j^2-2}}.$$

It follows that $|z_{j-1}|^2 + |z_j|^2$ is invariant under the transformation (2.7).

The Haar measure of SU(n) with respect to the above parameters is given as follows:

 0022-2488/85/081902-08$02.50

$$dV_n = \left[\prod_{i=1}^{n-1} \cos\frac{\theta_{ni}}{2}\left(\sin\frac{\theta_{ni}}{2}\right)^{2(n-i)-1} d\theta_{ni}\, d\phi_{ni}\right] \times d\psi_{nn-1}\, dV_{n-1}, \tag{2.8}$$

$$dV_2 = \cos(\theta_{21}/2)\sin(\theta_{21}/2)d\theta_{21}\, d\phi_{21}\, d\psi_{21}.$$

The volume of SU(n) is given by $V_n = (4\pi)^n\ V_{n-1}/[2(n-1)!]$ and $V_2 = (4\pi)^2/2$. Of course, the measure of U(n) is multiplied by $d\phi$ and the volume by 2π.

The expression (2.8) can be proved as follows. We consider the matrix $\Omega = dg^{(n)}\, g^{(n)\dagger}$, whose elements are one form of a linear combination of the differential of the parameters and invariant under the right shift.[6] Then Ω satisfies the relation

$$\Omega + \Omega^\dagger = 0. \tag{2.9}$$

It follows from (2.9) that the real parts of Ω are antisymmetric and the imaginary parts are symmetric. The number of the parameters of SU(n) is n^2-1 and then the matrix elements, except for one, which we take to be the (n,n)-th element, form the basis of the one-form.[6] The right-invariant volume element is given by the exterior product of (n^2-1) one-forms. It is, therefore, sufficient for us to calculate the determinant (Jacobian) formed from the coefficients of the differential of the angle parameters in order to obtain a measure. We make the $(n^2-1)\times(n^2-1)$ determinant in such a way that each column is arranged in its order of the angles in (2.5), the elements in each column are arranged in lexicographical order for the indices of SU($n-1$), and the $(1\,n),\ldots,(n-1n)$, $(n\,1),\ldots,(n\,n-1)$-th elements follow after.

By noting the relation $(\partial g^{(n-1)}/\partial\phi g^{(n-1)\dagger})_{jk} = 0$ for $j=n$ or $k=n$ with any possible angles ϕ and $(g^{(n-1)}\,\partial S_n^{(n)}/\partial\phi_{n1}\, S_n^{(n)\dagger} g^{(n-1)\dagger})_{jk} = 0$ for $j=n$, $k\neq n$ or $j\neq n$, $k=n$, we obtain a product of the $(n-1)^2\times(n-1)^2$ determinant whose elements consist of $(\partial g^{(n-1)}/\partial\phi\, g^{(n-1)\dagger})_{jk}$ and the $(2n-2)\times(2n-2)$ determinant whose elements consist of $(g^{(n-1)2}\, S_n^{(n)}/\partial\phi' S_n^{(n)\dagger} g^{(n-1)\dagger})_{jk}$ $(j<n, k=n$ and $j=n, k<n)$ with $\phi' = \theta_{n1},\ldots,\theta_{nn-1}, \phi_{n2},\ldots,\phi_{nn-1}, \psi_{nn-1}$. The latter determinant is given explicitly by

$$\Delta = \begin{vmatrix} (g^{(n-1)}A^1 g^{(n-1)\dagger})_{1\,n}, & \ldots, & (g^{(n-1)}A^{2n-2} g^{(n-1)\dagger})_{1\,n} \\ \vdots & & \vdots \\ (g^{(n-1)}A^1 g^{(n-1)\dagger})_{n-1\,n}, & \ldots, & (g^{(n-1)}A^{2n-2} g^{(n-1)\dagger})_{n-1\,n} \\ (g^{(n-1)}A^1 g^{(n-1)\dagger})_{n\,1}, & \ldots, & (g^{(n-1)}A^{2n-2} g^{(n-1)\dagger})_{n\,1} \\ \vdots & & \vdots \\ (g^{(n-1)}A^1 g^{(n-1)\dagger})_{n\,n-1}, & \ldots, & (g^{(n-1)}A^{2n-2} g^{(n-1)\dagger})_{n\,n-1} \end{vmatrix}, \tag{2.10}$$

where $A^j = \partial S_n^{(n)}/\partial\phi_j\, S_n^{(n)\dagger}$ with possible ϕ_j's. Equation (2.10) becomes, due to $\det(g^{(n-1)}) = 1$,

$$\Delta = \begin{vmatrix} (A^1)_{1\,n}, & \ldots, & (A^{2n-2})_{1\,n} \\ \vdots & & \vdots \\ (A^1)_{n-1\,n}, & \ldots, & (A^{2n-2})_{n-1\,n} \\ (A^1)_{n\,1}, & \ldots, & (A^{2n-2})_{n\,1} \\ \vdots & & \vdots \\ (A^1)_{n\,n-1}, & \ldots, & (A^{2n-2})_{n\,n-1} \end{vmatrix}. \tag{2.11}$$

Making use of the relations

$$(A^1)_{j\,n} = \left(\frac{\partial S_n^{(n)}}{\partial\theta_{n1}} s_n^{(n)\dagger}\right)_{jn} = \frac{1}{2}e^{i\phi_{n1}}\delta_{j\,n-1},$$

$$(A^1)_{n\,j} = -\frac{1}{2}e^{-i\phi_{n1}}\delta_{j\,n-1},$$

$$(A^k)_{j\,n} = -\sin(\theta_{n1}/2)e^{(i/2)\phi_{n1}} \times \left[e^{i\phi_{n1}\Gamma_{n-1}} e^{i\theta_{n1}E_{n^2-2}} \frac{\partial S_{n-1}^{(n)}}{\partial\phi_k} S_{n-1}^{(n)\dagger}\right]_{j\,n-1},$$

$$(A^k)_{nj} = -\sin\frac{\theta_{n1}}{2}e^{-(i/2)\phi_{n1}} \times \left[\frac{\partial S_{n-1}^{(n)}}{\partial\phi_k} S_{n-1}^{(n)\dagger} e^{-i\theta_{n1}E_{n^2-2}} e^{-i\phi_{n1}\Gamma_{n-1}}\right]_{n-1\,j},$$

$$\text{for } k\neq 1 \text{ and } \phi_k \neq \theta_{n1}, \tag{2.12}$$

$$S_n^{(n)} = e^{i\phi_{n1}\Gamma_{n-1}} e^{i\theta_{n1}E_{n^2-2}} S_{n-1}^{(n)},$$

we rewrite (2.11) as follows:

$$\Delta = \frac{1}{2}(-\sin(\theta_{n1}/2))^{2n-3} \begin{vmatrix} 0 \quad (VB^2)_{1\,n-1}, & \ldots, & (VB^{2n-2})_{1\,n-1} \\ \vdots & & \vdots \\ 0 \quad (VB^2)_{n-2\,n-1}, & \ldots, & (VB^{2n-2})_{n-2\,n-1} \\ e^{(i/2)\phi_{n1}}(VB^2)_{n-1\,n-1}, & \ldots, & (VB^{2n-2})_{n-1\,n-1} \\ 0 \quad (B^2V^\dagger)_{n-1\,1}, & \ldots, & (B^{2n-2}V^\dagger)_{n-1\,1} \\ \vdots & & \vdots \\ 0 \quad (B^2V^\dagger)_{n-1\,n-2}, & \ldots, & (B^{2n-2}V^\dagger)_{n-1\,n-2} \\ -e^{-(i/2)\phi_{n1}}(B^2V^\dagger)_{n-1\,n-1}, & \ldots, & (B^{2n-2}V^\dagger)_{n-1\,n-1} \end{vmatrix}, \tag{2.13}$$

where

$$V=\exp(i\phi_{n1}\Gamma_{n-1})\exp(i\theta_{n1}E_{n^2-2}),$$

$$B^k=\frac{\partial S_{n-1}^{(n)}}{\partial\phi_k}S_{n-1}^{(n)\dagger}.$$

The following relations hold:

$$(VB^k)_{jn-1}=\begin{cases}(B^k)_{jn-1} & \text{for } j<n-2,\\ \cos\frac{\theta_{n1}}{2}\quad e^{(i/2)\phi_{n1}}(B^k)_{n-1n-1} & \\ & \text{for } j=n-1,\end{cases}$$

$$(B^kV^\dagger)_{n-1j}=\begin{cases}(B^k)_{n-1j} & \text{for } j<n-2,\\ \cos\frac{\theta_{n1}}{2}\quad e^{-(i/2)\phi_{n1}}(B^k)_{n-1n-1} & \\ & \text{for } j=n-1.\end{cases} \tag{2.14}$$

We get from (2.13), by taking into account (2.14),

$$\Delta=(-1)^{n+1}\cos(\theta_{n1}/2)(\sin(\theta_{n1}/2))^{2n-3}$$

$$\times\begin{vmatrix}(B^2)_{1n-1}, & \dots, & (B^{2n-2})_{1n-1}\\ \vdots & & \vdots\\ (B^2)_{n-2n-1}, & \dots, & (B^{2n-2})_{n-2n-1}\\ (B^2)_{n-11}, & \dots, & (B^{2n-2})_{n-11}\\ \vdots & & \vdots\\ (B^2)_{n-1n-1}, & \dots, & (B^{2n-2})_{n-1n-1}\end{vmatrix}. \tag{2.15}$$

Finally, we use the relations

$$(B^2)_{jn-1}=\left(\frac{\partial S_{n-1}^{(n)}}{\partial\phi_{n2}}S_{n-1}^{(n)\dagger}\right)_{jn-1}=-\frac{i}{2}\delta_{jn-1},$$

$$(B^2)_{n-1j}=-(i/2)\delta_{jn-1},$$

to rewrite (2.15) in the form

$$\Delta=\frac{i}{2}(-1)^{n+1}\cos\frac{\theta_{n1}}{2}\left(\sin\frac{\theta_{n1}}{2}\right)^{2n-3}$$

$$\times\begin{vmatrix}(B^3)_{1n-1}, & \dots, & (B^{2n-2})_{1n-1}\\ \vdots & & \vdots\\ (B^3)_{n-2n-1}, & \dots, & (B^{2n-2})_{n-2n-1}\\ (B^3)_{n-11}, & \dots, & (B^{2n-2})_{n-11}\\ \vdots & & \vdots\\ (B^3)_{n-1n-2}, & \dots, & (B^{2n-2})_{n-1n-2}\end{vmatrix}. \tag{2.16}$$

It follows from (2.16) that the determinant on the right side corresponds to $n\to n-1$ in (2.11). We, therefore, obtain an expression for Δ except for a numerical factor

$$\Delta=\prod_{j=1}^{n-1}\cos\frac{\theta_{nj}}{2}\left(\sin\frac{\theta_{nj}}{2}\right)^{2(n-j)-1}. \tag{2.17}$$

By analogy with the above procedure, it is easily seen that the other $(n-1)^2\times(n-1)^2$ determinant becomes the weight function of the measure of SU $(n-1)$. A direct proof is given in the Appendix.

Thus our measure is given by (2.8). Though it is obvious that (2.8) is invariant under the right and the left shifts due to the theorem on the external form of the compact connected Lie group,[6] a direct proof is given in the Appendix.

III. *D*-MATRIX ELEMENTS

In this section, we discuss the *D*-matrix elements which will be used in the following section.

The representation operators D_j $(j=1,\dots,n^2-1)$ corresponding to E_j satisfy the same commutation relations as (2.4), i.e.,

$$[D_j,\ D_k]=i\,f_{jkl}D_l. \tag{3.1}$$

The representation *D*-matrix of SU (n) corresponding to the parametrization (2.5) is given by

$$D^{(n)}(g^{(n)})=D^{(n-1)}(g^{(n-1)})H(S_n^{(n)}),$$

$$H(S_n^{(n)})=\left(\prod_{j=n}^{2}e^{i\phi_{nn-j+1}T_{j-1}}e^{i\theta_{nn-j+1}D_{j^2-2}}\right)e^{i\psi_{nn-1}\Gamma_1}, \tag{3.2}$$

where

$$T_j=(D(e_{jj})-D(e_{j+1j+1}))/2$$

$$=(-\sqrt{j-1}D_{j^2-1}+\sqrt{j+1}D_{(j+1)^2-1})(1/\sqrt{2j}).$$

The following relations are easily recognized in the same way as in SO(n)[1]:

$$D_jD^{(n)}(g^{(n)})=\bar{J}_jD^{(n)}(g^{(n)}),$$

$$D^{(n)}(g^{(n)})D_j=J_jD^{(n)}(g^{(n)}), \tag{3.3}$$

where the differential operators $\bar{J}_j$ and J_j of the first- and the second-parameter groups are given by

$$\bar{J}_j=\sum(\bar{T}^{-1})_{jk}p_k$$

$$J_j=\sum(T^{-1})_{jk}\,p_k,$$

$$p_k=-\frac{i\partial}{\partial\theta_k},$$

$$(\bar{T})_{kj}=-2i\,\mathrm{tr}\left(E_j\frac{\partial g^{(n)}}{\partial\theta_k}g^{(n)\dagger}\right),$$

$$(T)_{kj}=-2i\,\mathrm{tr}\left(E_jg^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\right),$$

$$\theta_k\equiv(\phi_{jk},\ \theta_{jk},\ \psi_{jj-1}). \tag{3.4}$$

It is noted that the differential operator J_j's satisfy the same commutation relation as (3.1) but the $\bar{J}_j$'s satisfy the relation with the minus sign on the right side in (3.1).

The matrix elements of the unitary irreducible representations (UIR) of U(n) are given by the Gel'fand and Tsetlin basis[7]

$$|m_{jk}\rangle\equiv|\lambda_n,\dots,\lambda_1\rangle, \tag{3.5}$$

where λ_j stands for $(m_{1j},\dots,m_{jj})$. The non–negative integers m_{jk} are subject to the conditions $m_{jk+1}\geqslant m_{jk}\geqslant m_{j+1k+1}$. The dimensionality $N(\lambda_n)$ of the UIR of U (n) is given as follows[8]:

$$N(\lambda_n)=D(l_1,\dots,l_n)/D(n-1,\dots,0),$$

$$D(l_1,\dots,l_n)=\prod_{j<k}(l_j-l_k),\quad l_j\equiv m_{jn}+n-j. \tag{3.6}$$

The action of D_j on the basis (3.5) is given by[7,9]

$$(D_{k^2-3}+iD_{k^2-2})|m_{jk}\rangle=\sum_{j=1}^{k-1}A^j_{k-1k}|m_{jk-1}+1\rangle,$$
$$(D_{k^2-3}-iD_{k^2-2})|m_{jk}\rangle=\sum_{j=1}^{k-1}B^j_{kk-1}|m_{jk-1}-1\rangle,$$
$$D(e_{kk})|m_{jk}\rangle=\Big(\sum_j^k m_{jk}-\sum_j^{k-1}m_{jk-1}\Big)|m_{jk}\rangle, \quad (3.7)$$

where the representation operator $D(e_{jj})$ corresponds to E_{jj} and the explicit forms of the matrix elements A and B are omitted here.[7,9]

The representation D-matrix elements of SU(n) are defined as follows:

$$D^{(\lambda_n)}_{\{\lambda'_{n-1}\}\{\lambda_{n-1}\}}(g^{(n)}) = \langle\lambda_n\{\lambda'_n\}|D^{(n)}(g^{(n)})|\lambda_n\{\lambda_{n-1}\}\rangle, \quad (3.8)$$

where $\{\lambda_{n-1}\}$ means $(\lambda_{n-1},...,\lambda_1)$. The d-matrix elements are defined by

$$d^{(\lambda_n)}_{\lambda'_{n-1}(\lambda_{n-2})\lambda_{n-1}}(\theta) = \langle\lambda_n\lambda'_{n-1}\{\lambda_{n-2}\}|e^{i\theta D_{n^2-2}}|\lambda_n\{\lambda_{n-1}\}\rangle. \quad (3.9)$$

As is seen from the expression (3.2) and the definition (3.8), the D-matrix elements are determined by the d-matrix element (3.9) provided that those of SU($n-1$) are known.

The orthogonality and the completeness relations of the D-matrix elements are well known,[4] i.e.,

$$\int_{\mathrm{SU}(n)} dV_n\ \overline{D^{(\lambda_n)}_{\{\lambda_{n-1}\}\{\lambda'''_{n-1}\}}(g^{(n)})}\times D^{(\lambda'_n)}_{\{\lambda'_{n-1}\}\{\lambda''_{n-1}\}}(g^{(n)}) = \delta_{\{\lambda_n\}\{\lambda'_n\}}\delta_{\{\lambda'''_{n-1}\}\{\lambda''_{n-1}\}}[V_n/N(\lambda_n)], \quad (3.10)$$

$$\sum_{\lambda_n}\frac{N(\lambda_n)}{V_n}\sum_{\{\lambda'_{n-1}\}}\sum_{\{\lambda_{n-1}\}}\overline{D^{(\lambda_n)}_{\{\lambda'_{n-1}\}\{\lambda_{n-1}\}}(g^{(n)})}\times D^{(\lambda_n)}_{\{\lambda'_{n-1}\}\{\lambda_{n-1}\}}(g^{(n)\prime}) = \delta_{(n)}(\{\delta_n\},\{\phi'_n\})\delta_{(n)}(\{\theta_n\},\{\theta'_n\})\times\delta_{(n)}(\{\psi_n\},\{\psi'_n\}),$$

where $\delta_{\{\lambda_n\}\{\lambda'_n\}}$ stands for a product of Kronecker δ's in each of the indices, and the expressions for the $\delta_{(n)}$'s are given by

$$\delta_{(n)}(\{\phi_n\},\{\phi'_n\}) = \prod_{j=1}^{n-1}\delta(\phi_{nj}-\phi'_{nj})\delta_{(n-1)}(\{\phi_{n-1}\},\{\phi'_{n-1}\}),$$
$$\delta_{(2)}(\{\phi_2\},\{\phi'_2\})=\delta(\phi_{21}-\phi'_{21}),$$
$$\delta_{(n)}(\{\theta_n\},\{\theta'_n\}) = \Big[\prod_{j=1}^{n-1}\Big\{\cos\frac{\theta_{nj}}{2}\Big(\sin\frac{\theta_{nj}}{2}\Big)^{2(n-j)-1}\Big\}^{-1}\times\delta(\theta_{nj}-\theta'_{nj})\Big]\delta_{(n-1)}(\{\theta_{n-1}\},\{\theta'_{n-1}\}),$$
$$\delta_{(2)}(\{\theta_2\},\{\theta'_2\}) = \Big(\cos\frac{\theta_{21}}{2}\sin\frac{\theta_{21}}{2}\Big)^{-1}\delta(\theta_{21}-\theta'_{21}),$$
$$\delta_{(n)}(\{\psi_n\},\{\psi'_n\}) = \prod_{j=2}^{n}\delta(\psi_{jj-1}-\psi'_{jj-1}).$$

From (3.10), the following orthogonality and completeness relations are easily obtained:

$$\sum_{\lambda_{n-2}}N(\lambda_{n-2})\int_0^\pi d\theta\cos\frac{\theta}{2}\Big(\sin\frac{\theta}{2}\Big)^{2n-3}\times\overline{d^{(\lambda_n)}_{\lambda_{n-1}(\lambda_{n-2})\lambda'_{n-1}}(\theta)}\,d^{(\lambda'_n)}_{\lambda_{n-1}(\lambda_{n-2})\lambda'_{n-1}}(\theta) = \delta_{\lambda_n\lambda'_n}\frac{1}{n-1}\frac{N(\lambda_{n-1})N(\lambda'_{n-1})}{N(\lambda_n)}, \quad (3.11)$$

$$\sum_{\lambda_n}N(\lambda_n)\,\overline{d^{(\lambda_n)}_{\lambda_{n-1}(\lambda_{n-2})\lambda'_{n-1}}(\theta)}\,d^{(\lambda_n)}_{\lambda_{n-1}(\lambda'_{n-2})\lambda'_{n-1}}(\theta') = [\cos(\theta/2)(\sin(\theta/2))^{2n-3}]^{-1}\delta(\theta-\theta')\times\delta_{\lambda_{n-2}\lambda'_{n-2}}\frac{1}{n-1}\frac{N(\lambda_{n-1})N(\lambda'_{n-1})}{N(\lambda_{n-2})}.$$

Similar relations are known for those of SO(n).[1]

IV. DIFFERENTIAL OPERATORS

In this section, we discuss the expressions for $\bar{J}_j$ and J_j of the first- and the second-parameter groups, respectively. The expressions for these in terms of the matrix elements of U(n) as the parameters are given in Ref. 3.

The differential operators $\bar{J}_j$ and J_j are obtained from (3.4). We first discuss the expressions for $\bar{J}_j$ and J_j with $j=1,...,(n-1)^2-1$, i.e., those of SU($n-1$)$\subset$SU(n). It follows from (2.5) and the definition of $\bar{T}$ below (3.4) that the following relation holds:

$$\bar{T}_{jk} = \mathrm{tr}\Big(-2iE_k\frac{\partial g^{(n)}}{\partial\theta_j}g^{(n)\dagger}\Big) = \mathrm{tr}\Big(-2iE_k\frac{\partial g^{(n-1)}}{\partial\theta_j}g^{(n-1)\dagger}\Big), \quad (4.1)$$

for $\theta_j\neq(\theta_{nj},\phi_{nj},\psi_{nn-1})$ with $j=1,2,...,n-1$. Equation (4.1) means that the differential operators of the first-parameter group ($\bar{J}_j$) for $j=1,2,...,(n-1)^2-1$ are the same as for those of SU($n-1$). In order to consider J_j, we write $g^{(n)}$ in the form

$$g^{(n)}=g^{(n-1)}S^{(n)}_n,\quad g^{(2)}=S^{(2)}_2,$$
$$S^{(j)}_j = \Big(\prod_{l=1}^{k}e^{i\phi_{jl}\Gamma_{j-l}}e^{i\Theta_{jl}}E_{(j-l+1)^2-2}\Big)S^{(j)}_{j-k}. \quad (4.2)$$

Making use of the relation

$$(E_{j^2-2})_{st} = -(i/2)(\delta_{sj-1}\delta_{tj}-\delta_{sj}\delta_{tj-1}),$$

we obtain the relation

$$-i\Big(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_{jk}}\Big)_{st} = [(S^{(j)}_{j-k}S^{(j+1)}_{j+1}\cdots S^{(n)}_n)^\dagger E_{(j-k+1)^2-2}\times S^{(j)}_{j-k}S^{(j+1)}_{j+1}\cdots S^{(n)}_n]_{st}$$
$$= -(i/2)[(S^{(j)}_{j-k}S^{(j+1)}_{j+1}\cdots S^{(n)}_n)^\dagger_{sj-k}\times(S^{(j)}_{j-k}S^{(j+1)}_{j+1}\cdots S^{(n)}_n)_{j-k+1\,t} - (S^{(j)}_{j-k}S^{(j+1)}_{j+1}\cdots S^{(n)}_n)^\dagger_{sj-k+1}\times(S^{(j)}_{j-k}S^{(j+1)}_{j+1}\cdots S^{(n)}_n)_{j-k\,t}]. \quad (4.3)$$

It is easily seen that the following relations hold for $n\geqslant 2$:

$$(S_{j-k}^{(j)}S_{j+1}^{(j+1)}\cdots S_n^{(n)})_{j-k+1\,t} = (S_{j-k}^{(j)}S_j^{(j+1)}\cdots S_{n-1}^{(n)})_{j-k+1\,t},$$
$$(S_{j-k}^{(j)}S_{j+1}^{(j+1)}\cdots S_n^{(n)})_{j-k\,t} = (S_{j-k}^{(j)}S_j^{(j+1)}\cdots S_{n-1}^{(n)})_{j-k\,t}. \tag{4.4}$$

Taking into account (4.4), we can rewrite (4.3) in the form

$$-i\left(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_{jk}}\right)_{st} = \left[(S_{j-k}^{(j)}S_j^{(j+1)}\cdots S_{n-1}^{(n)})^\dagger E_{(j-k+1)^2-2} \times S_{j-k}^{(j)}S_j^{(j+1)}\cdots S_{n-1}^{(n)}\right]_{st}. \tag{4.5}$$

The right side of (4.5) is equal to $-i(g^{(n-1)\dagger}\,\partial g^{(n-1)}/\partial\theta_{j-1\,k-1})_{st}$ in which the angles $(\theta_{jk}, \phi_{jk}, \psi_{jj-1})$ are replaced by $(\theta_{j+1\,k+1}, \phi_{j+1\,k+1}, \psi_{j+1\,j})$. Similarly, the following relations are easily found for $n \geqslant 2$:

$$-i\left(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\phi_{jk}}\right)_{st} = -i\left[\left(g^{(n-1)\dagger}\frac{\partial g^{(n-1)}}{\partial\phi_{j-1\,k-1}}\right)_{st}\right]_{\{\theta_i\}\to\{\theta_{i+1}\}},$$
$$-i\left(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\psi_{k\,k-1}}\right)_{st} = -i\left[\left(g^{(n-1)\dagger}\frac{\partial g^{(n-1)}}{\partial\psi_{k-1\,k-2}}\right)_{st}\right]_{\{\theta_i\}\to\{\theta_{i+1}\}}, \tag{4.6}$$

where the notation $\{\theta_i\}\to\{\theta_{i+1}\}$ on the right side means the replacement $(\theta_{jk}, \phi_{jk}, \psi_{jj-1})\mapsto(\theta_{j+1\,k+1}, \phi_{j+1\,k+1}, \psi_{j+1\,j})$. It follows from (4.5) and (4.6) that the differential operators (J_j) for $j = 1,2,\ldots,((n-1)^2-1)$ are given by those of SU$(n-1)$ in which the angles $(\theta_{jk}, \phi_{jk}, \psi_{jj-1})$ are replaced by $(\theta_{j+1\,k+1}, \phi_{j+1\,k+1}, \psi_{j+1\,j})$. It, therefore, follows from the above discussion that in order to obtain the differential operators $\bar J_j$ and J_j of SU(n) it is sufficient for us to determine J_j $(\bar J_j)$ for $j = (n-1)^2,\ldots,n^2-1$ [or J_{jn} $(\bar J_{jn})$, $j = 1,2,\ldots,n-1$]. Furthermore, the operators J_{jn} $(\bar J_{jn})$ for $j < n-1$ are obtained from $J_{n-1\,n}$ and J_{n^2-1} $(=J_{n-1\,n-1})$ through the commutation relations.

A. The operators J_{n^2-1} and $\bar J_{n^2-1}$

Though the expression of J_{n^2-1} $(\bar J_{n^2-1})$ is obtained from (3.4), it is convenient to use the last of (3.7) and (3.8). The following relations are obtained from (3.8) and the last of (3.7):

$$J_{n^2-1}D^{(\lambda_n)}_{\{\lambda'_{n-1}\}\{\lambda_{n-1}\}}(g^{(n)}) = \frac{1}{\sqrt{2n(n-1)}}\left(n\sum_{k=1}^{n-1} m_{kn-1} - (n-1)\sum_{k=1}^{n} m_{kn}\right) \times D^{(\lambda_n)}_{\{\lambda'_{n-1}\}\{\lambda_{n-1}\}}(g^{(n)}).$$
$$\bar J_{n^2-1}D^{(\lambda_n)}_{\{\lambda'_{n-1}\}\{\lambda_{n-1}\}}(g^{(n)}) = \frac{1}{\sqrt{2n(n-1)}}\left(n\sum_{k=1}^{n-1} m'_{kn-1} - (n-1)\sum_{k=1}^{n} m_{kn}\right) \times D^{(\lambda_n)}_{\{\lambda'_{n-1}\}\{\lambda_{n-1}\}}(g^{(n)}). \tag{4.7}$$

The differential operator J_{n^2-1} $(\bar J_{n^2-1})$ will be given in terms of a linear combination of the differential operators of ϕ and ψ, because the D-matrix elements contain ϕ and ψ in the exponential form and must become the eigenfunction of J_{n^2-1} $(\bar J_{n^2-1})$. We therefore, assume the following expressions:

$$J_{n^2-1} = \frac{-2i}{\sqrt{2n(n-1)}} \times\left(\sum_{k=2}^{n}\sum_{j=1}^{k-1} a_{kj}\frac{\partial}{\partial\phi_{kj}} + \sum_{j=2}^{n} b_{jj-1}\frac{\partial}{\partial\psi_{jj-1}}\right),$$
$$\bar J_{n^2-1} = \frac{-2i}{\sqrt{2n(n-1)}} \times\left(\sum_{k=2}^{n}\sum_{j=1}^{k-1} \bar a_{kj}\frac{\partial}{\partial\phi_{kj}} + \sum_{j=2}^{n} \bar b_{jj-1}\frac{\partial}{\partial\psi_{jj-1}}\right). \tag{4.8}$$

Substituting (4.8) into (4.7) and comparing the coefficients of the m_{jk}'s on both sides after operating the differential operators, we obtain the following equations:

$$a_{21} = 0,\quad 2a_{31} - a_{21} - b_{21} = 0,\quad a_{n1} = n-1,$$
$$a_{n2} = -n,\quad b_{nn-1} = 0,\quad a_{nn-1} + b_{nn-1} = 0,$$
$$2a_{32} - a_{31} + 2b_{21} = 0,\quad a_{nj} = 0,\quad \text{for } 3 \leqslant j \leqslant n-2,$$
$$\left.\begin{array}{l}2a_{j1} - a_{j-1\,1} = 0,\\ 2a_{jj-1} + 2b_{j-1\,j-2} - a_{jj-2} = 0,\\ 2a_{j2} - a_{j1} - a_{j-1\,2} - b_{32}\delta_{j4} = 0,\end{array}\right\} \text{for } 4 \leqslant j \leqslant n,$$
$$2a_{jj-2} - a_{jj-3} - a_{j-1\,j-2} - b_{j-1\,j-2} = 0,\quad \text{for } 5 \leqslant j \leqslant n, \tag{4.9a}$$
$$2a_{kj} - a_{kj-1} - a_{k-1\,j} = 0,$$
$$\text{for } 6 \leqslant k \leqslant n \text{ and } 3 \leqslant j \leqslant k-3,$$

from the first of (4.7), and

$$\bar a_{21} = 0,\quad 2\bar a_{31} - \bar a_{21} - \bar b_{21} = 0,\quad 2\bar a_{n1} - \bar a_{n-1\,1} = n,$$
$$\bar a_{n1} = n-1,\quad 2\bar a_{32} - \bar a_{31} + 2\bar b_{21} = 0,\quad \bar b_{nn-1} = 0,$$
$$\bar a_{n2} = 0,\quad \bar a_{nn-1} + \bar b_{nn-1} = 0,\quad \bar a_{nj} = 0,$$
$$\text{for } 3 \leqslant j \leqslant n-1,$$
$$\left\{\begin{array}{l}2\bar a_{j1} - \bar a_{j-1\,1} = 0,\quad 2\bar a_{jj-1} + 2\bar b_{j-1\,j-2} - \bar a_{jj-2} = 0,\\ 2\bar a_{j2} - \bar a_{j1} - \bar a_{j-1\,2} - \bar b_{32}\delta_{j4} = 0,\quad \text{for } 4 \leqslant j \leqslant n,\end{array}\right. \tag{4.9b}$$
$$2\bar a_{jj-2} - \bar a_{jj-3} - \bar a_{j-1\,j-2} - \bar b_{j-1\,j-2} = 0,\quad \text{for } 5 \leqslant j \leqslant n,$$
$$2\bar a_{kj} - \bar a_{kj-1} - \bar a_{k-1\,j} = 0,$$
$$\text{for } 6 \leqslant k \leqslant n \text{ and } 3 \leqslant j \leqslant k-3,$$

from the second of (4.7). Equations (4.9a) and (4.9b) are easily solved as follows:

$$a_{jj-1} = 2^{n-2j+2}(-1)^j(3n-2j+3) \times\frac{(n-j)!}{(j-3)!(n-2j+3)!},$$
$$b_{jj-1} = 2^{n-2j+2}(-1)^j\frac{(n-j+1)!}{(j-2)!(n-2j+2)!},$$
$$a_{kj} = 2^{n-k-j+1}(-1)^{j-1}[n^2-(k-j+1)n+k+j-2] \times(n-k)!/[(j-1)!(n-k-j+2)!]$$
$$(k = j+2,\ldots,n;\ j = 1,\ldots,n-2),$$

$$\bar{a}_{jj-1} = 2^{n-2j+1}(-1)^j \frac{(3n-2j)(n-j-1)!}{(j-3)!(n-2j+2)!}$$

$$(j<n-1), \tag{4.10}$$

$$\bar{a}_{n1} = n-1, \quad \bar{a}_{nj} = 0 \quad (2\leqslant j<n-1),$$

$$\bar{b}_{jj-1} = 2^{n-2j+1}(-1)^j \frac{(n-j)!}{(j-2)!(n-2j+1)!},$$

$$\bar{a}_{kj} = 2^{n-k-j}(-1)^{j-1}[n^2-(k-j+3)n+2k]$$

$$\times(n-k-1)/[(j-1)!(n-k-j+1)!]$$

$$(k=j+2,...,n-1;\ j=1,...,n-3).$$

An inverse of a factorial of a negative number is to be put to zero. The expressions for J_{j^2-1} and $\bar{J}_{j^2-1}$ $(j=2,...,n-1)$ can, therefore, be given according to the procedure mentioned in the first part of this section. The expressions in the case of low n's are, of course, easily obtained directly from (3.4).

B. The operators $\bar{J}^{(\pm)}_{n-1\,n}$ and $\bar{J}^{(\pm)}_{n-1\,n}$

We rewrite (3.4) as follows:

$$-i\sum_{j=2}^{n}\sum_{l=1}^{j-1}\Big[\operatorname{tr}\Big(E_{lj}g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\Big)$$

$$-i\operatorname{tr}\Big(E_{(lj)}\,g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\Big)\Big]J^{(+)}_{lj}$$

$$-i\sum_{j=2}^{n}\sum_{l=1}^{j-1}\Big[\operatorname{tr}\Big(E_{lj}\,g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\Big)$$

$$+i\operatorname{tr}\Big(E_{(lj)}g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\Big)\Big]J^{(-)}_{lj}$$

$$-2i\sum_{j=2}^{n}\operatorname{tr}\Big(E_{j-1j-1}g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\Big)J_{j^2-1} = p_k, \tag{4.11}$$

where the notations $J^{(+)}_{jk} = J_{(k-1)^2+2j-2} + iJ_{(k-1)^2+2j-1}$ are used. The equations for $\bar{J}_j$ are obtained from (4.11) by the substitutions of $g^{(n)\dagger}\,\partial g^{(n)}/\partial\theta_k$ and $J^{(\pm)}_{ij}$ into $\partial g^{(n)}/\partial\theta_k\, g^{(n)\dagger}$ and $\bar{J}^{(\pm)}_j$. Equation (4.11) can be rewritten by taking into account the matrix elements of E_{ij} *and* $E_{(ij)}$ as follows:

$$-i\sum_{j=2}^{n}\sum_{l=1}^{j-1}\Big[\Big(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\Big)_{lj}J^{(+)}_{lj}$$

$$+\Big(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\Big)_{jl}J^{(-)}_{lj}\Big]$$

$$-i\sum_{j=2}^{n}\frac{2}{2j(j-1)}\Big[\sum_{l=1}^{j}\Big(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\Big)_{ll}$$

$$-j\Big(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_k}\Big)_{jj}\Big]J_{j^2-1} = p_k. \tag{4.12}$$

As is seen from (4.2)–(4.6), it is sufficient for us to consider the cases of $\theta_j \rightarrow (\theta_{j+1\,1},\phi_{j+1\,1})$ and $J_{j\,n}$ $(j=1,...,n-1)$. It is straightforward and elementary to obtain the expression for $J^{(+)}_{n-1\,n}$ from (4.12), and therefore we give only the result

$$J^{(+)}_{n-1\,n} = \sum_{j=1}^{n-1}\Delta^{(1)}_j\Big(-i\frac{\partial}{\partial\theta_{n-j+1\,1}}\Big)$$

$$+\sum_{j=1}^{n-1}\Delta^{(2)}_j\Big(-i\frac{\partial}{\partial\phi_{n-j+1\,1}}\Big)$$

$$-\sum_{j=2}^{n-1}\sum_{l=1}^{j-1}\sum_{k=1}^{n-1}[\Delta^{(1)}_k\{a_{kl}(j)J^{(+)}_{lj}+a^*_{kl}(j)J^{(-)}_{lj}\}$$

$$+\Delta^{(2)}_k\{b_{kl}(j)J^{(+)}_{lj}+b^*_{kl}(j)J^{(-)}_{lj}\}] \tag{4.13}$$

$$-\sum_{j=2}^{n}\frac{2}{\sqrt{2j(j-1)}}\sum_{k=1}^{n-1}\{c_{kj}\Delta^{(1)}_k$$

$$+c'_{kj}\Delta^{(2)}_k\}J_{j^2-1},$$

where

$$a_{kl}(j) = a^*_{kj}(l) = -i\Big(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\theta_{n-k+1\,1}}\Big)_{lj},$$

$$b_{kl}(j) = b^*_{kj}(l) = -i\Big(g^{(n)\dagger}\frac{\partial g^{(n)}}{\partial\phi_{n-k+1\,1}}\Big)_{lj},$$

$$c_{kj} = \sum_{l=1}^{j}a_{kl}(l) - ja_{kj}(j),$$

$$c'_{kj} = \sum_{l=1}^{j}b_{kl}(l) - jb_{kj}(j),$$

$$\Delta^{(1)}_j = ie^{(i/2)\psi_{2\,1}}[K_{n-j}/(A_{n-j})_{n-j+1\,n}],$$

$$\Delta^{(2)}_j = e^{(i/2)\psi_{2\,1}}\Big[\frac{1}{(A_{n-j})_{n-j+1\,n}}\frac{K_{n-j}}{\sin\theta_{n-j+1\,1}}$$

$$+\sum_{l=1}^{n-j-1}\frac{K_{n-j-l}}{(A_{n-j-l})_{n-j-l+1\,n}}\frac{\cot\theta_{n-j-l+1\,1}}{2^l}\Big],$$

$$(A_k)_{jn} = \Big(\prod_{p=k+2}^{n}\sin\frac{\theta_{p1}}{2}e^{(i/2)\phi_{p1}}\Big)\delta_{jk+1}, \quad \text{for } k\geqslant 2,$$

$$(A_1)_{jn} = \Big(\prod_{p=3}^{n}\sin\frac{\theta_{p1}}{2}e^{(i/2)\phi_{p1}}\Big)e^{-(i/2)\psi_{2\,1}}\delta_{j2},$$

$$K_{n-1} = \begin{cases}\cos\dfrac{\theta_{n2}}{2}e^{-(i/2)(\phi_{n2}+\psi_{2\,1})}, & \text{for } n>3,\\[2mm] \cos\dfrac{\theta_{3\,2}}{2}e^{-(i/2)(\psi_{2\,1}+\phi_{3\,2}+\psi_{3\,2})}, & \text{for } n=3,\end{cases}$$

$$K_j = \begin{cases}\Big(\prod_{p=3}^{n}\sin\dfrac{\theta_{p2}}{2}e^{(i/2)\phi_{p2}}\Big)e^{-(i/2)\phi_{3\,2}}, & \text{for } j=1,\\[2mm] \Big(\prod_{p=4}^{n}\sin\dfrac{\theta_{p2}}{2}e^{(i/2)\phi_{p2}}\Big)e^{-(i/2)(\phi_{3\,2}+\psi_{3\,2}+\psi_{2\,1})}, & \text{for } j=2,\\[2mm] \Big(\prod_{p=j+2}^{n}\sin\dfrac{\theta_{p2}}{2}e^{(i/2)\phi_{p2}}\Big)e^{-(i/2)(\psi_{2\,1}+\phi_{j+1\,2})} & \\ \quad\times\cos(\theta_{j+1\,2}/2), & \text{for } 3\leqslant j\leqslant n-2.\end{cases}$$

The expression for $J^{(-)}_{n-1\,n}(J_{n^2-3}-iJ_{n^2-2})$ is obtained from (4.13) by replacing $\partial/\partial\theta_{jk}$ and $\exp(ix)$ with $-\partial/\partial\theta_{jk}$ and $\exp(-ix)$, respectively.

Similarly, the expression for $\bar{J}^{(+)}_{n-1\,n}(=\bar{J}_{n^2-3}+i\bar{J}_{n^2-2})$ is given as follows:

$$\bar{J}^{(+)}_{n-1n} = \sum_{j=1}^{n-1}\bar{\Delta}^{(1)}_j p_j + \sum_{j=1}^{n-1}\bar{\Delta}^{(2)}_j p'_{j+1}$$

$$-\sum_{j=2}^{n-1}\sum_{l=1}^{j-1}\sum_{k=1}^{n-1}[\bar{\Delta}^{(1)}_k\{\bar{a}_{kl}(j)\bar{J}^{(+)}_{lj}+\bar{a}^*_{kl}(j)\bar{J}^{(-)}_{lj}\}$$

$$+\bar{\Delta}^{(2)}_k\{\bar{b}_{k+1\,l}(j)\bar{J}^{(+)}_{lj}+\bar{b}^*_{k+1\,l}(j)\bar{J}^{(-)}_{lj}\}]$$

$$-\sum_{j=2}^{n}\frac{2}{\sqrt{2j(j-1)}}\sum_{k=1}^{n-1}[\bar{\Delta}^{(1)}_k\bar{c}_{kj}$$

$$+\bar{\Delta}^{(2)}_k\bar{c}'_{k+1\,j}]\bar{J}_{j^2-1}, \tag{4.14}$$

where

$$p_j = -i\frac{\partial}{\partial\theta_{nj}},\quad p'_j = -i\frac{\partial}{\partial\phi_{nj}}\quad (j=1,2,\ldots,n-1),$$

$$p'_n = -i\frac{\partial}{\partial\psi_{n\,n-1}},$$

$$\bar a_{kl}(j) = \bar a^*_{kj}(l) = -i\left(\frac{\partial g^{(n)}}{\partial\theta_{nk}}g^{(n)\dagger}\right)_{lj},$$

$$\bar b_{kl}(j) = \bar b^*_{kj}(l) = -i\left(\frac{\partial g^{(n)}}{\partial\phi_{nk}}g^{(n)\dagger}\right)_{lj},$$

$$\bar b_{nk}(j) = \bar b^*_{nj}(k) = -i\left(\frac{\partial g^{(n)}}{\partial\psi_{n\,n-1}}g^{(n)\dagger}\right)_{kj},$$

$$\bar c_{kj} = \sum_{l=1}^{j}\bar a_{kl}(l) - j\,\bar a_{kj}(j),$$

$$\bar c'_{kj} = \sum_{l=1}^{j}\bar b_{kl}(l) - j\bar b_{kj}(j),$$

$$\bar\Delta^{(1)}_j = ie^{-i\phi_{nj}}\frac{(S^{(n-1)}_{n-1})^*_{n-1\,n-j}}{(U^{(n)}_{j-1})^*_{n\,n-j+1}},$$

$$\bar\Delta^{(2)}_j = \frac{2e^{-i\phi_{nj}}}{\sin\theta_{nj}}\frac{(S^{(n-1)}_{n-1})^*_{n-1\,n-j}}{(U^{(n)}_{j-1})^*_{n\,n-j+1}} - \tan\frac{\theta_{n\,j+1}}{2}\frac{(S^{(n-1)}_{n-1})^*_{n-1\,n-j-1}}{(U^{(n)}_{j})^*_{n\,n-j}}$$

for $j = 1,\ldots,n-3$,

$$\bar\Delta^{(2)}_{n-2} = -\frac{2e^{-i\phi_{n\,n-2}}}{\sin\theta_{n\,n-2}}\frac{(S^{(n-1)}_{n-1})^*_{n-1\,2}}{(U^{(n)}_{n-3})^*_{n3}} - \cot\theta_{n\,n-1}e^{-i\phi_{n\,n-1}}\frac{(S^{(n-1)}_{n-1})^*_{n-1\,1}}{(U^{(n)}_{n-2})^*_{n2}},$$

$$\bar\Delta^{(2)}_{n-1} = -\frac{e^{-i\phi_{n\,n-1}}}{\sin\theta_{n\,n-1}}\frac{(S^{(n-1)}_{n-1})^*_{n-1\,1}}{(U^{(n)}_{n-2})^*_{n2}},$$

$$S^{(n)}_n = U^{(n)}_{k-1}\left(\prod_{l=k}^{n-1}e^{i\phi_{nl}\Gamma_{n-l}}e^{i\theta_{nl}E_{(n-l+1)^2-2}}\right)\times e^{i\psi_{n\,n-1}\Gamma_1},$$

$$U^{(n)}_{k-1} = \prod_{l=1}^{k-1}e^{i\phi_{nl}\Gamma_{n-l}}e^{i\theta_{nl}E_{(n-l+1)^2-2}},$$

$$(U^{(n)}_{k-1})_{nj} = \prod_{l=1}^{k-1}\left(-e^{-(i/2)\phi_{nl}}\sin\frac{\theta_{nl}}{2}\right)\delta_{j\,n-k+1} - \left(\prod_{l=1}^{k-1}\left(-\cos\frac{\theta_{nl}}{2}e^{-(i/2)\phi_{nl}}\right)\right)\delta_{j\,n-k+2}.$$

The expression for $\bar J^{(-)}_{n-1\,n}$ is also obtained from (4.14) in the same way as in $J^{(-)}_{n-1\,n}$.

APPENDIX: PROOF OF BOTH SHIFT INVARIANCES OF THE MEASURE

In this appendix, we show that the determinant mentioned below (2.9) becomes the measure of SU$(n-1)$ and then the right shift invariant measure (2.8) is also invariant under the left shift.

The $(n-1)^2\times(n-1)^2$ determinant mentioned below (2.9) may be written as follows:

$$\begin{vmatrix} (A_1)_{1\,1}, & \ldots, & (A_{(n-1)^2-1})_{1\,1} & (C)_{1\,1} \\ \vdots & & \vdots & \vdots \\ (A_1)_{n-1\,n-1}, & \ldots, & (A_{(n-1)^2-1})_{n-1\,n-1} & (C)_{n-1\,n-1}\end{vmatrix}, \tag{A1}$$

where

$$A_j = \frac{\partial g^{(n-1)}}{\partial\phi_j}g^{(n-1)\dagger}\quad\text{for possible }\phi_j\quad (j=1,\ldots,(n-1)^2-1),$$

$$C = g^{(n-1)}\frac{\partial S^{(n)}_n}{\partial\phi_{nl}}S^{(n)\dagger}_n g^{(n-1)\dagger}.$$

Taking into account the property of the determinant and $g^{(n-1)}g^{(n-1)\dagger} = g^{(n-1)\dagger}g^{(n-1)} = 1$ with $\det g^{(n-1)} = 1$, we rewrite (A1) in the form

$$\begin{vmatrix} (B_1)_{1\,1}, & \ldots, & (B_{(n-1)^2-1})_{1\,1} & (C_1)_{11} \\ \vdots & & \vdots & \vdots \\ (B_1)_{n-1\,n-1}, & \ldots, & (B_{(n-1)^2-1})_{n-1\,n-1} & (C_1)_{n-1\,n-1}\end{vmatrix}, \tag{A2}$$

where

$$B_j = g^{(n-1)\dagger}\frac{\partial g^{(n-1)}}{\partial\phi_j},\quad C_1 = \frac{\partial S^{(n)}_n}{\partial\phi_{nl}}S^{(n)\dagger}_n.$$

Making use of the relation $(C_1)_{jk} = i(\delta_{jk}\delta_{jn-1} - \delta_{jk}\delta_{nj})/2$, we obtain from (A2)

$$\frac{i}{2}\begin{vmatrix} (B_1)_{1\,1}, & \ldots, & (B_{(n-1)^2-1})_{1\,1} \\ \vdots & & \vdots \\ (B_1)_{n-1\,n-2}, & \ldots, & (B_{(n-1)^2-1})_{n-1\,n-2}\end{vmatrix}. \tag{A3}$$

Equation (A3) gives the left-invariant measure of SU$(n-1)$.

The quantity $\Omega = dg^{(n)}g^{(n)\dagger}$ becomes for the left shift $g^{(n)}\to g^{(n)\prime} = Vg^{(n)}$ for $V\in \mathrm{SU}_{(n)}$,

$$\Omega' = V\Omega V^\dagger. \tag{A4}$$

The $(n^2-1)\times(n^2-1)$ determinant necessary for us is written in the form

$$\begin{vmatrix} (\Omega'_1)_{11}, & \ldots, & (\Omega'_{n^2-1})_{1\,1} \\ \vdots & & \vdots \\ (\Omega'_1)_{n1}, & \ldots, & (\Omega'_{n^2-1})_{n1} \\ (\Omega'_1)_{1\,2}, & \ldots, & (\Omega'_{n^2-1})_{1\,2} \\ \vdots & & \vdots \\ (\Omega'_1)_{n\,n-1}, & \ldots, & (\Omega'_{n^2-1})_{n\,n-1}\end{vmatrix}$$

$$= (-1)^{n^2}\begin{vmatrix} 0 & (\Omega'_1)_{1\,1} & \ldots, & (\Omega'_{n^2-1})_{1\,1} \\ \vdots & \vdots & & \vdots \\ 0 & (\Omega'_1)_{n\,n-1}, & \ldots, & (\Omega'_{n^2-1})_{n\,n-1} \\ 1 & (\Omega'_1)_{n\,n}, & \ldots, & (\Omega'_{n^2-1})_{n\,n}\end{vmatrix}, \tag{A5}$$

where the n^2-1 elements except for the (n,n)-th element in each Ω_j are ordered lexicographically and the second expression is obvious from the property of the determinant. The right side of (A5) can be rewritten as a product of the determinant of the V's, which consists of the n V's in the diagonal, and becomes, by taking into account det $V = 1$,

$$(-1)^{n^2}\begin{vmatrix} 0 & (\Omega_1 V^\dagger)_{11}, & \dots, & (\Omega_{n^2-1} V^\dagger)_{11} \\ \vdots & \vdots & & \vdots \\ 0 & (\Omega_1 V^\dagger)_{n-1\,1}, & \dots, & (\Omega_{n^2-1} V^\dagger)_{n-1\,1} \\ V^*_{n1} & (\Omega_1 V^\dagger)_{1n}, & \dots, & (\Omega_{n^2-1} V^\dagger)_{1\,n} \\ \vdots & \vdots & & \vdots \\ V^*_{nn} & (\Omega_1 V^\dagger)_{n\,n}, & \dots, & (\Omega_{n^2-1} V^\dagger)_{n\,n} \end{vmatrix}, \tag{A6}$$

where $VV^\dagger = V^\dagger V = 1$ is considered. Similarly, (A6) is rewritten as follows:

$$\begin{vmatrix} (\Omega_1)_{11} & \dots, & (\Omega_{n^2-1})_{11} \\ \vdots & & \vdots \\ (\Omega_1)_{n\,n-1}, & \dots, & (\Omega_{n^2-1})_{n\,n-1} \end{vmatrix}, \tag{A7}$$

which shows that our measure is left invariant.

[1]K. B. Wolf, J. Math. Phys. **12**, 197 (1971); T. Maekawa, J. Math. Phys. **20**, 691 (1977).

[2]F. D. Murnaghan, *The Unitary and Rotation Groups* (Spartan, Washington, D. C., 1962); A. Maduemezia, J. Math. Phys. **12**, 1681 (1971).

[3]J. D. Louck, Am. J. Phys. **38**, 3 (1970).

[4]N. Ja. Vilenkin, "Special functions and the theory of group representations," Mathematical Monographs, translated by V. N. Singh (Am. Math. Soc., Providence, RI, 1968), Vol. 22.

[5]G. Racah, "Group theory and spectroscopy," lecture notes, Princeton, 1951.

[6]H. Flanders, *Differential Forms with Applications to the Physical Sciences* (Academic, New York, 1963).

[7]I. M. Gel'fand and M. L. Tsetlin, Dokl. Akad. Nauk. SSSR **71**, 825, 1017 (1950).

[8]H. Weyl, *The Theory of Groups and Quantum Mechanics* (Dover, New York, 1950).

[9]G. E. Baird and L. C. Biedenharn, J. Math. Phys. **4**, 1449 (1963).

Formula for invariant integrations on SU(n)

Takayoshi Maekawa
Department of Physics, Kumamoto Univeristy, Kumamoto 860, Japan

(Received 21 January 1985; accepted for publication 12 April 1985)

The explicit formula for evaluating integrals of a product of the matrix elements of SU(n) is given in terms of Kronecker delta symbols.

I. PRELIMINARIES

The integrals involving the matrix elements of SU(n), which are necessary for the strong coupling expansion of lattice gauge theory in the case of SU(3) (see Ref. 1), are given by the graphical method.[2] In this paper, we give the explicit formula to the integrals by using the well-known group theoretical procedure.[3]

A parametrization of the elements of SU(n) and the Haar measure are given as follows[4]:

$$g^{(n)} = g^{(n-1)} S_n^{(n)}, \tag{1}$$

$$S_n^{(n)} = \left(\prod_{j=n}^{2} e^{i\phi_{nn-j+1}\Gamma_{j-1}} e^{i\theta_{nn-j+1}E_{(j^2-2)}} \right) e^{i\psi_{nn-1}\Gamma_1},$$

$$g^{(2)} = e^{i\phi_{21}\Gamma_1} e^{i\theta_{21}E_2} e^{i\psi_{21}\Gamma_1},$$

$$dV_n = \left[\prod_{i=1}^{n-1} \cos\frac{\theta_{ni}}{2} \left(\sin\frac{\theta_{ni}}{2} \right)^{2(n-i)-1} d\theta_{ni}\, d\phi_{ni} \right] \times d\psi_{nn-1}\, dV_{n-1}, \tag{2}$$

$$dV_2 = \cos(\theta_{21}/2) \sin(\theta_{21}/2) \mathrm{d}\theta_{21}\, \mathrm{d}\phi_{21}\, \mathrm{d}\psi_{21},$$

$$0 \leqslant \theta_{jk} \leqslant \pi, \quad 0 \leqslant \phi_{jk} < 4\pi, \quad 0 \leqslant \psi_{jj-1} < 2\pi.$$

The matrix elements of (1) are, for example, given by the expressions

$$(R_j)_{j-1\,j-1} = \cos(\theta/2) e^{i\phi/2},$$
$$(R_j)_{jj} = \cos(\theta/2) e^{-i\phi/2},$$
$$(R_j)_{j-1\,j} = \sin(\theta/2) e^{i\phi/2},$$
$$(R_j)_{j\,j-1} = -\sin(\theta/2) e^{-i\phi/2},$$
$$(R_j)_{kl} = \delta_{kl}, \quad \text{for } k \neq j, j-1,$$

where

$$R_j = e^{i\phi\Gamma_{j-1}} e^{i\theta E_{j^2-2}}.$$

If we introduce $dg = dV_n/V_n$ with $V_n = (4\pi)^n V_{n-1}/[2(n-1)!]$ and $V_2 = (4\pi)^2/2$, the measure dg is normalized to one on SU(n), i.e.,

$$\int_{\mathrm{SU}(n)} dg = 1. \tag{3}$$

It is our problem to calculate the integrals of the form

$$\int_{\mathrm{SU}(n)} dg\, g_{i_1j_1} \cdots g_{i_sj_s} g^*_{k_1l_1} \cdots g^*_{k_tl_t}, \tag{4}$$

where the superscript n on g is omitted. The integral (4) is easily calculated if we decompose a direct product quantity $g_{i_1j_1} \cdots g_{i_sj_s}$ into the irreducible components of the representation. Therefore, we first show that the representation given by g_{ij} is equivalent to a vector representation of SU(n).

The elements of the unitary irreducible representations (UIR) of U(n) are given by the Gel'fand and Tsetlin basis[5]

$$|m_{jk}\rangle = |\lambda_n, \ldots, \lambda_1\rangle, \tag{5}$$

where λ_j denotes $(m_{1j}, \ldots, m_{jj})$. The non-negative integers m_{jk} are subject to the conditions $m_{j\,k+1} \geqslant m_{j\,k} \geqslant m_{j+1\,k+1}$. The dimension of the UIR of U(n) is given by[3]

$$N(\lambda_n) = \frac{\Pi_{i<j}(t_i - t_j)}{(n-1)! \cdots 2!1!}, \quad t_j = m_{jn} + n - j. \tag{6}$$

It is noted that the representations of SU(n) are given by taking $m_{nn} = 0$, and the vector representation with the dimension n is given by $m_{1n} = 1$ and $m_{jn} = 0$ ($j = 2, \ldots, n$).

The representation D matrix elements of U(n) are defined as follows:

$$D^{(\lambda_n)}_{\{\lambda'_{n-1}\}\{\lambda_{n-1}\}}(g^{(n)}) = \langle \lambda_n \{\lambda'_{n-1}\} | D^{(n)}(g^{(n)}) | \lambda_n \{\lambda_{n-1}\} \rangle, \tag{7}$$

where $\{\lambda_{n-1}\}$ means $(\lambda_{n-1}, \ldots, \lambda_1)$ and $D^{(n)}(g^{(n)})$ denotes the representation matrix corresponding to (1). The orthogonality relation for (7) is given by[4,6]

$$\int_{\mathrm{SU}(n)} \overline{D^{(\lambda'_n)}_{\{\lambda'_{n-1}\}\{\lambda''_{n-1}\}}}(g) D^{(\lambda_n)}_{\{\lambda_{n-1}\}\{\lambda'''_{n-1}\}}(g)\, dg = \frac{1}{N(\lambda_n)} \delta_{\{\lambda'_n\}\{\lambda_n\}} \delta_{\{\lambda''_{n-1}\}\{\lambda'''_{n-1}\}}. \tag{8}$$

As noted below (6), the vector representation is given by taking $m_{1n} = 1$ and $m_{jn} = 0$ ($j = 2, \ldots, n$) and has dimension n. Then we may write the basis corresponding to (5) as follows:

$$|1,(m)\rangle \equiv |1, m_{n-1}, m_{n-2}, \ldots, m_1\rangle, \tag{9}$$

where (m) stands for $(m_{n-1}, \ldots, m_1)$ and m_j takes the values from 0 to m_{j+1}, which takes 0 or 1. There are n independent bases corresponding to the dimension n. The matrix elements of the vector representation are given in terms of (9) by

$$\langle 1, (m') | g^{(n)} | 1, (m) \rangle. \tag{10}$$

The matrix elements g_{ij} are specified by the basis $|j\rangle$ ($j = 1, \ldots, n$), which is given by a column vector with one on the jth and zeros on the other $n-1$ components. As we have n independent bases for both (9) and $|j\rangle$, we may write the relation between them as follows:

$$|j\rangle = \sum_{(m)} U_{j(m)} |1, (m)\rangle. \tag{11}$$

It follows from (11) that the coefficients $U_{j(m)}$ satisfy the unitarity conditions, i.e.,

$$\delta_{jk} = \langle j | k \rangle = \sum_{(m)} U^*_{j(m)} U_{k(m)},$$
$$\delta_{(m')(m)} = \langle 1,(m') | 1,(m) \rangle = \sum_j U^*_{j(m')} U_{j(m)}. \tag{12}$$

 0022-2488/85/081910-04$02.50

The matrix elements g_{ij} are given in terms of (10) as follows:

$$g_{ij} = \langle i|g^{(n)}|j\rangle$$
$$= \sum_{(m')}\sum_{(m)} U^*_{i(m')} U_{j(m)} \langle 1,(m')|g^{(n)}|1,(m)\rangle. \tag{13}$$

It is, therefore, evident from (13) that the representation g_{ij} is equivalent to the vector representation and the g_{ij}'s satisfy the following relations:

$$\int_{\mathrm{SU}(n)} g_{ij}\, dg = 0, \quad \int_{\mathrm{SU}(n)} g_{ij} g^*_{kl}\, dg = \frac{1}{n}\delta_{ik}\delta_{jl}, \tag{14}$$

where the relations (8) and (12) are used.

In order to decompose the direct product of the g_{ij}'s into irreducible ones, we explain Young's diagram briefly.[3] For a given integer s, a partition of s is considered as follows:

$$s = m_1 + \cdots + m_r, \quad m_1 \geqslant m_2 \geqslant \cdots \geqslant m_r > 0, \tag{15}$$

where $m_1,\cdots,m_r$ are all integers. To each partition of (15) corresponds the Young's diagram B. The diagram B has $f_{(m)}$ standard tableaux, which are defined by the condition that the numbers $(1,2,\ldots,s)$ increase in each row of B from left to right and in every column of B downwards,

$$f_{(m)} = s!\,\frac{\Pi_{i<j}(l_i - l_j)}{l_1!\cdots l_r!}, \tag{16}$$

where $(m) = (m_1,m_2,\ldots,m_r)$ and $l_j = m_j + r - j (j = 1,2,\ldots,r)$. We introduce Young's symmetrizer to one of the standard tableaux as follows:

$$\sum \epsilon_q q \sum p, \tag{17}$$

where p denotes any permutation which interchanges only the numbers of each row among themselves, Σp means the sum over all p, and q denotes the vertical permutation with plus or minus sign according to whether the permutation q is even or odd. Using (16) and (17), we can make the decomposition of the unit element as follows[3]:

$$1 = \sum \frac{f_{(m)}}{s!} e_B, \tag{18}$$

where the summation is taken over all possible partitions of s in (15) and e_B is the sum of the Young's symmetrizers corresponding to the standard tableaux of the diagram B.

II. CONSTRUCTION OF COMPUTATION FORMULA OF INTEGRAL

Making use of (18), we decompose the quantity contained in (4) into the irreducible components as follows:

$$g_{i_1 j_1} \cdots g_{i_s j_s} = \sum \frac{f_{(m)}}{s!} e_B\, g_{i_1 j_1} \cdots g_{i_s j_s}, \tag{19}$$

where the permutation in e_B is taken over the indices of j's for fixed i's. For example, e_B corresponding to the case of $s = 3$ with the partition $3 = 2 + 1$ and its action on the indices are given by

$$e_B = (1 - (13))(1 + (12)) + (1 - (12))(1 + (13)), \tag{20}$$

$$\begin{array}{|c|c|}\hline & \\ \hline & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array} \qquad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array} \qquad \begin{array}{|c|c|}\hline 1 & 3 \\ \hline 2 & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array}$$

$$\begin{aligned} e_B g_{i_1 j_1} g_{i_2 j_2} g_{i_3 j_3} = {} & g_{i_1 j_1} g_{i_2 j_2} g_{i_3 j_3} + g_{i_1 j_2} g_{i_2 j_1} g_{i_3 j_3} \\ & - g_{i_1 j_3} g_{i_2 j_2} g_{i_3 j_1} - g_{i_1 j_2} g_{i_2 j_3} g_{i_3 j_1} \\ & + g_{i_1 j_1} g_{i_2 j_2} g_{i_3 j_3} - g_{i_1 j_2} g_{i_2 j_1} g_{i_3 j_3} \\ & + g_{i_1 j_3} g_{i_2 j_2} g_{i_3 j_1} - g_{i_1 j_3} g_{i_2 j_1} g_{i_3 j_2}. \end{aligned}$$

It is, of course, known that the irreducible quantities corresponding to the different standard tableaux of the same diagram are equivalent to each other. Though it is possible to introduce the symmetrizer with respect to the i's in (19), it is sufficient for us to consider (19).

The matrix elements of the irreducible unitary representation of the compact group satisfy the following orthogonal relation[6]:

$$\int_g t^{(\alpha)}_{ij}(g) t^{(\beta)*}_{mn}(g) dg = \frac{1}{d_\alpha}\delta_{\alpha\beta}\delta_{im}\delta_{jn}, \tag{21}$$

where α is a parameter characterizing the irreducible representation, i and j specify the matrix elements of the representation matrix, and d_α is the dimension of the irreducible representation α. In the case of SU(n), the partition (15) gives the $f_{(m)}$ equivalent irreducible representations and the dimension of the irreducible representation corresponding to a standard tableau is given as follows[3]:

$$d_{(m)} = \prod_{i<j}(p_i - p_j)/[(n-1)!\cdots 2!1!], \tag{22}$$

where $(m) = (m_1,m_2,\ldots,m_r,0,\ldots,0)$ with $r \leqslant n$ and $n - r$ zeros, and $p_j = m_j + n - j$ $(j = 1,2,\ldots,n)$. Of course, the partition (m) corresponds to $(m_{1n},\ldots,m_{rn},0,\ldots,0)$ in the Gel'fand and Tsetlin basis and (22) to (6).

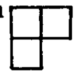
It follows from (14) that the indices i and j of g are used to specify the matrix elements of the fundamental representation of SU(n) that has the dimension n. The left side of (19), the direct product of g's, can be considered as the representation matrix elements such as $t_{(i_1\cdots i_s)(j_1\cdots j_s)}(g)$, which is reducible. Therefore, it is considered that (19) expresses the decomposition of the representation matrix elements into the irreducible components in which the equivalent representations are included. Taking into account (21) and the fact that the different partitions of (15) give nonequivalent representations, we easily calculate the integral (4).

As the right side of (19) can contain a scalar only for $s = n$ $(s < 2n)$, which corresponds to one column diagram, we obtain

$$\begin{aligned} & \int_{\mathrm{SU}(n)} g_{i_1 j_1} \cdots g_{i_s j_s}\, dg \\ & = \delta_{sn}\frac{1}{n!}\int_{\mathrm{SU}(n)} e_B g_{i_1 j_1} \cdots g_{i_n j_n}\, dg \\ & = \frac{1}{n!}\delta_{sn}\int_{\mathrm{SU}(n)} \begin{vmatrix} g_{i_1 j_1} & \cdots & g_{i_1 j_n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ g_{i_n j_1} & \cdots & g_{i_n j_n} \end{vmatrix} dg \\ & = \frac{1}{n!}\delta_{sn}\epsilon_{i_1\cdots i_n}\epsilon_{j_1\cdots j_n} \\ & = \frac{1}{n!}\delta_{sn} \begin{vmatrix} \delta_{i_1 j_1} & \cdots & \delta_{i_1 j_n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \delta_{i_n j_1} & \cdots & \delta_{i_n j_n} \end{vmatrix}, \end{aligned} \tag{23}$$

which agrees with that given for $s = n$ in Ref. 2. Here, $e_B = \Sigma\epsilon_q q$ in (23) corresponds to a column diagram and means all possible permutations over the indices of the j's with plus or minus one according to even or odd permutation q. Similarly, we can give the following formula to the integral (4) for $0<s,\ t\leqslant n$:

$$\int_{SU(n)} g_{i_1j_1}\cdots g_{i_sj_s} g^*_{k_1l_1}\cdots g^*_{k_tl_t}\,dg \tag{24}$$
$$=\delta_{st}\sum_{(m)}\frac{(f_{(m)})^2}{(s!)^2}\frac{1}{d_{(m)}}$$
$$\times\sum_P\{P\delta_{i_1k_1}\cdots\delta_{i_sk_s}(e_B\delta_{j_1l_1}\cdots\delta_{j_sl_s})\},$$

where e_B denotes the sum of the Young's symmetrizers formed from the standard tableaux corresponding to the diagram B and the permutations in e_B act on the indices of l's for fixed j's, $\Sigma_{(m)}$ means the sum over all possible partitions of $s\,(=m_1+\cdots+m_r, r\leqslant n)$ as in (15), and Σ_P means the sum in the same way over the indices k's and l's for fixed i's and j's. The last statement means, for instance, for $e_B=(1-(12))(1+(13))+(1-(13))(1+(12))$ corresponding to the diagram ,

$$\sum_P P\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}e_B\delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}$$
$$=\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}e_B\delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}+\delta_{i_1k_1}\delta_{i_2k_3}\delta_{i_3k_2}e_B\delta_{j_1l_1}\delta_{j_2l_3}\delta_{j_3l_2}+\delta_{i_1k_2}\delta_{i_2k_1}\delta_{i_3k_3}e_B\delta_{j_1l_2}\delta_{j_2l_1}\delta_{j_3l_3}$$
$$+\delta_{i_1k_2}\delta_{i_2k_3}\delta_{i_3k_1}e_B\delta_{j_1l_2}\delta_{j_2l_3}\delta_{j_3l_1}+\delta_{i_1k_3}\delta_{i_2k_1}\delta_{i_3k_2}e_B\delta_{j_1l_3}\delta_{j_2l_1}\delta_{j_3l_2}+\delta_{i_1k_3}\delta_{i_2k_2}\delta_{i_3k_1}e_B\delta_{j_1l_3}\delta_{j_2l_2}\delta_{j_3l_1},$$
$$e_B\delta_{j_1'l_1}\delta_{j_2'l_2}\delta_{j_3'l_3}=2\delta_{j_1'l_1}\delta_{j_2'l_2}\delta_{j_3'l_3}-\delta_{j_1'l_3}\delta_{j_2'l_1}\delta_{j_3'l_2}-\delta_{j_1'l_2}\delta_{j_2'l_3}\delta_{j_3'l_1}.$$

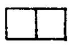
Though the integral (24) is given for $s,t\leqslant n$, those for $s>n$ $(t>n)$ or $s,t>n$ are easily found. For instance, in the case of $s=n+t$ and $t<n$, the partition of s which contributes to the integral is $(m)=(m_1,m_2,\ldots m_{n-1},1)$ and the representations given by the partition are equivalent to those given by $(m')=(m_1-1,m_2-1,\ldots,m_{n-1},0)$ with $t=m_1'+\cdots+m_{n-1}',m_1'\geqslant m_2'\geqslant\cdots\geqslant m_{n-1}'\geqslant 0$. The explicit form of the decomposition into the irreducible components is given as follows:

$$g_{i_1j_1}\cdots g_{i_sj_s}\cong\sum\frac{f_{(m)}}{s!}\epsilon_{i_1\cdots i_n}\epsilon_{j_1\cdots j_n}e_Bg_{i_{n+1}j_{n+1}}\cdots g_{i_sj_s}, \tag{25}$$

where on the right side only the typical term, which contributes to the integrals, is written and therefore we must add over the terms corresponding to the standard tableaux. Then the integral is easily found because only the same form of the integral as in (24) is needed.

III. SIMPLE EXAMPLES

Some simple examples for the integral (15) are given below.

(1) For $s=2$. There are two diagrams  and . The dimensions of their irreducible representations are

$$d_{(2,0,\ldots,0)}=n(n+1)/2,\quad d_{(1,1,0,\ldots,0)}=n(n-1)/2.$$

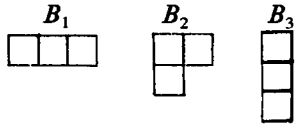
The integral (24) gives

$$\int_{SU(n)}g_{i_1j_1}g_{i_2j_2}g^*_{k_1l_1}g^*_{k_2l_2}\,dg=\frac{1}{(2)^2}\frac{1}{d_{(2,0,\ldots,0)}}\sum P\delta_{i_1k_1}\delta_{i_2k_2}(1+(12))\delta_{j_1l_1}\delta_{j_2l_2}$$
$$+\frac{1}{(2)^2}\frac{1}{d_{(1,1,0,\ldots,0)}}\sum P\delta_{i_1k_1}\delta_{i_2k_2}(1-(12))\delta_{j_1l_1}\delta_{j_2l_2}$$
$$=\frac{1}{2n(n+1)}(\delta_{i_1k_1}\delta_{i_2k_2}+\delta_{i_1k_2}\delta_{i_2k_1})(\delta_{j_1l_1}\delta_{j_2l_2}+\delta_{j_1l_2}\delta_{j_2l_1})$$
$$+\frac{1}{2n(n-1)}(\delta_{i_1k_1}\delta_{i_2k_2}-\delta_{i_1k_2}\delta_{i_2k_1})(\delta_{j_1l_1}\delta_{j_2l_2}-\delta_{j_2l_1}\delta_{j_1l_2}), \tag{26}$$

which leads to the same expression as in Ref. 2.

(2) For $s=3$. There are three diagrams:

B_1 B_2 B_3

.

The dimensions of their irreducible representations are, respectively,

$$d_{(3,0,\ldots,0)}=n(n+1)(n+2)/3!,\quad d_{(2,1,0,\ldots,0)}=n(n^2-1)/3,\quad d_{(1,1,1,0,\ldots,0)}=n(n-1)(n-2)/3!.$$

The integral (24) gives

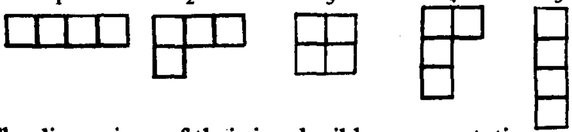
$$\int_{SU(n)}g_{i_1j_1}g_{i_2j_2}g_{i_3j_3}g^*_{k_1l_1}g^*_{k_2l_2}g^*_{k_3l_3}\,dg$$
$$=\frac{1}{3!n(n+1)(n+2)}\sum P\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}e_{B_1}\delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}+\frac{1}{3n(n^2-1)}\sum P\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}e_{B_2}\delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}$$
$$+\frac{1}{3!n(n-1)(n-2)}\sum P\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}e_{B_3}\delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}, \tag{27}$$

where

$$e_{B_1} = 1 + (12) + (13) + (23) + (123) + (132),$$
$$e_{B_2} = (1-(12))(1+(13)) + (1-(13))(1+(12)),$$
$$e_{B_3} = 1 - (12) - (13) - (23) + (123) + (132).$$

(3) For $s = 4$. There are five diagrams:

B_1 B_2 B_3 B_4 B_5



The dimensions of their irreducible representations are, respectively,

$$d_{(4,0,\ldots,0)} = n(n+1)(n+2)(n+3)/4!, \quad d_{(3,1,0,\ldots,0)} = n(n-1)(n+1)(n+2)/8,$$
$$d_{(2,2,0,\ldots,0)} = n^2(n^2-1)/12, \quad d_{(2,1,1,0,\ldots,0)} = n(n^2-1)(n-2)/8, \quad d_{(1,1,1,1,0,\ldots,0)} = n(n-1)(n-2)(n-3)/4!.$$

The expression (16) gives

$$f_{(4,0,\ldots,0)} = f_{(1,1,1,1,0,\ldots,0)} = 1, \quad f_{(2,2,0,\ldots,0)} = 2, \quad f_{(3,1,0,\ldots,0)} = f_{(2,1,1,0,\ldots,0)} = 3.$$

The integral (24) gives

$$\begin{aligned}
\int_{SU(n)} & g_{i_1j_1} g_{i_2j_2} g_{i_3j_3} g_{i_4j_4} g^*_{k_1l_1} g^*_{k_2l_2} g^*_{k_3l_3} g^*_{k_4l_4}\, dg \\
&= \frac{1}{4!n(n+1)(n+2)(n+3)} \sum P\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}\delta_{i_4k_4} e_{B_1} \delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}\delta_{j_4l_4} \\
&\quad + \frac{1}{8n(n-1)(n+1)(n+2)} \sum P\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}\delta_{i_4k_4} e_{B_2} \delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}\delta_{j_4l_4} \\
&\quad + \frac{1}{12n^2(n^2-1)} \sum P\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}\delta_{i_4k_4} e_{B_3} \delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}\delta_{j_4l_4} \\
&\quad + \frac{1}{8n(n+1)(n-1)(n-2)} \sum P\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}\delta_{i_4k_4} e_{B_4} \delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}\delta_{j_4l_4} \\
&\quad + \frac{1}{4!n(n-1)(n-2)(n-3)} \sum P\delta_{i_1k_1}\delta_{i_2k_2}\delta_{i_3k_3}\delta_{i_4k_4} e_{B_5} \delta_{j_1l_1}\delta_{j_2l_2}\delta_{j_3l_3}\delta_{j_4l_4},
\end{aligned} \tag{28}$$

where

$$\begin{aligned}
e_{B_1} &= 1 + (12) + (13) + (14) + (23) + (24) + (34) + (12)(34) + (13)(24) + (14)(23) + (123) + (132) + (124) \\
&\quad + (142) + (134) + (143) + (234) + (243) + (1234) + (1243) + (1324) + (1342) + (1423) + (1432), \\
e_{B_2} &= (1-(12))(1+(13)+(14)+(34)+(134)+(143)) + (1-(13))(1+(12)+(14)+(24)+(124)+(142)) \\
&\quad + (1-(14))(1+(12)+(13)+(23)+(123)+(132)), \\
e_{B_3} &= (1-(12))(1-(34))(1+(13))(1+(24)) + (1-(13))(1-(24))(1+(12))(1+(34)), \\
e_{B_4} &= (1-(12)-(14)-(24)+(124)+(142))(1+(13)) + (1-(12)-(13)-(23)+(123)+(132))(1+(14)), \\
e_{B_5} &= 1 - (12) - (13) - (14) - (23) - (24) - (34) + (12)(34) + (13)(24) + (14)(23) + (123) + (132) + (124) \\
&\quad + (142) + (134) + (143) + (234) + (243) - (1234) - (1243) - (1324) - (1342) - (1423) - (1432).
\end{aligned}$$

It is noted that the expressions corresponding to a row and a column on the right side of (24) can be rewritten, respectively, as follows:

$$\frac{(n-1)!}{(n+s-1)!s!}\left(\sum P\delta_{i_1k_1}\cdots\delta_{i_sk_s}\right)\left(\sum P\delta_{j_1l_1}\cdots\delta_{j_sl_s}\right),$$
$$\frac{(n-s)!}{n!s!}\left(\sum \epsilon_p P\delta_{i_1k_1}\cdots\delta_{i_sk_s}\right)\left(\sum \epsilon_p P\delta_{j_1l_1}\cdots\delta_{j_sl_s}\right), \tag{29}$$

where ΣP means the sum over all permutations of the indices of k's (l's) for fixed i's (j's), and similarly $\Sigma\epsilon_p P$ means the same as ΣP except for plus or minus one according to whether P is an even or odd permutation.

[1]K. Wilson, Phys. Rev. D. **10**, 2445 (1974); L. P. Kadanoff, Rev. Mod. Phys. **49**, 2679 (1977).

[2]M. Creutz, Rev. Mod. Phys. **50**, 561 (1978); J. Math. Phys. **19**, 2046 (1978).

[3]H. Boerner, *Representations of Groups with Special Consideration for the Needs of Modern Physics* (North-Holland, Amsterdam, 1970).

[4]T. Maekawa, J. Math Phys. **26**, 1902 (1985).

[5]I. M. Gel'fand and M. L. Tsetlin, Dokl. Akad. Nauk SSSR **71**, 825 (1950).

[6]N. Ja. Vilenkin, *Special Functions and the Theory of Group Representations, Mathematical Monographs Vol. 22*, translated by V. N. Singh (Am. Math. Soc., Providence, RI, 1968).

New expressions for the eigenvalues of the invariant operators of the general linear and the orthosymplectic Lie superalgebras

C. O. Nwachuku[a]
International Centre for Theoretical Physics, Trieste, Italy
M. A. Rashid
Department of Mathematics, Ahmadu Bello University, Zaria, Nigeria

(Received 27 August 1984; accepted for publication 25 January 1985)

We obtain expansions for the eigenvalues C_p of the invariant operators (Casimir operators) of the general linear, and the orthosymplectic Lie superalgebras in terms of *products* of suitably defined graded power sums P_k. The resulting expressions are closed and provide unified formulas for computing the C_p's for those superalgebras as well as their corresponding Lie algebras. The formulas are remarkably simple to suggest that the power sums used in this text could play a more basic role in the understanding of the pattern of the expansion coefficients. Explicit illustrations are given for the various series for $p \leqslant 8$.

I. INTRODUCTION

The problem of constructing the generalized invariant operators of the general linear, the special linear, and the orthosymplectic Lie superalgebras has been successfully accomplished.[1–4] The related problem of finding the generalized eigenvalues of these invariants in simple and convenient forms continues to receive attention.[5–15]

Recently Scheunert,[6] and Bincer,[7] working in tensor basis, were able to obtain expressions for the eigenvalues of these superalgebras in a particularly useful closed form, analogous to those of Perelomov and Popov[8] for the unitary, and the present authors[13] for the orthogonal and the symplectic Lie algebras [see Eq. (2.8) of this paper].

As recognized by earlier authors,[9,14] this form of expression though useful for studying the analytic structures of the C_p, is not particularly suited for practical calculations even for low values of p, since the summation as well as the product runs through to the rank of the algebra. The equation also contains fractional terms, whereas the C_p is a polynomial in the λ's, and there must be cancellations. Some successful attempts have been made in the past for the various series of the classical Lie algebras to rewrite the C_p using generating functions, as a polynomial expansion in the λ's. Perhaps the most convenient form so far as a result of this endeavor is the power sums *product* expansion formula for the C_p, obtained for the u(n) by Popov,[10] and for the o(n) and the sp($2n$) by one of the present authors.[15]

In what follows we attempt to express directly the eigenvalues of the general linear u(n, m), the special linear su(n, m) and the orthosymplectic osp(N, $2m$) Lie superalgebras, as sums of products of suitably defined power sums, obtaining expressions for these Lie superalgebras analogous to those of Refs. (10) and (15) for the Lie algebras, and thus tying together the treatment of this aspect of the problem.

In so doing we have chosen to expand in terms of the naturally occurring graded power sum P_k defined in Eq. (2.15). With this power sum, half of the terms in the coefficients of the expansion vanish resulting in considerable time saved. The final result appears remarkably simple to suggest that the P_k's and therefore their nongraded limits P_k^0 could play a useful role in the study of the general patterns for these coefficients. Our result also affords a direct proof of the *Ansatz* obtained by Balantekin and Bars[2,5] for the su(n), and conjectured by them for the o(N). Finally, for the case of the sp($2n$) we obtain the corresponding *Ansatz*.

In order to fix the notations and keep the presentation as self-contained as practicable, we outline briefly in Sec. II some of the earlier results which we need to derive our main equations in Sec. III. Illustrative examples are outlined in Sec. IV, which also contains pertinent discussions. There are appendices which outline the properties of the coefficients $Q_l(\nu)$ and $\beta_p(\nu)$ for completeness.

II. NOTATIONS AND DEFINITIONS

The infinitesimal generators $X^i_{\ j}$, in the canonical two index form (Racah basis) of the general linear, and the orthosymplectic Lie superalgebras satisfy the commutation relations, which can be written in a single form as

$$\begin{aligned}\langle X^i_{\ j}, X^k_{\ l}\rangle = {}& \delta^k_j X^i_{\ l} - (-1)^{(\eta_i+\eta_j)(\eta_k+\eta_l)}\delta^i_l X^k_{\ j} \\ & - (-1)^{\eta_i\eta_j}\epsilon_{-i}\epsilon_j(\delta^k_{-i} X_l^{\ -j} \\ & - (-1)^{(\eta_i+\eta_j)(\eta_k+\eta_l)}\,\delta_l^{-j} X^k_{\ -i}),\end{aligned} \tag{2.1}$$

where the supercommutator $\langle\ ,\ \rangle$ is defined by

$$\langle X^i_{\ j}, X^k_{\ l}\rangle = X^i_{\ j} X^k_{\ l} - (-1)^{(\eta_i+\eta_j)(\eta_k+\eta_l)} X^k_{\ l} X^i_{\ j}, \tag{2.2}$$

with the index grading

$$\eta_i = \begin{cases} 0, & -n \leqslant i \leqslant n, \\ 1, & n+1 \leqslant |i| \leqslant n+m, \end{cases} \tag{2.3}$$

and

$$\epsilon_i = \begin{cases} 1, & -n \leqslant i \leqslant n \\ \operatorname{sgn} i, & n+1 \leqslant |i| \leqslant n+m. \end{cases} \tag{2.4}$$

Equation (2.1) as it stands in full with the indices running from $-n-m$ to $n+m$ describes the orthosymplectic Lie superalgebra osp(N, $2m$). Here N stands for $2n$ or $2n+1$, and zero is included only when N is odd. If the η_i's are set

[a] On sabbatical leave from the Department of Mathematics, University of Benin, Benin City, Nigeria.

 0022-2488/85/081914-07$02.50

equal to zero (one) we recover the Lie algebra o(N) [sp($2n$)]. If the indices are restricted to vary from 1 to $n+m$ only, the general linear Lie superalgebra u(n, m), is obtained. In this case the Lie algebra u(n) is recovered by setting the η_i's equal to zero.

The pth-order invariant (Casimir) operators $\hat{C}_p$ for each of the superalgebras are defined by

$$\hat{C}_p = \sum_{i_1,\dots,i_p} X^{i_p}_{i_1}\sigma_{i_1}X^{i_1}_{i_2}\sigma_{i_2}\cdots X^{i_{p-1}}_{i_p}, \quad p\geqslant 1, \tag{2.5}$$

$$\hat{C}_0 = \sum_i \sigma_i, \quad \text{where } \sigma_i = (-1)^{\eta_i}, \tag{2.6}$$

and the summation runs over the appropriate ranges of the indices in each case. Clearly, these operators commute accordingly with all the generators of the algebra, and by the generalized Schur lemma, they are constant multiples of the identity matrix in any irreducible representation, except possibly when the dimensions of the bosonic and fermionic subspaces of the superalgebra are equal. This case is therefore excluded from consideration in this paper.[16]

For representations with unique highest weight, by acting recursively p times on the highest weight with the operator $T^i_j(p)$ defined by

$$T^i_j(p) = \sum_{i_1} (X^{p-1})^i_{i_1}\sigma_{i_1}X^{i_1}_j, \tag{2.7}$$

and contracting the i and j indices afterwards, one obtains the expression for the eigenvalues C_p of the Casimir operators of the superalgebras as sums of the elements of the corresponding Perelomov–Popov matrix A_{ij} raised to the power p. On diagonalizing the A_{ij}, the C_p can be written in the form[6,7]

$$C_p = \sum_i \sigma_i \lambda_i^p \zeta_i \prod_{j\neq \pm i,0}\left(1 - \frac{\sigma_j}{\lambda_i - \lambda_j}\right), \tag{2.8}$$

where the summation and the product run over the respective ranges as indicated earlier for each superalgebra.

In Eq. (2.8)

$$\zeta_i = \begin{cases} 1, & \text{for the u}(n, m), \\ 1 + \dfrac{\sigma_i - 1}{\lambda_{-i} - \lambda_i}, & \text{for the osp}(2n, 2m), \\ 1 + (1-\delta_{i0})\dfrac{\sigma_i + 1}{\lambda_{-i} - \lambda_i}, & \text{for the osp}(2n+1, 2m). \end{cases} \tag{2.9}$$

Here the λ_i's are the diagonal elements of A_{ij} given by

$$\lambda_i = \begin{cases} \sigma_i f_i + r_i + \frac{1}{2}(n - m - \sigma_i), & \text{for the u}(n, m), \\ \sigma_i f_i + r_i + \frac{1}{2}(2n - 2m - 1 - \sigma_i), & \text{for the osp}(2n, 2m), \\ \sigma_i f_i + r_i + \frac{1}{2}(2n - 2m - \sigma_i + \delta_{i0}), & \text{for the osp}(2n+1, 2m), \end{cases} \tag{2.10}$$

where f_i is the number of boxes in the ith row of the Young supertableaux corresponding to the highest weight of the irreducible representation of interest. The quantities r_i take different values for the different superalgebras. For the u(n, m),

$$r_i = \frac{1}{2}\left(\sum_{j>i}\sigma_j - \sum_{j<i}\sigma_j\right), \tag{2.11}$$

while for the osp ($2n$, $2m$), and the osp($2n+1$, $2m$), respectively,

$$r_i = \sum_{j>i\geqslant 1}\sigma_j + \frac{1}{2}(\sigma_i - 1), \tag{2.12}$$

and

$$r_i = \sum_{j>i\geqslant 0}\sigma_j + \frac{1}{2}\sigma_i, \tag{2.13}$$

with the subsidiary conditions $r_{-i} = -r_i$ in the two latter cases.

It remains to define the graded power sums. For this purpose it is convenient to introduce the symbol

$$\rho_i = \lambda_i - \sigma_i f_i. \tag{2.14}$$

Then for each of the superalgebras, the graded power sums P_k are defined by

$$P_k = \sum_i \sigma_i\left[\left(\lambda_i + \frac{\sigma_i}{2}\right)^k - \left(\rho_i + \frac{\sigma_i}{2}\right)^k\right], \tag{2.15}$$

with $P_1 = 0$ for the su(n, m) and the osp(N, $2m$). Here the summation again runs over the ranges of i.

III. THE PRODUCT POWER SUM EXPANSION FORMULAS

A. The integral representation and generating functions

In this subsection it is convenient to separate the cases.

(i) The u(n, m) and the su(n,m). For the u(n, m) and the su(n, m), Eq. (2.8) together with Eq. (2.9) can be recast in the integral form

$$C_p = -\frac{1}{2\pi i}\oint d\lambda\, \lambda^p \prod_{i=1}^{n+m}\left(1 - \frac{\sigma_i}{\lambda - \lambda_i}\right), \tag{3.1}$$

where the integration is over a large circle with the origin as center on the complex λ plane, containing all the poles. By setting $\lambda = 1/z$, we have

$$C_p = -\frac{1}{2\pi i}\oint dz\, \frac{f(z)}{z^{p+2}}, \tag{3.2}$$

which implies the identity

$$1 - f(z) = \sum_{p=0}^{\infty} C_p z^{p+1}, \tag{3.3}$$

where

$$f(z) = \prod_{i=1}^{n+m}\left(1 - \frac{\sigma_i z}{1 - \lambda_i z}\right), \tag{3.4}$$

and the integration in Eq. (3.2) is along a closed path containing the origin but excluding all the poles of $f(z)$.

Furthermore, let $f_0(z)$ be the corresponding function $f(z)$ for the identity representation for which $f_i = 0$ for all i. Then

$$f_0(z) = \prod_{i=1}^{n+m}\left(1 - \frac{\sigma_i z}{1-\rho_i z}\right) \tag{3.5}$$

$$= \exp\left(-\sum \frac{z^k}{k}(n-m)^k\right). \tag{3.6}$$

(ii) The osp($2n$, $2m$). For this case, Eqs. (2.8) and (2.9) become

$$C_p = \sum_{i=-n-m}^{n+m} \sigma_i \lambda_i^p \frac{\lambda_i - n + m + 1}{\lambda_i - n + m + \frac{1}{2}} \times \prod_{\substack{j=-n-m \\ \neq i}}^{n+m}\left(1 - \frac{\sigma_j}{\lambda_i - \lambda_j}\right) \tag{3.7}$$

$$= -\frac{1}{2\pi i}\oint d\lambda\, \lambda^p \frac{\lambda - n + m + 1}{\lambda - n + m + \frac{1}{2}} \times \prod_{i=-n-m}^{n+m}\left(1 - \frac{\sigma_i}{\lambda - \lambda_i}\right) + \frac{1}{2}\left(n - m - \frac{1}{2}\right)^p. \tag{3.8}$$

Here, the extra term in Eq. (3.8) takes care of the pole in the integrand at $\lambda = n - m - \frac{1}{2}$ which is not included in the summation Eq. (3.7); and we have made use of the identity

$$\prod_{i=-n-m}^{n+m}\left(1 - \frac{\sigma_i}{n - m - \frac{1}{2} - \lambda_i}\right) = 1. \tag{3.9}$$

In terms of the variable $z = 1/\lambda$, Eq. (3.8) becomes

$$C_p = -\frac{1}{2\pi i}\oint \frac{dz}{z^{p+2}} f(z) + \frac{1}{2}\left(n - m - \frac{1}{2}\right)^p, \tag{3.10}$$

which implies the identity

$$1 - f(z) = \sum_{p=0}^{\infty} C_p z^{p+1} - \frac{1}{2}\sum_{p=0}^{\infty}\left(n - m - \frac{1}{2}\right)^p z^{p+1}, \tag{3.11}$$

where

$$f(z) = \frac{1 - (n - m - 1)z}{1 - (n - m - \frac{1}{2})z}\prod_{i=-n-m}^{n+m}\left(1 - \frac{\sigma_i z}{1 - \lambda_i z}\right). \tag{3.12}$$

Furthermore, as in the previous case

$$f_0(z) = f(z)|_{f_i = 0} = \exp\left(-\sum_{k\geq 1}\frac{z^k}{k}[(2n - 2m - 1)^k + (n - m)^k - (n - m - \tfrac{1}{2})^k]\right). \tag{3.13}$$

(iii) The osp($2n+1$, $2m$). In a similar way, the C_p for the osp($2n+1$, $2m$) can be written in the form

$$C_p = \sum_{i=-n-m}^{n+m} \sigma_i \lambda_i^p \frac{\lambda_i - n + m - \frac{1}{2}}{\lambda_i - n + m - 1} \times \prod_{\substack{j=-n-m \\ \neq i}}^{n+m}\left(1 - \frac{\sigma_j}{\lambda_i - \lambda_j}\right) + \frac{1}{2}\lambda_0^p \tag{3.14}$$

$$= -\frac{1}{2\pi i}\oint \frac{dz}{z^{p+2}} f(z) + \frac{1}{2}(n - m)^p, \tag{3.15}$$

where

$$f(z) = \frac{1 - (n - m + \frac{1}{2})z}{1 - (n - m + 1)z}\prod_{i=-n-m}^{n+m}\left(1 - \frac{\sigma_i z}{1 - \lambda_i z}\right), \tag{3.16}$$

and the corresponding identity

$$\prod_{i=-n-m}^{n+m}\left(1 - \frac{\sigma_i z}{n - m - \lambda_i}\right) = 1, \tag{3.17}$$

has been used.

Equation (3.16) implies the Taylor sum

$$1 - f(z) = \sum_{p=0}^{\infty} C_p z^{p+1} - \frac{1}{2}\sum_{p=0}^{\infty}(n - m)^p z^{p+1}. \tag{3.18}$$

The corresponding $f_0(z)$ is given by

$$f_0(z) = \exp\left(-\sum_{k=1}\frac{z^k}{k}[2n - 2m)^k + (n - m + \tfrac{1}{2})^k - (n - m)^k]\right). \tag{3.19}$$

B. The expansion formula

If for each superalgebra we take the logarithm of $f(z)$ and use the corresponding relations, Eqs. (3.6), (3.13), and (3.19), respectively, an expansion involving the exponential of functions of power sums can be obtained in a form given by Scheunert[6] which is analogous to the results in Refs. 9 and 14 for the u(n), and the o(N) and sp($2n$), respectively. These expressions are in themselves tedious to handle. Instead, we seek an expansion in terms of the elementary products of the power sums P_k in which the combining coefficients are completely determined.

With $f(z)$ and $f_0(z)$ defined in the preceding subsection, it is now convenient to treat the three cases together. For each superalgebra, let $F(z)$ be defined by the relation

$$f(z) = f_0(z)F(z), \tag{3.20}$$

so that from Eqs. (3.4) and (3.5), (3.12) and (3.13), (3.16) and (3.19), we have

$$\begin{aligned}\ln F(z) &= \sum_i\left\{\ln\left[1 - \frac{\sigma_i z}{1 - \lambda_i z}\right] - \ln\left[1 - \frac{\sigma_i z}{1 - \rho_i z}\right]\right\} \\ &= -\sum_{k\geq 1}\frac{z^k}{k}\sum_i\{[(\lambda_i + \sigma_i)^k - \lambda_i^k] - [(\rho_i + \sigma_i)^k - \rho_i^k]\} \\ &= -\sum_{k\geq 1}\frac{z^k}{k}\sum_{l=1}^{k}\binom{k}{l}\sum_i\left(\left(\frac{\sigma_i}{2}\right)^l - \left(-\frac{\sigma_i}{2}\right)^l\right) \times[(\lambda_i + \sigma_i/2)^{k-l} - (\rho_i + \sigma_i/2)^{k-l}] \\ &= -\sum_{k\geq 1}\frac{z^k}{k}\sum_{l=1}^{k}\binom{k}{l} 2^{1-l}\xi^l P_{k-l},\end{aligned} \tag{3.21}$$

where

$$\xi^l = \tfrac{1}{2}(1 - (-)^l) = \delta_{lT}, \quad l \geq 1, \tag{3.22}$$

and T is the set of all odd positive integers.

If we extend the definition of ξ^l, to include $l = 0$, by requiring that $\xi^0 = 1$, Eq. (3.21) can be written in a more convenient form as

TABLE I. $C_p (p\leqslant 8)$ for the su(n, m), $d = n - m$.

K	P_k	C_2	C_3	C_4	C_5	C_6	C_7	C_8
2	P_2	1	$-d$	1/2	$-d/2$	3/16	$-3d/16$	1/16
3	P_3		1	$-d$	5/6	$-5d/6$	7/16	$-7d/16$
4	P_4			1	$-d$	5/4	$-5d/4$	7/8
5	P_5				1	$-d$	7/4	$-7d/4$
	P_2^2				$-1/2$	$d/2$	$-25/32$	$25d/32$
6	P_6					1	$-d$	7/3
	P_3P_2					-1	d	$-25/12$
7	P_7						1	$-d$
	P_4P_2						-1	d
	P_3^2						$-1/2$	$d/2$
8	P_8							1
	P_5P_2							-1
	P_4P_3							-1
	P_2^3							1/6

$$\ln F(z) = -\sum_{k\geqslant 1}^{\infty} z^{k+1} P_k \psi_k(z), \tag{3.23}$$

where $\psi_k(z)$ has the formal form

$$\psi_k(z) = \sum_{l=1}^{\infty} \binom{k+l}{k} \frac{\xi^{l-1}(z/2)^{l-1}}{k+l}$$
$$\equiv \frac{2}{kz}\left[\left(1-\xi\frac{z}{2}\right)^{-k} - 1\right]. \tag{3.24}$$

From Eq. (3.23),

$$F(z) = \prod_{k=1}^{\infty} \sum_{\nu_k=0}^{\infty} z^{(k+1)\nu_k} \frac{(-)^{\nu_k}}{\nu_k!} P_k^{\nu_k} [\psi_k(z)]^{\nu_k}. \tag{3.25}$$

Now let us introduce the abbreviations

$$\bar{\nu} = \sum_k \nu_k, \tag{3.26}$$

$$[\nu!] = \prod_k \nu_k!, \tag{3.27}$$

and let (ν) denote the set of all non-negative integers satisfying the constraint equation

$$K+1 = 2\nu_1 + 3\nu_2 + \cdots + (k+1)\nu_k, \tag{3.28}$$

for some integer K [$\nu_1 = 0$ for the su(n, m) and the osp $(N, 2m)$]. Equation (3.25) can be written in a form in which all the coefficients of a typical product term $P_1^{\nu_1} \cdots P_k^{\nu_k}$ are collected together:

$$F(z) = \sum_{l\geqslant 0} \sum_{(\nu)} \frac{(-)^{\bar{\nu}}}{[\nu!]} Q_l(\nu) P_1^{\nu_1} \cdots P_k^{\nu_k} z^{K+l+1}, \tag{3.29}$$

where $Q_l(\nu)$ is the rank-independent function satisfying

$$\sum_{l\geqslant 0} Q_l(\nu) z^l = [\Psi_1(z)]^{\nu_1} \cdots [\Psi_k(z)]^{\nu_k}. \tag{3.30}$$

The $Q_l(\nu)$'s play a universal role in this problem in the sense that for a well-defined set of power sums and for a given l and (ν), they have the same value for all the Lie superalgebras considered, and also for their corresponding Lie algebras. It is this remarkable property that enables considerable unification in the treatment of this aspect of the problem.

Equation (3.30) implies the following special values for the $Q_l(\nu)$'s:

$$Q_0(\nu) = 1, \quad \text{for all } (\nu), \tag{3.31}$$

$$Q_l(0) = Q_{-l}(\nu) = 0, \quad l \geqslant 1, \tag{3.32}$$

$$Q_{2r+1}(\nu) = 0, \quad r \geqslant 0. \tag{3.33}$$

TABLE II[a]. "Correction effects" for the $C_p (p\leqslant 8)$ for the u(n, m), $d = n - m$.

K	P_k	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
1	P_1	1	$-d$	1/4	$-d/4$	1/16	$-a/16$	1/64	$-d/64$
3	P_1^2			$-1/2$	$d/2$	$-3/8$	$3d/8$	$-5/32$	$5d/32$
4	P_2P_1				-1	d	$-9/8$	$9d/8$	$-5/8$
5	P_3P_1					-1	d	$-19/12$	$19d/12$
	P_1^3					1/6	$-d/6$	1/2	$-d/2$
6	P_4P_1						-1	d	$-17/8$
	$P_2P_1^2$						1/2	$-d/2$	1
7	P_5P_1							-1	d
	$P_3P_1^2$							1/2	$-d/2$
	$P_2^2P_1$							1/2	$-d/2$
	P_1^4							$-1/24$	$d/24$
8	P_6P_1								-1
	$P_4P_1^2$								1/2
	$P_3P_2P_1$								1
	$P_2P_1^3$								$-1/6$

[a] The C_p for the u(n, m) is obtained by adding the terms here for a given p to the corresponding terms in Table I for the su(n, m).

Finally, on substituting Eqs. (3.29) and (3.20) into Eqs. (3.3), (3.11), and (3.18) for the various cases, we obtain the desired expansion for the C_p,

$$C_p = -\sum_{(\nu)} \beta_p(\nu) P_1^{\nu_1} \cdots P_k^{\nu_k}, \quad p \geqslant 1. \tag{3.34}$$

In this equation, for the u(n, m),

$$\beta_p(\nu) = \frac{(-)^{\bar{\nu}}}{[\nu!]} (Q_l(\nu) - dQ_{l-1}(\nu)), \tag{3.35}$$

where

$$d = n - m, \quad l = p - K \geqslant 0, \quad p \geqslant K \geqslant 1$$
$$[p \geqslant K \geqslant 2 \text{ for the su}(n, m)], \tag{3.36}$$

and for the osp(N, $2m$)

$$\beta_p(\nu) = \frac{(-)^{\bar{\nu}}}{[\nu!]} \left[Q_l(\nu) - \left(d - \frac{1}{2}\right) Q_{l-1}(\nu) + \frac{1}{2} \sum_{t=0}^{l-2} Q_t(\nu) \left(\frac{d-1}{2}\right)^{l-t-1} \right] \tag{3.37}$$

with

$$d = N - 2m, \quad l = p - K \geqslant 0, \quad p \geqslant K \geqslant 2. \tag{3.38}$$

IV. RESULTS AND DISCUSSIONS

Further properties of the coefficients $Q_l(\nu)$ and $\beta_p(\nu)$ and illustrations are given in the appendices for more specific cases. Using these we compute the C_p for $p \leqslant 8$. The cases of the su(n, m) and the osp(N, $2m$) are further simplified by the fact that $P_1 = 0 = \nu_1$. Table I contains the C_p's for the su(n, m), and Table II the "correction effects" due to the u(n, m). In Table III we show the corresponding values for the osp(N, $2m$).

Our results agree with those of Scheunert[6] when expressed in terms of the power sums

$$Q_k = \sum_{i=1}^{n+m} \sigma_i (l_i^k - r_i^k), \tag{4.1}$$

using the relation

$$P_k = (2) \sum_{l=1}^{k} \binom{k}{l} \left(\frac{d}{2}\right)^{k-l} Q_k \tag{4.2}$$

for the u(n, m) and the osp(N, $2m$), the factor of 2 being applicable only in the case of the osp(N, $2m$). In Eq. (4.1),

$$l_i = \sigma_i f_i + r_i. \tag{4.3}$$

In the nongraded limit in which $\sigma_i = 1$, and d is replaced by $n(N)$ for the u(n) [o(N)]

$$P_k \to P_k^0 = \sum_i \left[\left(\lambda_i^0 + \frac{1}{2}\right)^k - \left(\rho_i^0 + \frac{1}{2}\right)^k \right], \tag{4.4}$$

where λ_i^0 and ρ_i^0 are the corresponding values of λ_i and ρ_i in this limit. We recover the results of Popov for the u(n) and those of Ref. (15) for the o(N), by using the relations

$$P_k^0 = \sum_{l=1}^{k} \binom{k}{l} \left(\frac{1}{2}\right)^{k-l} S_l, \tag{4.5}$$

$$S_l = \sum_i (\lambda_i^{0l} - \rho_i^{0l}), \quad S_1 = 0, \quad \text{for su}(n), \text{o}(N). \tag{4.6}$$

This is a direct proof of the *Ansatz* of Ref. 5 for the u(n) and the o(N).

TABLE III. $C_p (p < 8)$ for the osp(N, $2m$), $d = N - 2m$.

K	P_k	C_2	C_3	C_4	C_5	C_6	C_7	C_8
2	P_2	1	$\frac{1}{2} - d$	$\frac{1}{4}(1 + d)$	$\frac{1}{8}(3 - 6d + d^2)$	$\frac{1}{16}(13 + 5d - 3d^2 + d^3)$	$+\frac{1}{32}(6 - 14d + 4d^2 - 4d^3 + d^4)$	$\frac{1}{64}(-2 + 14d - 16d^2 + 12d^3 - 5d^4 + d^5)$
3	P_3		1	$\frac{1}{2} - d$	$\frac{1}{4}(\frac{7}{3} + d)$	$\frac{1}{8}(\frac{13}{3} - \frac{26}{3}d + d^2)$	$+\frac{1}{16}(\frac{8}{3} + \frac{19}{3}d - \frac{9}{3}d^2 + d^3)$	$\frac{1}{32}(\frac{34}{3} - \frac{74}{3}d + \frac{25}{3}d^2 - 4d^3 + d^4)$
4	P_4			1	$\frac{1}{2} - d$	$\frac{1}{4}(4 + d)$	$\frac{1}{8}(6 - 7d + d^2)$	$\frac{1}{16}(8 + 8d - 3d^2 + d^3)$
5	P_5				1	$\frac{1}{2} - d$	$\frac{1}{4}(6 + d)$	$\frac{1}{8}(8 - 16d + d^2)$
	P_2^2				$-\frac{1}{2}$	$-\frac{1}{2}(\frac{1}{2} - d)$	$-\frac{1}{8}(\frac{21}{4} + d)$	$-\frac{1}{8}(\frac{21}{4} - \frac{29}{2}d + d^2)$
6	P_6					1	$\frac{1}{2} - d$	$\frac{1}{4}(\frac{25}{3} + d)$
	P_3P_2					-1	$-(\frac{1}{2} - d)$	$-\frac{1}{4}(\frac{22}{3} + d)$
7	P_7						1	$\frac{1}{2} - d$
	P_4P_2						-1	$-(\frac{1}{2} - d)$
	P_3^2						$-\frac{1}{2}$	$-\frac{1}{2}(\frac{1}{2} - d)$
8	P_8							1
	P_5P_2							-1
	P_4P_3							-1
	P_2^3							$\frac{1}{6}$

If we replace σ_i by -1, λ_i by $-\lambda_i$, and d by $-2m$, we obtain for the case of the sp(2m),

$$P_k \to (-)^{k+1} \sum_{l=2}^{k} \binom{k}{l} \left(\frac{1}{2}\right)^{k-l} S_l, \tag{4.7}$$

according to which

$$C_p^{\mathrm{osp}(2n,\,2m)} \underset{n\to 0}{\longrightarrow} C_p^0 = (-)^{p+1} C_p^N, \tag{4.8}$$

where C_p^N is the eigenvalue of the Casimir operator of order p for the sp(2m) in Table II of Ref. 15. This provides the equivalent *Ansatz* for obtaining the osp(2n, 2m) results from those of the sp(2m) and vice versa: multiply the C_p of sp(2m) by $(-)^{p+1}$ and replace $2m$ by $d = 2n - 2m$.

Generally, when compared with earlier results, our results show that the C_p's appear in their simplest form revealing an interesting pattern when expressed in terms of the graded power sums P_k, or equivalently P_k^0 for the corresponding Lie algebras, a fact hardly apparent if one confined one's investigations to the Lie algebras and blindly applied existing *Ansätze*. Also in terms of the P_k or P_k^0 half the coefficients $Q_l(\nu)$ appearing in the main expansion formula vanish resulting in considerable time savings. These power sums could play a more significant role in the detailed study of the general pattern of these coefficients.

ACKNOWLEDGMENTS

One of the authors (C.O.N.) would like to thank Professor Abdus Salam, the International Atomic Energy Agency, and the United Nations Educational, Scientific, and Cultural Organization (UNESCO) for hospitality at the International Centre for Theoretical Physics, Trieste.

APPENDIX A: THE COEFFICIENTS $Q_l(\nu)$

More specific forms for the coefficients $Q_l(\nu)$ can be established using the relation

$$\begin{aligned}\Psi_{k_1}(z)\cdots\Psi_{k_t}(z) = \frac{(2|z|)^{t-1}}{k_1\cdots k_t}\{&(k_1+k_2+\cdots+k_t)\Psi_{k_1+k_2+\cdots+k_t}(z) - [(k_1+\cdots+k_{t-1})\Psi_{k_1+\cdots+k_{t-1}}(z)\\ &-(k_1+\cdots+k_{t-2})\Psi^{(z)}_{k_1+\cdots k_{t-2}} + \cdots + (-)^t k_1\Psi_{k_1}(z) + \text{com. } k_1, k_2,\ldots,k_t\,]\},\end{aligned} \tag{A1}$$

where com. means all possible combinations of $k_1,\ldots,k_t$ in all the terms appearing in the square bracket [ ]. This relation can readily be established by induction on t using the defining equation (3.24).

Now let $Q_l(P_{k_1}\cdots P_{k_t})$ denote $Q_l(\nu)$ for the particular set (ν) which satisifes the constraint equation (3.28) for the values $\nu_{k_1} = \nu_{k_2} = \cdots = \nu_{k_t} = 1$, all other ν's being equal to zero. Then using Eqs. (3.30) to (3.33) in (A1), we obtain, for $l = 2r$ (r positive integer),

$$\begin{aligned}Q_{2r}(P_{k_1}\cdots P_{k_t}) = \frac{2^{-2r}}{k_1\cdots k_t}\Bigg\{&\binom{k_1+\cdots+k_t+2r+t-1}{2r+t} - \Bigg[\binom{k_1+\cdots+k_{t-1}+2r+t-1}{2r+t}\\ &-\binom{k_1+\cdots+k_{t-2}+2r+t-1}{2r+t} + \cdots + (-)^t\binom{k_1+2r+t-1}{2r+t} + \text{com. } k_1,\ldots,k_t\Bigg]\Bigg\}.\end{aligned} \tag{A2}$$

Equation (A2) determines the values of all the $Q_l(\nu)$ needed for the calculation of $\beta_p(\nu)$. We present three examples.

(i) For $t = 1$, set $k_1 = k$,

$$Q_{2r}(P_k) = \frac{2^{-2r}}{k}\binom{k+2r}{2r+1}. \tag{A3}$$

(ii) For $t = 2$,

$$Q_{2r}(P_{k_1}P_{k_2}) = \frac{2^{-2r}}{k_1k_2}\left\{\binom{k_1+k_2+2r+1}{2r+2} - \binom{k_1+2r+1}{2r+2} - \binom{k_2+2r+1}{2r+2}\right\}. \tag{A4}$$

If, in Eq. (A4), $k_1 = k_2 = k$, we have

$$Q_{2r}(P_k^2) = \frac{2^{-2r}}{k^2}\left\{\binom{2k+2r+1}{2r+2} - 2\binom{k+2r+1}{2r+2}\right\}. \tag{A5}$$

(iii) For $t = 3$,

$$\begin{aligned}Q_{2r}(P_{k_1}P_{k_2}P_{k_3}) = \frac{2^{-2r}}{k_1k_2k_3}\Bigg\{&\binom{k_1+k_2+k_3+2r+2}{2r+3} - \binom{k_1+k_2+2r+2}{2r+3} - \binom{k_1+k_3+2r+2}{2r+3}\\ &-\binom{k_2+k_3+2r+2}{2r+3} + \binom{k_1+2r+2}{2r+3} + \binom{k_2+2r+2}{2r+3} + \binom{k_3+2r+2}{2r+3}\Bigg\},\end{aligned} \tag{A6}$$

with quantities $Q_{2r}(P_{k_1}P_k^2)$ and $Q_{2r}(P_k^3)$ obtained by setting $k_2 = k_3 = k$, and $k_1 = k_2 = k_3 = k$, respectively, in this equation.

APPENDIX B: THE COEFFICIENTS $\beta_p(\nu)$

These are determined directly from Eqs. (3.35) and (3.37). It is, however, sometimes convenient to use simpler forms derived from these equations as illustrated below for the case of u(n, m).

(i) The $\beta_p(P_k)$: The sum P_k corresponds to the partition (ν) for which $\nu_k = 1$, $\nu_i = 0$, $i \neq k$, which satisfies the constraint equation (3.28) for $K = k$. Thus, $l = p - k \geqslant 0$, and from Eq. (3.25)

$$\beta_p(P_k) = \begin{cases} \dfrac{2^{k-p}}{k}\dbinom{p}{k-1}, & p-k \text{ even}, \quad \text{(B1a)} \\ \dfrac{2^{k-p+1}d}{k}\dbinom{p-1}{k-1}, & p-k \text{ odd}. \quad \text{(B1b)} \end{cases}$$

(ii) The $\beta_p(P_{k_1}P_{k_2})$: The product sum $P_{k_1}P_{k_2}$ occurs when $\nu_{k_1} = \nu_{k_2} = 1$, $\nu_i = 0$, $i \neq k_1, k_2$. In this case $K = k_1 + k_2 + 1$, and $l = p - (k_1 + k_2 + 1) \geqslant 0$ so that

$$\beta_p(P_{k_1}P_{k_2}) = \frac{2^{k_1+k_2-p+1}}{k_1k_2}\left[\binom{p}{k_1+k_2-1} - \binom{p-k_2}{k_1-1} - \binom{p-k_1}{k_2-1}\right], \quad p-k_1-k_2-1 \text{ even}, \tag{B2a}$$

$$= \frac{2^{k_1+k_2-p+2}d}{k_1k_2}\left[\binom{p-k_2-1}{k_1-1} + \binom{p-k_1-1}{k_2-1} - \binom{p-1}{k_1+k_2-1}\right], \quad p-k_1-k_2-1 \text{ odd}. \tag{B2b}$$

From Eqs. (B2) $\beta_p(P_k^2)$ is obtained by setting $k_1 = k_2 = k$ and dividing through by 2!; and so on for higher products.

[1]P. D. Jarvis and H. S. Green, J. Math. Phys. **20**, 2115 (1979).
[2]B. Balantekin and I. Bars, J. Math. Phys. **22**, 1149 (1981).
[3]M. Bednar and V. Sachl, J. Math. Phys. **19**, 1487 (1978).
[4]For the construction of generalized Casimir operators for the classical Lie algebras see, for example, G. Racah, reprinted in Springer Tracts Mod. Phys. **37**, 28 (1965); H. B. G. Casimir, Proc. Roy. Acad. Amsterdam **34**, 844 (1931); L. C. Biedenharn and J. D. Louck, Commun. Math. Phys. **8**, 89 (1968).
[5]B. Balantekin, J. Math. Phys. **24**, 486 (1982); see also Refs. 1–3.
[6]M. Scheunert, J. Math. Phys. **24**, 2681 (1983).
[7]A. M. Bincer, J. Math. Phys. **24**, 2546 (1983).
[8]A. M. Perelomov and V. S. Popov, J. Nucl. Phys. USSR **3**, 924 (1966); **3**, 1127 (1966).
[9]A. M. Perelomov and V. S. Popov, J. Nucl. Phys. USSR **5**, 693 (1967).
[10]V. S. Popov, Theor. Mat. Fiz. **29**, 357 (1976) [Theor. Math. Phys. **29**, 1122 (1976) (translated August 1977)].
[11]J. D. Louck and L. C. Biedenharn, J. Math. Phys. **11**, 2368 (1970).
[12]S. Okubo, J. Math. Phys. **16**, 528 (1975).
[13]C. O. Nwachuku and M. A. Rashid, J. Math. Phys. **17**, 1611 (1976).
[14]C. O. Nwachuku and M. A. Rashid, J. Math. Phys. **18**, 1387 (1977).
[15]C. O. Nwachuku, J. Math. Phys. **20**, 1260 (1979).
[16]V. Rittenberg, "A guide to Lie superalgebras," invited talk at the VI International Colloquium on Group Theoretical Methods in Physics, Tübingen, 18–22 July 1977.

The symmetric and antisymmetric structure constants for SU(6)

Uri Sarid
Department of Physics, University of Arizona, Tucson, Arizona 85721

(Received 25 July 1984; accepted for publication 4 April 1985)

The 124 completely antisymmetric f_{ijk} and 221 completely symmetric d_{ijk} (nonzero) structure constants for a simple representation of SU(6) are tabulated. The basis matrices λ_i used to generate the structure constants are also given.

I. INTRODUCTION

The structure constants proposed for SU(3) by Gell-Mann[1] in 1961 and later[2] extended to SU(4) are generalized herein to SU(6). With the recent probable identification[3] of bound states of the sixth quark t at CERN's UA1 detector, the significance of SU(6) is now evident. The symmetric d_{ijk} and antisymmetric f_{ijk} structure constants, which determine the Lie algebra associated with this group, are used as fundamental couplings in a variety of ways. Working with the d's and f's usually eliminates the need for Clebsch–Gordan coefficients and with them the difficulties due to phase conventions. As a prominent example, the charge operator acts on any hadron state $|H^j\rangle$ through $Q^i|H^j\rangle = if^{ijk}|H^k\rangle$, and the commutator of two charge operators is given by $[Q_i,Q_j] = if_{ijk}Q_k$, this latter property generalizing to the commutation relations with the other components of the current operator J_j^μ. The structure constants also figure prominently in the quark energy densities $u^i = \bar{q}\lambda^i q$ for a scalar and $v^i = \bar{q}\lambda^i\gamma_5 q$ for a pseudoscalar, which therefore obey commutation relations with the charge operator $[Q^i,u^j] = if^{ijk}u^k$, $[Q_5^i,v^j] = id^{ijk}u^k$, and so on.

II. THEORY

We choose as the representation the equivalent of the familiar Pauli matrices τ_i extended to six dimensions and labeled λ_i. Typically the columns and rows are labeled by the six quark flavors u-d-s-c-b-t in order of increasing mass, and we assume this convention unless otherwise mentioned. But since the f_{ijk} and d_{ijk} are derived strictly from the λ_i without regard for the row and column labeling, one need only make sure to use the proper indices i,j,k from the λ_i corresponding to the particular application.

To generate the λ_i we begin with λ_1, λ_2, and λ_3 as the τ_1, τ_2, and τ_3 matrices, respectively, in the $\{u,d\}$ subspace. The normalized identity matrix λ_0 is included from the extension to U(6) via U(1)$\times$SU(6). The τ_1 and τ_2 in the $\{u,s\}$ and $\{d,s\}$ subspaces are next, followed by the diagonal counterpart of τ_3 in the $\{u,d,s\}$ subspace. Thus each quark flavor is added until the 36 λ_i, for $i = 0,1,\ldots,35$, are generated. Note that appropriate normalization factors are needed for the diagonal matrices to preserve the relations

$$\operatorname{tr}\lambda_i\lambda_j = 2\delta_{ij}. \tag{1}$$

The λ_i matrices are listed more compactly below by either designating the subspace in which they take the form of a τ matrix, or (if diagonal) by explicitly writing down $\bar{q}\lambda_i q$. We label their rows by the basis row vectors $\bar{u}, \bar{d}, \bar{s}, \bar{c}, \bar{b}$, and $\bar{t}$ and their columns by the basis column vectors u, d, s, c, b, and t to comply with the standard quark assignments. Thus the first two matrices conventionally written

$$\lambda_0 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\lambda_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

appear as $\bar{q}\lambda_0 q = (1/\sqrt{3})(\bar{u}u + \bar{d}d + \bar{s}s + \bar{c}c + \bar{b}b + \bar{t}t)$ and $\lambda_1 = \tau_1(u,d)$. The λ_i used to generate the SU(6) structure constants are

$$\bar{q}\lambda_0 q = (1/\sqrt{3})(\bar{u}u + \bar{d}d + \bar{s}s + \bar{c}c + \bar{b}b + \bar{t}t),$$
$$\lambda_1 = \tau_1(u,d), \quad \lambda_2 = \tau_2(u,d),$$
$$\bar{q}\lambda_3 q = \bar{u}u - \bar{d}d,$$
$$\lambda_4 = \tau_1(u,s), \quad \lambda_5 = \tau_2(u,s),$$
$$\lambda_6 = \tau_1(d,s), \quad \lambda_7 = \tau_2(d,s),$$
$$\bar{q}\lambda_8 q = (1/\sqrt{3})(\bar{u}u + \bar{d}d - 2\bar{s}s),$$
$$\lambda_9 = \tau_1(u,c), \quad \lambda_{10} = \tau_2(u,c),$$
$$\lambda_{11} = \tau_1(d,c), \quad \lambda_{12} = \tau_2(d,c),$$
$$\lambda_{13} = \tau_1(s,c), \quad \lambda_{14} = \tau_2(s,c),$$
$$\bar{q}\lambda_{15} q = (1/\sqrt{6})(\bar{u}u + \bar{d}d + \bar{s}s - 3\bar{c}c),$$
$$\lambda_{16} = \tau_1(u,b), \quad \lambda_{17} = \tau_2(u,b),$$
$$\lambda_{18} = \tau_1(d,b), \quad \lambda_{19} = \tau_2(d,b),$$
$$\lambda_{20} = \tau_1(s,b), \quad \lambda_{21} = \tau_2(s,b),$$
$$\lambda_{22} = \tau_1(c,b), \quad \lambda_{23} = \tau_2(c,b),$$
$$\bar{q}\lambda_{24} q = (1/\sqrt{10})(\bar{u}u + \bar{d}d + \bar{s}s + \bar{c}c - 4\bar{b}b),$$
$$\lambda_{25} = \tau_1(u,t), \quad \lambda_{26} = \tau_2(u,t),$$
$$\lambda_{27} = \tau_1(d,t), \quad \lambda_{28} = \tau_2(d,t),$$
$$\lambda_{29} = \tau_1(s,t), \quad \lambda_{30} = \tau_2(s,t),$$
$$\lambda_{31} = \tau_1(c,t), \quad \lambda_{32} = \tau_2(c,t),$$
$$\lambda_{33} = \tau_1(b,t), \quad \lambda_{34} = \tau_2(b,t),$$
$$\bar{q}\lambda_{35} q = (1/\sqrt{15})(\bar{u}u + \bar{d}d + \bar{s}s + \bar{c}c + \bar{b}b - 5\bar{t}t).$$

Since the λ_i for SU(n) contain the λ_i for SU(m) where $m<n$, the structure constants f_{ijk} and d_{ijk} tabulated below may be

 0022-2488/85/081921-04$02.50

TABLE I. The symmetric structure constants for SU(6).

$i\ j\ k$	d_{ijk}	$i\ j\ k$	d_{ijk}	$i\ j\ k$	d_{ijk}	$i\ j\ k$	d_{ijk}
0 j k	$1/\sqrt{3}\delta_{jk}$						
1 1 8	$1/\sqrt{3}$	5 5 15	$1/\sqrt{6}$	9 17 23	$1/2$	15 32 32	$-3/2\sqrt{6}$
1 1 15	$1/\sqrt{6}$	5 5 24	$1/\sqrt{10}$	9 25 31	$1/2$	16 16 24	$-3/2\sqrt{10}$
1 1 24	$1/\sqrt{10}$	5 5 35	$1/\sqrt{15}$	9 26 32	$1/2$	16 16 35	$1/\sqrt{15}$
1 1 35	$1/\sqrt{15}$	5 9 14	$-1/2$	10 10 15	$-1/\sqrt{6}$	16 25 33	$1/2$
1 4 6	$1/2$	5 10 13	$1/2$	10 10 24	$1/\sqrt{10}$	16 26 34	$1/2$
1 5 7	$1/2$	5 16 21	$-1/2$	10 10 35	$1/\sqrt{15}$	17 17 24	$-3/2\sqrt{10}$
1 9 11	$1/2$	5 17 20	$1/2$	10 16 23	$-1/2$	17 17 35	$1/\sqrt{15}$
1 10 12	$1/2$	5 25 30	$-1/2$	10 17 22	$1/2$	17 25 34	$-1/2$
1 16 18	$1/2$	5 26 29	$1/2$	10 25 32	$-1/2$	17 26 33	$1/2$
1 17 19	$1/2$	6 6 8	$-1/2\sqrt{3}$	10 26 31	$1/2$	18 18 24	$-3/2\sqrt{10}$
1 25 27	$1/2$	6 6 15	$1/\sqrt{6}$	11 11 15	$-1/\sqrt{6}$	18 18 35	$1/\sqrt{15}$
1 26 28	$1/2$	6 6 24	$1/\sqrt{10}$	11 11 24	$1/\sqrt{10}$	18 27 33	$1/2$
2 2 8	$1/\sqrt{3}$	6 6 35	$1/\sqrt{15}$	11 11 35	$1/\sqrt{15}$	18 28 34	$1/2$
2 2 15	$1/\sqrt{6}$	6 11 13	$1/2$	11 18 22	$1/2$	19 19 24	$-3/2\sqrt{10}$
2 2 24	$1/\sqrt{10}$	6 12 14	$1/2$	11 19 23	$1/2$	19 19 35	$1/\sqrt{15}$
2 2 35	$1/\sqrt{15}$	6 18 20	$1/2$	11 27 31	$1/2$	19 27 34	$-1/2$
2 4 7	$-1/2$	6 19 21	$1/2$	11 28 32	$1/2$	19 28 33	$1/2$
2 5 6	$1/2$	6 27 29	$1/2$	12 12 15	$-1/\sqrt{6}$	20 20 24	$-3/2\sqrt{10}$
2 9 12	$-1/2$	6 28 30	$1/2$	12 12 24	$1/\sqrt{10}$	20 20 35	$1/\sqrt{15}$
2 10 11	$1/2$	7 7 8	$-1/2\sqrt{3}$	12 18 23	$-1/2$	20 29 33	$1/2$
2 16 19	$-1/2$	7 7 15	$1/\sqrt{6}$	12 19 22	$1/2$	20 30 34	$1/2$
2 17 18	$1/2$	7 7 24	$1/\sqrt{10}$	12 27 32	$-1/2$	21 21 24	$-3/2\sqrt{10}$
2 25 28	$-1/2$	7 7 35	$1/\sqrt{15}$	12 28 31	$1/2$	21 21 35	$1/\sqrt{15}$
2 26 27	$1/2$	7 11 14	$-1/2$	13 13 15	$-1/\sqrt{6}$	21 29 34	$-1/2$
3 3 8	$1/\sqrt{3}$	7 12 13	$1/2$	13 13 24	$1/\sqrt{10}$	21 30 33	$1/2$
3 3 15	$1/\sqrt{6}$	7 18 21	$-1/2$	13 13 35	$1/\sqrt{15}$	22 22 24	$-3/2\sqrt{10}$
3 3 24	$1/\sqrt{10}$	7 19 20	$1/2$	13 20 22	$1/2$	22 22 35	$1/\sqrt{15}$
3 3 35	$1/\sqrt{15}$	7 27 30	$-1/2$	13 21 23	$1/2$	22 31 33	$1/2$
3 4 4	$1/2$	7 28 29	$1/2$	13 29 31	$1/2$	22 32 34	$1/2$
3 5 5	$1/2$	8 8 8	$-1/\sqrt{3}$	13 30 32	$1/2$	23 23 24	$-3/2\sqrt{10}$
3 6 6	$-1/2$	8 8 15	$1/\sqrt{6}$	14 14 15	$-1/\sqrt{6}$	23 23 35	$1/\sqrt{15}$
3 7 7	$-1/2$	8 8 24	$1/\sqrt{10}$	14 14 24	$1/\sqrt{10}$	23 31 34	$-1/2$
3 9 9	$1/2$	8 8 35	$1/\sqrt{15}$	14 14 35	$1/\sqrt{15}$	23 32 33	$1/2$
3 10 10	$1/2$	8 9 9	$1/2\sqrt{3}$	14 20 23	$-1/2$	24 24 24	$-3/\sqrt{10}$
3 11 11	$-1/2$	8 10 10	$1/2\sqrt{3}$	14 21 22	$1/2$	24 24 35	$1/\sqrt{15}$
3 12 12	$-1/2$	8 11 11	$1/2\sqrt{3}$	14 29 32	$-1/2$	24 25 25	$1/2\sqrt{10}$
3 16 16	$1/2$	8 12 12	$1/2\sqrt{3}$	14 30 31	$1/2$	24 26 26	$1/2\sqrt{10}$
3 17 17	$1/2$	8 13 13	$-1/\sqrt{3}$	15 15 15	$-2/\sqrt{6}$	24 27 27	$1/2\sqrt{10}$
3 18 18	$-1/2$	8 14 14	$-1/\sqrt{3}$	15 15 24	$1/\sqrt{10}$	24 28 28	$1/2\sqrt{10}$
3 19 19	$-1/2$	8 16 16	$1/2\sqrt{3}$	15 15 35	$1/\sqrt{15}$	24 29 29	$1/2\sqrt{10}$
3 25 25	$1/2$	8 17 17	$1/2\sqrt{3}$	15 16 16	$1/2\sqrt{6}$	24 30 30	$1/2\sqrt{10}$
3 26 26	$1/2$	8 18 18	$1/2\sqrt{3}$	15 17 17	$1/2\sqrt{6}$	24 31 31	$1/2\sqrt{10}$
3 27 27	$-1/2$	8 19 19	$1/2\sqrt{3}$	15 18 18	$1/2\sqrt{6}$	24 32 32	$1/2\sqrt{10}$
3 28 28	$-1/2$	8 20 20	$-1/\sqrt{3}$	15 19 19	$1/2\sqrt{6}$	24 33 33	$-2/\sqrt{10}$
4 4 8	$-1/2\sqrt{3}$	8 21 21	$-1/\sqrt{3}$	15 20 20	$1/2\sqrt{6}$	24 34 34	$-2/\sqrt{10}$
4 4 15	$1/\sqrt{6}$	8 25 25	$1/2\sqrt{3}$	15 21 21	$1/2\sqrt{6}$	25 25 35	$-2/\sqrt{15}$
4 4 24	$1/\sqrt{10}$	8 26 26	$1/2\sqrt{3}$	15 22 22	$-3/2\sqrt{6}$	26 26 35	$-2/\sqrt{15}$
4 4 35	$1/\sqrt{15}$	8 27 27	$1/2\sqrt{3}$	15 23 23	$-3/2\sqrt{6}$	27 27 35	$-2/\sqrt{15}$
4 9 13	$1/2$	8 28 28	$1/2\sqrt{3}$	15 25 25	$1/2\sqrt{6}$	28 28 35	$-2/\sqrt{15}$
4 10 14	$1/2$	8 29 29	$-1/\sqrt{3}$	15 26 26	$1/2\sqrt{6}$	29 29 35	$-2/\sqrt{15}$
4 16 20	$1/2$	8 30 30	$-1/\sqrt{3}$	15 27 27	$1/2\sqrt{6}$	30 30 35	$-2/\sqrt{15}$
4 17 21	$1/2$	9 9 15	$-1/\sqrt{6}$	15 28 28	$1/2\sqrt{6}$	31 31 35	$-2/\sqrt{15}$
4 25 29	$1/2$	9 9 24	$1/\sqrt{10}$	15 29 29	$1/2\sqrt{6}$	32 32 35	$-2/\sqrt{15}$
4 26 30	$1/2$	9 9 35	$1/\sqrt{15}$	15 30 30	$1/2\sqrt{6}$	33 33 35	$-2/\sqrt{15}$
5 5 8	$-1/2\sqrt{3}$	9 16 22	$1/2$	15 31 31	$-3/2\sqrt{6}$	34 34 35	$-2/\sqrt{15}$

applied to SU(m) by simply restricting the indices $i,j,k \leq m^2-1$. From the defining commutation relation $[\lambda_i, \lambda_j] = 2if_{ijk}\lambda_k$ we solve for f_{ijk} by multiplying by λ_k, and taking the trace, using (1) to simplify the resulting expression. This immediately shows

$$f_{ijk} = -(i/4)\mathrm{tr}([\lambda_i,\lambda_j]\lambda_{ik}). \qquad (2)$$

Similarly, by using the defining anticommutation relation $\{\lambda_i,\lambda_j\} = 2d_{ijk}\lambda_k$ we arrive at

$$d_{ijk} = \tfrac{1}{4}\mathrm{tr}(\{\lambda_i,\lambda_j\}\lambda_k). \qquad (3)$$

Note that, as for example in Gell-Mann's original formulation[1] for SU(3), the anticommutator may be given as $\{\lambda_i,\lambda_j\} = \frac{4}{3}\delta_{ij} + 2d_{ijk}\lambda_k$, $i,j,k = 1,...,8$. Then for SU(6) the generalization is $\{\lambda_i,\lambda_j\} = \frac{4}{6}\delta_{ij} + 2d_{ijk}\lambda_k$, $i,j,k = 1,...,35$. If we let i,j, and k run from 0 to 35 as for U(6), the δ_{ij} term can be

TABLE II. The antisymmetric structure constants for SU(6).

$i\ j\ k$	f_{ijk}	$i\ j\ k$	f_{ijk}	$i\ j\ k$	f_{ijk}
1 2 3	1	8 16 17	$1/2\sqrt{3}$	21 30 34	$1/2$
1 4 7	$1/2$	8 18 19	$1/2\sqrt{3}$	22 23 24	$\sqrt{10}/4$
1 5 6	$-1/2$	8 20 21	$-1/\sqrt{3}$	22 31 34	$1/2$
1 9 12	$1/2$	8 25 26	$1/2\sqrt{3}$	22 32 33	$-1/2$
		8 27 28	$1/2\sqrt{3}$	23 31 33	$1/2$
1 10 11	$-1/2$	8 29 30	$-1/\sqrt{3}$	23 32 34	$1/2$
1 16 19	$1/2$	9 10 15	$\sqrt{2}/\sqrt{3}$	24 25 26	$1/2\sqrt{10}$
1 17 18	$-1/2$	9 16 23	$1/2$	24 27 28	$1/2\sqrt{10}$
1 25 28	$1/2$	9 17 22	$-1/2$	24 29 30	$1/2\sqrt{10}$
1 26 27	$-1/2$	9 25 32	$1/2$	24 31 32	$1/2\sqrt{10}$
2 4 6	$1/2$	9 26 31	$-1/2$	24 33 34	$-2/\sqrt{10}$
2 5 7	$1/2$	10 16 22	$1/2$	25 26 35	$3/\sqrt{15}$
2 9 11	$1/2$	10 17 23	$1/2$	27 28 35	$3/\sqrt{15}$
2 10 12	$1/2$	10 25 31	$1/2$	29 30 35	$3/\sqrt{15}$
2 16 18	$1/2$	10 26 32	$1/2$	31 32 35	$3/\sqrt{15}$
2 17 19	$1/2$	11 12 15	$\sqrt{2}/\sqrt{3}$		
2 25 27	$1/2$	11 18 23	$1/2$		
2 26 28	$1/2$	11 19 22	$-1/2$		
3 4 5	$1/2$	11 27 32	$1/2$		
3 6 7	$-1/2$	11 28 31	$-1/2$		
3 9 10	$1/2$	12 18 22	$1/2$		
3 11 12	$-1/2$	12 19 23	$1/2$		
3 16 17	$1/2$	12 27 31	$1/2$		
3 18 19	$-1/2$	12 28 32	$1/2$		
3 25 26	$1/2$	13 14 15	$\sqrt{2}/\sqrt{3}$		
3 27 28	$-1/2$	13 20 23	$1/2$		
4 5 8	$\sqrt{3}/2$	13 21 22	$-1/2$		
4 9 14	$1/2$	13 29 32	$1/2$		
4 10 13	$-1/2$	13 30 31	$-1/2$		
4 16 21	$1/2$	14 20 22	$1/2$		
4 17 20	$-1/2$	14 21 23	$1/2$		
4 25 30	$1/2$	14 29 31	$1/2$		
4 26 29	$-1/2$	14 30 32	$1/2$		
5 9 13	$1/2$	15 16 17	$1/2\sqrt{6}$		
5 10 14	$1/2$	15 18 19	$1/2\sqrt{6}$		
5 16 20	$1/2$	15 20 21	$1/2\sqrt{6}$		
5 17 21	$1/2$	15 22 23	$-3/2\sqrt{6}$		
5 25 29	$1/2$	15 25 26	$1/2\sqrt{6}$		
5 26 30	$1/2$	15 27 28	$1/2\sqrt{6}$		
6 7 8	$\sqrt{3}/2$	15 29 30	$1/2\sqrt{6}$		
6 11 14	$1/2$	15 31 32	$-3/2\sqrt{6}$		
6 12 13	$-1/2$	16 17 24	$\sqrt{10}/4$		
6 18 21	$1/2$	16 25 34	$1/2$		
6 19 20	$-1/2$	16 26 33	$-1/2$		
6 27 30	$1/2$	17 25 33	$1/2$		
6 28 29	$-1/2$	17 26 34	$1/2$		
7 11 13	$1/2$	18 19 24	$\sqrt{10}/4$		
7 12 14	$1/2$	18 27 34	$1/2$		
7 18 20	$1/2$	18 28 33	$-1/2$		
7 19 21	$1/2$	19 27 33	$1/2$		
7 27 29	$1/2$	19 28 34	$1/2$		
7 28 30	$1/2$	20 21 24	$\sqrt{10}/4$		
8 9 10	$1/2\sqrt{3}$	20 29 34	$1/2$		
8 11 12	$1/2\sqrt{3}$	20 30 33	$-1/2$		
8 13 14	$-1/\sqrt{3}$	21 29 33	$1/2$		

dropped.[4] The d_{ijk} are thus shown for all of U(6)—for i, j, or $k = 0$, we simply get from (3) and (1) that $d_{0jk} = (1/\sqrt{3})\delta_{jk}$, the $1/\sqrt{3}$ coming from the normalization factor for λ_0. The f_{ijk}, of course, will not involve λ_0. Due to the complete symmetry of the d's and complete antisymmetry of the f's, it suffices to consider only those cases where $i \leq j \leq k$ for d_{ijk} and $i < j < k$ for f_{ijk}. When individual f_{ijk} and d_{ijk} are computed by hand, it is usually easier to simply multiply out the left-hand sides of the defining relations and arrive at the structure constants by inspection; however, for computer calculations, (2) and (3) are easier to use despite being more time consuming. Equations (2) and (3) may be directly programmed even on a microcomputer to generate the structure constants for higher dimensions, for example, using the language FORTH as was done for the Tables I and II.

III. CONCLUSION

We have computed the nonzero symmetric and antisymmetric structure constants d_{ijk} and f_{ijk} for the group SU(6). The generating basis matrices λ_i are listed, using the simple representation of Gell-Mann generalized from the usual representation of the τ_i SU(2) matrices.

[1]M. Gell-Mann, "The eightfold way: A theory of strong interaction symmetry," California Institute of Technology Synchrotron Laboratory Report CTSL-20 (1961), unpublished; M. Gell-Mann and Y. Ne'eman, *The Eightfold Way* (Benjamin, New York, 1964).
[2]Z. Maki, T. Maskawa, and I. Umemura, "Quartet scheme of hadrons in chiral U(4)×U(4)," Prog. Theor. Phys. **47**, 1682 (1972).
[3]UA1 Collaboration, presented at the XXII International Conference on High Energy Physics, Leipzig, July 1984.
[4]M. Gell-Mann, "Symmetries of baryons and mesons," Phys. Rev. **125**, 1067 (1962).

Canonical map approach to channeling stability in crystals

A. W. Sáenz
Naval Research Laboratory, Washington, D.C. 20375 and Physics Department, Catholic University, Washington, D.C. 20064

(Received 21 March 1984; accepted for publication 21 December 1984)

We state and prove rigorous mathematical results on the orbital stability of certain rectilinear trajectories of sufficiently energetic particles subjected to appropriate periodic potentials. This is done in the context of nontrivial classical Hamiltonian models, nonrelativistic and relativistic, in two space dimensions. The main steps involved in the proofs are the derivation of the asymptotic form of certain canonical maps in the plane in the limit of large particle energies and the application of a version of Moser's twist theorem. When suitably specialized, these results establish rigorously for the first time that the pertinent straight-line channeling trajectories of fast particles in two-dimensional rigid crystal lattices have this stability property under reasonable conditions on the crystal potential.

I. INTRODUCTION

Consider the following question of classical mechanics. Let a fast positively charged particle travel along the midline between two lines of atoms of a rectangular atomic array in the plane. Is this rectilinear motion stable in some sense under reasonable assumptions on the repulsive interactions between the fast particle and the atoms of the array? Numerical studies[1,2] suggest that the answer is affirmative. The present paper was motivated by the desire to investigate in a mathematically rigorous way problems of this type in the context of well-defined Hamiltonian models.

Motions of fast charged particles in crystals which, e.g., remain confined between lines or planes of atoms, at least over distances which are large compared to the relevant atomic spacings, are called channeling motions, or simply channeling. This is an old subject in physics which even now is not well understood mathematically. In 1912, Stark[3] suggested on purely intuitive grounds the existence of channeling of fast positive particles in crystals. Independently of Stark, the channeling phenomenon was discovered by Robinson and Oen[4] in 1963 by computer-simulation calculations. Since then, the literature on channeling has become very large (see, e.g., the bibliographies in Refs. 1 or 2) and much physical insight has been gained on the phenomenon. Knowledge of the above literature[1,2] is not needed to understand the present work, which is largely self-contained and devoted to the basic theoretical questions of classical mechanics concerned.

In this paper, we will state and prove rigorous results on the stability of certain rectilinear motions occurring in the context of two nontrivial classical Hamiltonian models in two space dimensions.[5] These models, termed the NR model and the R model, describe nonrelativistic and relativistic motion, respectively, of a particle in an external potential $V(x_1, x_2)$. We assume that the function V is real valued and analytic in the real variables x_1, x_2[6] in some neighborhood of the line $x_2 = 0$ in $\mathbb{R}$, is periodic in x_1 with period unity, is such that $\partial V(x_1,0)\partial x_2 = 0$ for $x_1 \in \mathbb{R}$, and satisfies natural stability conditions (see Sec. II for an exact statement of these conditions). Since the x_2 component of the force on the particle vanishes along the entire line $x_2 = 0$, these hypotheses entail that at large enough E the equations of motion pertaining to either of these models have solutions whose corresponding orbits are half-infinite straight-line segments of the form $\rho < x_1 < \infty, x_2 = 0$ which are traversed in the time interval $0 < t < \infty$.

Our main result—Theorem 1—is a stability theorem from which one can infer directly that this rectilinear motion is orbitally stable for fixed ρ for both the NR model and the R model, provided that the energy E of the particle is sufficiently large. Theorem 1 is stated in Sec. II, and in that section we also outline the strategy of its proof. Its assertions for the NR model and the R model are proved in Secs. III and IV, respectively. The stability proof for each of these models has two major steps. The first is to derive the asymptotic behavior in the limit $E \to \infty$ of a certain natural canonical map defined in terms of suitable Poincaré surfaces of section. The second is to apply a version of Moser's twist theorem for analytic maps[7] to the asymptotic form of the map in question (see Ref. 8 for the original version of the twist theorem and Refs. 9–13 for related results and further bibliography). The version of this theorem used in this paper is stated in the Appendix for the reader's convenience.

When suitably specialized, the above orbital stability results yield the first rigorous statements on channeling stability in the nonrelativistic and relativistic regimes for appropriate two-dimensional rigid crystal lattices, without making the customary continuum-model idealization of replacing the interaction potential between an energetic injected particle and a line of atoms by an average potential.

Key results of Secs. III and IV used to prove Theorem 1 entail that for sufficiently large E there exist abundant motions for the NR model and the R model which wind around the line $x_2 = 0$ and which can be viewed as quasiperiodic by identifying all hyperplanes $x_1 = \rho + n(|n| = 0,1,2,...)$. From previous numerical work,[1,2] one expects that chaotic motions, in some sense, also exist for these models. However, no rigorous results on the existence of chaos are known in this context, and no systematic numerical experiments to detect chaotic motion for these dynamical systems have been reported in the literature.[14,15]

The Poincaré surface-of-section method exploited in the present paper and in the work in Refs. 14 and 15 is of great potential usefulness for systematic analytical and numerical studies of regular and chaotic motions of fast particles in crystals, described by appropriate Hamiltonian models. In view of this, it is very surprising that in the literature no mention is made by other investigators of the use of this method in channeling research.

By the approach of Secs. III and IV, we have proved a generalized version of Theorem 1 which applies to a nonrelativistic and a relativistic model encompassing the NR model and the R model, respectively, as special cases. This generalization of the latter two models consists in replacing the assumption that $\partial V(x_1,0)/\partial x_2 = 0$ for $x_1 \in \mathbb{R}$ by the condition that $\int_0^1 \partial V(x_1,0)/\partial x_2 \cdot dx_1 = 0$, which demands only that the x_2 component of the force on the particle of interest vanish *on the average* along the line $x_2 = 0$. We have shown for these generalized models that at sufficiently high E there exist distinguished solutions of the equations of motion which exist for all time and are orbitally stable. The corresponding trajectories in configuration space are slightly deformed versions of the line $x_2 = 0$ for large enough E. This weakening of our assumptions on V is of both physical and mathematical interest. In particular, it is of interest physically because it models what happens in many real channeling phenomena. This generalization of Theorem 1 will be discussed in detail in a future publication.[16]

II. STATEMENT OF PRINCIPAL THEOREM AND STRATEGY OF PROOF

A. Statement of main result

We begin by defining the two-dimensional chaneling models alluded to in the Introduction. In suitable units, the Hamiltonians $H_{\rm NR}$, $H_{\rm R}$ for the NR model and the R model, respectively, are smooth real-valued functions defined for each $(x_1, x_2, p_1, p_2) \in \mathbb{R}^4$ such that (x_1, x_2) is in a neighborhood of the line $x_2 = 0$ in $\mathbb{R}^2$:

$$H_{\rm NR}(x_1, x_2, p_1, p_2) = \tfrac{1}{2}(p_1^2 + p_2^2) + V(x_1, x_2), \tag{2.1a}$$

$$H_{\rm R}(x_1, x_2, p_1, p_2) = \sqrt{p_1^2 + p_2^2 + 1} + V(x_1, x_2). \tag{2.1b}$$

The symbol H will stand for $H_{\rm NR}$ or $H_{\rm R}$.

In this paper, the potential V in (2.1a) and (2.1b) will be a fixed function which will be assumed to satisfy conditions (I)–(III) below.

(I) V is a real-valued function which is analytic in the real variables x_1, x_2[6] in the closed strip

$$\mathscr{S} = \{(x_1, x_2) \in \mathbb{R}^2 : |x_2| \leqslant \kappa\}, \tag{2.2}$$

for some constant $\kappa > 0$.

(II) In the strip (2.2), V is periodic in x_1 with period unity and

$$\frac{\partial V(x_1,0)}{\partial x_2} = 0, \quad x_1 \in \mathbb{R}. \tag{2.3}$$

(III) The following inequalities hold:

$$A_2 > 0, \tag{2.4a}$$

$$A_2 A_4 - \tfrac{5}{3} A_3^2 \neq 0, \tag{2.4b}$$

where

$$A_j \equiv \int_0^1 \frac{\partial^j V(x_1,0)}{\partial x_2^j} \cdot dx_1. \tag{2.4c}$$

The analyticity requirement in (I) was imposed to simplify the proofs; with some more work one should be able to prove Theorem 1 below for suitable potentials of an appropriate C^r class. We do not need to assume that V is periodic in x_2, since only trajectories close to the line $x_2 = 0$ are of interest here. The inequalities (2.4a) and (2.4b) are stability conditions. Condition (2.4a) is very natural physically; conditions of the same type are satisfied in continuum models of classical channeling (see, e.g., Ref. 1 or Chap. 10 of Ref. 2). Requirement (2.4b) is very weak; we imposed it in order that the relevant canonical maps were analytically conjugate to twist maps to which the version of the twist theorem in the Appendix was applicable.

Unless an explicit statement to the contrary is made, the discussions in the remainder of this sections apply to both cases $H = H_{\rm NR}$, $H = H_{\rm R}$.

For each quadruple ρ, ξ, η, E of real numbers with $(\rho, \xi) \in \mathscr{S}$ for which they exist, we denote by $x_i(t;\zeta,E)$, $p_i(t;\zeta,E)$ $(i = 1,2)$ functions having the following properties at each t in a maximal open interval of $\mathbf{R}$ containing the origin: they are differentiable in t, are such that $(x_i(t;\zeta,E), x_2(t;\zeta,E)) \in \mathscr{S}$, they satisfy the Hamiltonian equations of motion

$$\dot{x}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial x_i}, \quad i = 1,2 \tag{2.5}$$

and the initial conditions

$$\begin{aligned} &x_1(0;\zeta,E) = \rho, \quad x_2(0;\zeta,E) = \xi, \\ &p_1(0;\zeta,E) > 0, \quad p_2(0;\zeta,E) = \eta, \end{aligned} \tag{2.6}$$

and they lie on the energy surface $H = E$. Here $\zeta = (\xi,\eta)$ and we will keep ρ *fixed* in this paper.

Equation (2.3) states that the x_2 component of the force on the particle vanishes along the entire line $x_2 = 0$. Hence if the initial conditions (2.6) hold for $\xi = \eta = 0$ at some energy

$$E > V_0 \equiv \max_{|u| < \infty} V(u,0) < \infty, \tag{2.7}$$

then the particle traverses a semi-infinite rectilinear orbit in the phase space $\mathbf{R}^4$, the points (x_1, p_1, x_2, p_2) of this orbit being those on $H = E$ having coordinates $\rho < x_1 < \infty$, $p_1 > 0$, $x_2 = p_2 = 0$. The last inequality in (2.7) follows by (I) and the periodicity property in (II). Theorem 1—our principal result—and elementary considerations using the relevant energy integrals show that for large enough energy this rectilinear motion is *orbitally stable* for fixed ρ: if $|\xi|$, $|\eta|$ are small enough and E varies by a sufficiently small amount, then the distance of each point of the phase-space orbit traversed during this interval from the above rectilinear orbit is as small as desired in the cases $H = H_{\rm NR}$, $H = H_{\rm R}$.

Theorem 1: Let (I), (II), and (III) hold, and choose an arbitrary $\epsilon > 0$. Then there is an energy $E_0 = E_0(H) > V_0$ which is independent of ϵ and the following properties hold. For each $E^0 > E_0$ there exist positive constants $\sigma = \sigma(\epsilon, E^0, H)$, $e = e(\epsilon, E^0, H)$ such that when $t \geqslant 0$, $|\xi| < \sigma$, $|\eta| < \sigma$, and $E^0 - e < E < E^0 + e$; the above solution $x_i(t;\zeta,E)$, $p_i(t;\zeta,E)$ $(i = 1,2)$ exists, is unique, and satisfies $|x_2(t;\zeta,E)| < \epsilon$, $|p_2(t;\zeta,E)| < \epsilon$.

Remark: In Ref. 5, a less general theorem than this was announced for a version of the NR model in which V was assumed to be even in x_2 in the pertinent strip.

B. Strategy of proof of Theorem 1

Recall that V is a fixed potential with properties (I)–(III). The discussions in this subsection will be frequently informal. More precise definitions and statements will be made later.

The points (x_1, x_2, p_1, p_2) with $p_1 > 0$ lying on a *given* energy surface $H = E$ in the phase space $\mathbb{R}^4$ of the system (2.5) will be labeled (x_1, x_2, p_2) for convenience. In the spirit of Poincaré, we consider two hyperplanes (surfaces of section) $x_1 = \rho$ and $x_1 = \rho + 1$ in $\mathbb{R}^4$ at unit distance apart, motivated by the assumption that V has period unity in x_1. For large enough E, there is a well-defined canonical map[17] which sends each point (ρ,ξ,η) on the first hyperplane for which (ξ,η) is in a suitably small neighborhood of $0\in\mathbb{R}$ to the unique point $(\rho + 1,\xi',\eta')$ on the second hyperplane at which the phase-space orbit of system (2.5), emanating from (ρ,ξ,η) at $t = 0$, cuts the latter hyperplane. That is, for each such ζ and E, ξ' and η' are the values of $x_2(t;\zeta,E)$ and $p_2(t;\zeta,E)$, respectively, at the unique value of t at which this intersection occurs; $x_1(t;\zeta,E) = \rho + 1$. This map, which we denote by $\mathscr{P}(E)$ for short (suppressing its dependence on $H = H_{\rm NR}, H_{\rm R}$) has the form

$$\begin{aligned} \xi' &= \alpha(E)\xi + \beta(E)\eta + f(\xi,\eta,E), \\ \eta' &= \gamma(E)\xi + \delta(E)\eta + g(\xi,\eta,E), \end{aligned} \tag{2.8}$$

where $\alpha, \beta,\gamma,\delta$ are real-valued analytic functions of the real variable E for high enough E and are independent of ξ and η, and f,g are real-valued analytic functions of ξ,η, and E at each such E when (ξ,η) is in some neighborhood of the origin, the power series of f,g in ξ,η starting with second-degree terms. Since $\mathscr{P}(E)$ is an area- and orientation-preserving map at those E at which it exists, the matrix

$$M(E) = \begin{pmatrix} \alpha(E) & \beta(E) \\ \gamma(E) & \delta(E) \end{pmatrix} \tag{2.9}$$

has determinant equal to unity at each such E.

Theorem 1 follows by an easy standard argument once we have shown that the fixed point $0\in\mathbb{R}^2$ is stable under $\mathscr{P}(E)$ for sufficiently large E, in the sense of Eq. (3.38) of Sec. III D. A major step in proving this stability result is to derive asymptotic formulas for the derivatives

$$\left.\frac{\partial^{j+k}\xi'}{\partial\xi^j\,\partial\eta^k}\right|_0, \quad \left.\frac{\partial^{j+k}\eta'}{\partial\xi^j\,\partial\eta^k}\right|_0, \quad j\geqslant 1, \quad k\geqslant 1, \quad j+k\leqslant 3, \tag{2.10}$$

in the limit $E\to\infty$, where the zero subscript means that they are evaluated at $\xi = \eta = 0$. We derived such formulas by the simple approach explained in Secs. III B, III C, and IV B. This involved estimating solutions of certain systems of integral equations. These systems are equivalent to systems of variational equations of the first, second, and third orders (corresponding to the above rectilinear solution along $x_2 = 0$) plus the relevant initial conditions. In this paper, it will be unnecessary to consider these variational equations.

One can show (see Lemmas 3.3 and 4.3 and their proofs) that the eigenvalues of $M(E)$, if labeled appropriately (to insure analyticity), have the following properties for large enough E.

(i) They are of the form $\lambda(E)$, $\overline{\lambda(E)}$, with $\lambda(E)\neq\pm 1$, $|\lambda(E)| = 1$, [i.e., $0\in\mathbb{R}$ is an elliptic fixed point of $\mathscr{P}(E)$ at high enough E] and are such that λ is an analytic function of E at each such E.

(ii) One has $\lambda^3(E)$, $\lambda^4(E)\neq 1$.

By these properties and the stated analyticity properties of the map (2.8), we can express this canonical map in the Birkhoff normal form[18]

$$z' = \lambda(E)z[1 + \mu(E)|z|^2] + O(|z|^4), \quad z\to 0, \tag{2.11}$$

for sufficiently large E by a simultaneous analytic change of variables from ξ,η (respectively, ξ',η') to the complex variables $z, \bar z$ (respectively, $z', \bar z'$). At each of the latter E values, the first twist coefficient $\mu(E)$ is a real-valued analytic function of E, indepenent of $z,\bar z$. This coefficient depends on how one chooses a matrix which diagonalizes $M(E)$ at large enough E. If one selects this matrix for the NR model and the R model as is done in this paper (see Secs. III C and IV B), $\mu(E)$ is given asymptotically by

$$\mu(E) = KE^{-1/2} + O(E^{-1}), \quad \text{NR model}, \tag{2.12a}$$

$$\mu(E) = K'E^{-1} + O(E^{-3/2}), \quad \text{R model}, \tag{2.12b}$$

for $E\to\infty$, where K, K' are nonzero constants depending only on A_2, A_3, A_4 in (2.4c). Equations (2.12a) and (2.12b) are derived by using the asymptotic formulas, alluded to above, for the derivatives (2.10) (see Lemmas 3.5 and 4.5 and their proofs).

The point of Eqs. (2.12) is that they entail that $\mu(E)\neq 0$ for both of these models when E is large enough. The following discussion applies to the cases $H = H_{\rm NR}$ and $H = H_{\rm R}$. By $\mu(E)\neq 0$ and certain analyticity properties of a mapping induced by (2.11), one concludes that Moser's twist theorem (in the version stated in the Appendix) applies to that induced mapping. By this theorem and elementary arguments, we infer that $\mathscr{P}(E)$ has an analytic closed invariant curve in each punctured neighborhood of the point $0\in\mathbb{R}^2$ when E is sufficiently large. A standard argument now shows that this point is stable under $\mathscr{P}(E)$ in the desired sense at each such E, and thus that Theorem 1 is true. These considerations are presented in detail only for the case $H = H_{\rm NR}$ (see Sec. III D), since the corresponding stability proof for $H = H_{\rm R}$ is similar.

III. PROOF OF THEOREM 1 FOR THE NR MODEL

The present section is divided into four subsections, III A–III D. In Sec. III A, we state certain existence, analyticity, and asymptotic results for the NR model which will be used in subsequent proofs. In Secs. III B and III C we give asymptotic estimates for the canonical map (2.8) for this model in the limit $E\to\infty$. A proof of Theorem 1 for the NR model is given in Sec. III D.

In the remainder of this paper, it should be understood that equations in which the symbol $O(E^{-r})$ occurs for some $r > 0$ hold in the limit $E\to\infty$. In most cases, this limit will not be explicitly stated, in the interest of brevity.

Throughout the present section, $x_i(t;\zeta,E)$, $p_i(t;\zeta,E)$ $(i = 1,2)$ will always denote a solution of (2.5) as defined in Sec. II A, specialized to $H = H_{\rm NR}$.

A. Auxiliary results

Suppose that ζ,E are such that $(\rho,\zeta)\in\mathscr{S}$ [see (2.2)] and

$$E>\eta^2/2+V(\rho,\zeta), \tag{3.1}$$

this last inequality being necessary in order for the condition $p_1(0;\zeta,E)>0$ in (2.6) to hold in the present case $H=H_{\rm NR}$. Then by standard results on ordinary differential equations, one knows that such a solution $x_i(t;\zeta,E)$, $p_i(t;\zeta,E)(i=1,2)$ exists and is unique at all t in some maximum interval $I(\zeta,E)\subset\mathbb{R}$ containing the origin. At each ζ,E of the last-mentioned type, the functions $x_i(t;\zeta,E)$ satisfy the integral equations

$$x_i(t;\xi,E)=\xi_i+\eta_i\,t+\int_0^t(t-s)f_i(x(s;\xi,E))\,ds,\quad i=1,2, \tag{3.2}$$

for $t\in I(\zeta,E)$. In (3.2), $x(s;\zeta,E)=(x_1(s;\zeta,E),x_2(s;\zeta,E))$,

$$\xi_1=\rho,\quad \xi_2=\zeta,$$
$$\eta_1=\eta_1(\zeta,E)=\sqrt{2[E-V(\rho,\zeta)]-\eta^2}, \tag{3.3}$$
$$\eta_2=\eta,$$

and

$$f_i(x)=-\frac{\partial V(x)}{\partial x_i},\quad i=1,2. \tag{3.4}$$

In order to derive the asymptotic behavior of the canonical map $\mathscr{P}_{\rm NR}(E)$ for $E\to\infty$, we need certain results on existence and uniqueness of analytic solutions of Eq. (3.2). These results are stated in Lemma 3.1 below.

By condition (I) and the periodicity property in (II), the potential V has the following property.

(I′) V has an analytic extension, also denoted by V, such that $V(z_1,z_2)$ is analytic in the complex variables (z_1,z_2) (Ref. 6) at each point (z_1,z_2) of a closed region

$$\mathscr{R}=\{(z_1,z_2)\in\mathbb{C}^2:|{\rm Im}\,z_1|\leqslant a_0,|z_2|\leqslant a_0\}, \tag{3.5}$$

where a_0 is a positive constant not larger than κ in (2.2).

By (I′),

$$V_1\equiv\max_{(z_1,z_2)\in\mathscr{R}}|V(z_1,z_2)|<\infty,$$

$$N\equiv\max_{i=1,2}\max_{(z_1,z_2)\in\mathscr{R}}\left|\frac{\partial V(z_1,z_2)}{\partial z_i}\right|<\infty. \tag{3.6}$$

Notice also that (3.1) holds under the assumptions of the next lemma.

Lemma 3.1: Choose positive constants a,b with $a<a_0$, and select a sufficiently large positive constant E_1 so that

$$E_1>V_1+\tfrac{1}{2}b^2,$$

in particular. Then the solution $x_i(t;\zeta,E)$, $p_i(t;\zeta,E)(i=1,2)$ exists, is unique, and is such that each of the latter four functions is analytic in t,ξ,η,E for $(\zeta,(t,E))\in\mathscr{U}\times\mathscr{V}\subset\mathbb{R}^4$. Here

$$\mathscr{U}=\{(u_1,u_2)\in\mathbb{R}^2:|u_1|<a,|u_2|<b\},$$

$$\mathscr{V}=\{(t,E)\in\mathbb{R}^2:|t|<t_0,E>E_1\}, \tag{3.7}$$

where t_0 is the unique positive root of

$$Nt^2+2bt-(a_0-a)=0.$$

This solution has the property

$$(x_1(t;\zeta,E),x_2(t;\zeta,E))\in\mathscr{R}\cap\mathbb{R}\subset\mathscr{S}$$
$$\text{if } (\zeta,(t,E))\in\mathscr{U}\times\mathscr{V}.$$

Remark: One of the main points of this lemma is that the stated properties of this solution hold on the interval $|t|<t_0$, with t_0 a positive number *independent* of E. A much better existence theorem for Eqs. (2.5) at high enough E follows by using the theory of averaging as expounded, e.g., in Ref. 19. However, Lemma 3.1 amply suffices for our purposes. Similar comments apply to the existence portion of Lemma 4.1.

Proof of Lemma 3.1: This lemma is a real-valued version of appropriate existence and uniqueness results on solutions of (3.2) in the complex domain. These results are themselves versions of familiar classical results and were proved by the standard method of successive approximations. We omit further details. ∎

Let $E_2\geqslant E_1$ be a sufficiently large constant. Then for each $\zeta\in\mathscr{U}$ and $E>E_2\geqslant E_1$, the function $t\mapsto u$ with

$$u=x_1(t;\zeta,E) \tag{3.8}$$

maps $(-t_0,t_0)$ bijectively onto an interval $J(\zeta,E)\subset\mathbb{R}$ containing $[\rho,\rho+1]$. Moreover, $t=g(u;\zeta,E)$ for $u\in J(\zeta,E)$ at each such ζ,E where at these values of its arguments g is a real-valued analytic function of u,ξ,η,E which is a strictly increasing function of u. To prove these assertions, we observe first that

$$\dot{x}_1(t;\zeta,E)=\sqrt{2[E-V(x(t;\zeta,E))]-\dot{x}_2^2(x(t;\zeta,E))}$$
$$=(2E)^{1/2}+O(E^{-1/2})>0, \tag{3.9}$$

for $E\to\infty$, uniformly with respect to t,ζ for $t\in(-t_0,t_0)$, $\zeta\in\mathscr{U}$, as follows by using, in particular, the integral of energy and the fact that $V(x(t;\zeta,E))$ and $\dot{x}_2(t;\zeta,E)$ are bounded at each such t,ζ for $E>E_1$ [by (I) and Lemma 3.1] so that there exists $E_2\geqslant E_1$ so large that the quantity inside the square root of (3.9) is positive, and by invoking (2.6) and a continuity argument. Thus the invertibility of $t\mapsto u$ is guaranteed at *all* such t,ζ,E. The proof of the assertions in the second and third sentences of this paragraph is completed by making use of the analytic version of the implicit function theorem, in particular.

These properties of the diffeomorphism $t\mapsto u$ entail that at each $\zeta\in\mathscr{U}$, $E>E_2$ there is a unique time $\tau(\zeta,E)$ at which the solution curve of system (2.5) for $H=H_{\rm NR}$, originating at a point $(x_1=\rho,x_2=\zeta,p_2=\eta)$ in the intersection of the surface of section $x_1=\rho$ with the energy surface $H_{\rm NR}=E$ at $t=0$, intersects the surface of section $x_1=\rho+1$. That is, the equation

$$x_1(\tau(\zeta,E);\zeta,E)=\rho+1 \tag{3.10}$$

has a unique solution $\tau(\zeta,E)$ at each such ζ, E. One can easily show, by (3.9) in particular, that

$$\tau(\zeta,E)=(2E)^{-1/2}+O(E^{-3/2}),$$

for $E\to\infty$, uniformly in ζ on $\mathscr{U}$. This result will not be needed here.

Our main interest in the diffeomorphism $t\mapsto u$ is that it is essential for deriving the integral equations (3.12) below, whose use greatly lightens the labor of obtaining the required asymptotic estimates of the derivatives (2.10).

Define the functions X,P,P_1 as follows in terms of the solution $x_i(t;\zeta,E), p_i(t;\zeta,E)(i=1,2)$:

$$X(u;\zeta,E)=x_2(t;\zeta,E), \tag{3.11a}$$

$$P(u;\zeta,E)=p_2(t;\zeta,E), \tag{3.11b}$$

$$P_1(u;\zeta,E)=p_1(t;\zeta,E) = \sqrt{2[E-V(u,X(u;\zeta,E))]-P^2(u;\zeta,E)}. \tag{3.11c}$$

for $u\in J(\zeta,E)$, $\zeta\in\mathscr{U}$, $E>E_2$. By Lemma 3.1, (3.11), and the fact that $t\mapsto u$ is a diffeomorphism with the stated analyticity and other properties, X,P,P_1 are analytic functions of u,ξ,η, E at the latter values of u,ζ,E.

At each such u,ζ,E, the functions X and P obey a second-order system of Hamiltonian differential equations

$$\frac{dX(u/\zeta,E)}{du}=\frac{P(u;\zeta,E)}{P_1(u;\zeta,E)}, \tag{3.12a}$$

$$\frac{dP(u;\zeta,E)}{du}=\frac{f_2(u;\zeta,E)}{P_1(u;\zeta,E)}. \tag{3.12b}$$

These equations follow directly by changing the independent variable from t to u in the equations of motion of type (2.5) or $H=H_{\rm NR}$, whose left sides are $\dot x_2(t;\zeta,E)$ and $\dot p_2(t;\zeta,E)$, and using (3.11). This is a simple example of isoenergetic reduction.[20] The equations of real interest here are the integral equations

$$X(u;\zeta,E)=\xi+\int_\rho^u \frac{P(u';\zeta,E)}{P_1(u';\zeta,E)}\,du', \tag{3.13a}$$

$$P(u;\zeta,E)=\eta+\int_\rho^u \frac{f_2(u';\zeta,E)}{P_1(u';\zeta,E)}\,du', \tag{3.13b}$$

equivalent to the differential equations obeyed by X and P plus the relevant initial conditions (2.6), and which hold at the latter u,ζ,E values.

For our purposes, the advantage of working with Eqs. (3.13) is that it avoids the need to estimate derivatives of $\tau(\zeta,E)$ and $x_1(\tau(\zeta,E);\zeta,E)$ with respect to ξ,η in order to obtain the needed estimates of the derivatives (2.10) of orders greater than unity, as would be necessary when using (3.2) and the analogous integral equations for $p_i(t;\zeta,E)(i=1,2)$ to accomplish this objective. This represents a considerable simplification.

B. Asymptotic formulas concerning the linear portion of $\mathscr{P}_{\rm NR}(E)$

Before deriving these formulas, we will define the mapping $\mathscr{P}_{\rm NR}(E)$, which is a precise version for $H=H_{\rm NR}$ of the mapping $\mathscr{P}(E)$ defined informally in Sec. II B.

Now,

$$\begin{aligned} X(\rho+1;\zeta,E)&=x_2(\tau(\zeta,E);\zeta,E),\\ P(\rho+1;\zeta,E)&=p_2(\tau(\zeta,E);\zeta,E), \end{aligned} \tag{3.14}$$

for $\zeta\in\mathscr{U}$, $E>E_2\geqslant E_1$, by the bijective property of the map $t\mapsto u$, (3.10), and (3.11a), where $\mathscr{U}$, E_1,E_2 are defined in (3.7), Lemma 3.1, and the first sentence of the paragraph containing (3.8). Therefore, $\mathscr{P}_{\rm NR}(E)$ can be defined as follows.

Definition: For each $E>E_2$, $\mathscr{P}_{\rm NR}(E)$ is a mapping with domain $\mathscr{U}$ which sends every $(\xi,\eta)\in\mathscr{U}$ into

$$(\xi',\eta')=(X(\rho+1;\zeta,E),P(\rho+1;\zeta,E))\in\mathbb{R}^2.$$

Remarks: (1) $\mathscr{P}_{\rm NR}(E)$ is well defined at each $E>E_2$ and is a canonical map[17] such that writing $\xi'=\phi_1(\xi,\eta,E)$, $\eta'=\phi_2(\xi,\eta,E)$, the functions ϕ_1,ϕ_2 are analytic in ξ, η, E for $(\xi,\eta)\in\mathscr{U}$, $E>E_2$. These properties follow by the analyticity properties of X and P, and a well-known theorem.[21]

(2) Standard arguments using the periodicity of V in (II) show that each iterate $(\mathscr{P}_{\rm NR}(E))^n(n\geqslant 1)$ of $\mathscr{P}_{\rm NR}(E)$, if it exists, has the following property. Consider a sufficiently large E and let $(\xi_n,\eta_n)=(\mathscr{P}_{\rm NR}(E))^n(\xi,\eta)$. Then (in the notation of Sec. II B) $(x_1=\rho+n, x_2=\xi_n, p_2=\eta_n)$ $(n\geqslant 1)$ is the point in the intersection of the surface of section $x_1=\rho+n$ with the energy surface $H_{\rm NR}=E$ at which the solution curve of the system (2.5) for $H=H_{\rm NR}$, emanating at $t=0$ from the point $(x_1=\rho, x_2=\xi, p_2=\eta)$ on this energy surface, intersects the latter surface of section.

(3) The property of the iterates of $\mathscr{P}_{\rm NR}(E)$ stated in the previous remark, together with a suitable version of Moser's twist theorem, will allow us to go from the considerations of this and the next subsection, mostly limited to *finite* time intervals, to the stability result for the case $H=H_{\rm NR}$ for an *infinite* time interval asserted by Theorem 1.

The next two lemmas deal with the linear part of $\mathscr{P}_{\rm NR}(E)$, i.e., with the functions $\alpha,\beta,\gamma,\delta$ given by

$$\begin{aligned} \alpha(E)&=\frac{\partial X(\rho+1;0,E)}{\partial\xi}, & \beta(E)&=\frac{\partial X(\rho+1;0,E)}{\partial\eta},\\ \gamma(E)&=\frac{\partial P(\rho+1;0,E)}{\partial\xi}, & \delta(E)&=\frac{\partial P(\rho+1;0,E)}{\partial\eta}, \end{aligned} \tag{3.15}$$

which are analytic in E for $E>E_2$ by the analyticity properties of X and P.

Lemma 3.2: One has for the functions $\alpha,\beta,\gamma,\delta$ in (3.15)

$$\alpha(E)=1+\frac{1}{2E}\int_\rho^{\rho+1}(1+\rho-u)\,\frac{\partial f_2(u,0)}{\partial x_2}\,du+O(E^{-2}), \tag{3.16a}$$

$$\beta(E)=(2E)^{-1/2}+O(E^{-3/2}), \tag{3.16b}$$

$$\gamma(E)=(2E)^{-1/2}\int_0^1\frac{\partial f_2(u,0)}{\partial x_2}\,du+O(E^{-3/2}), \tag{3.16c}$$

$$\delta(E)=1+\frac{1}{2E}\int_\rho^{\rho+1}(u-\rho)\,\frac{\partial f_2(u,0)}{\partial x_2}\,du+O(E^{-2}). \tag{3.16d}$$

Proof: Assume that the hypotheses of the lemma are made. We will show that (3.16a) and (3.16c) hold, (3.16b) and (3.16d) being provable similarly.

To prove (3.16a) and (3.16c), we differentiate Eqs. (3.13), with respect to ξ, thus obtaining for $u\in[\rho,\rho+1]$, $E>E_2$

$$\frac{\partial X(u;0,E)}{\partial\xi}=1+\int_\rho^u\frac{1}{P_1(u';0,E)}\,\frac{\partial P(u';0,E)}{\partial\xi}\,du', \tag{3.17a}$$

$$\frac{\partial P(u;0,E)}{\partial\xi}=\int_\rho^u\frac{1}{P_1(u';0,E)}\,\frac{\partial f_2(u';0)}{\partial x_2}\times\frac{\partial X(u';0,E)}{\partial\xi}\,du', \tag{3.17b}$$

where, of course, $\partial f_2/\partial x_2$ is the derivative of f_2 with respect to the second variable and where we have used (2.3), (3.11b), and the equations

$$X((u;0,E)=0,\quad P(u;0,E)=0, \tag{3.18}$$

which hold for $u\in\mathbb{R}$, $E>E_2$. These equations follow by (3.11a) and the existence, for sufficiently large E, of a rectilinear solution of Eqs. (2.5) for $H=H_{\rm NR}$, such that $x_2(t;0,E)=p_2(t;0,E)=0$ $(t\in\mathbb{R})$.

The proof that the right sides of (3.17a) and (3.17b) have the respective asymptotic forms (3.16a) and (3.16c) when $u=\rho+1$ will be carried out in two main steps.

(1) In this step, we will prove that

$$\frac{\partial X(u;0,E)}{\partial\xi}=1+O(E^{-1}), \tag{3.19a}$$

$$\frac{\partial P(u;0,E)}{\partial\xi}=O(E^{-1/2}), \tag{3.19b}$$

for $E\to\infty$, uniformly with respect to u on $[\rho,\rho+1]$.

By (3.9), (3.11c), and obvious properties of $t\mapsto u$,

$$P_1(u;\zeta,E)=(2E)^{1/2}+O(E^{-1/2}), \tag{3.9'}$$

in the latter limit, uniformly in u,ζ for $u\in[\rho,\rho+1]$, $\zeta\in\mathscr{U}$.

By (3.17), (3.9'), and the boundedness of $\partial f_2(u,0)/\partial x_2$ [recall (I) and (3.4)],

$$\left|\frac{\partial X(u;0,E)}{\partial\xi}\right|\leqslant 1+CE^{-1/2}\int_\rho^u\left|\frac{\partial P(u';0,E)}{\partial\xi}\right|du',$$

$$\left|\frac{\partial P(u;0,E)}{\partial\xi}\right|\leqslant C'E^{-1/2}\int_\rho^u\left|\frac{\partial X(u';0,E)}{\partial\xi}\right|du'$$

at the latter u, E if E is sufficiently large, the positive numbers C, C' being independent of u,E. Substituting the second of these inequalities in the first and either integrating the resulting inequality over $[u,\rho]$ or, more elegantly, applying the usual Gronwall estimate, we see that

$$\frac{\partial X(u;0,E)}{\partial\xi}=O(1),$$

for $E\to\infty$ in the above uniform sense, whence (3.19b) obtains in the same sense in this limit. That (3.19a) holds in the desired sense now follows by (3.9), (3.17a), and (3.19b).

(2) Using (3.17), the fact that (3.9') and (3.19) are true in the specified uniform sense, and the periodicity property of V in (II), the desired asymptotic results (3.16a) and (3.16b) emerge readily. ■

Let $M_{\rm NR}(E)$ be the matrix (2.9), with elements defined by (3.15). The next lemma concerns properties of the eigenvalues of $M_{\rm NR}(E)$.

Lemma 3.3: At each $E>E_3$, where $E_3\geqslant E_2$ is a sufficiently large constant, the eigenvalues of $M_{\rm NR}(E)$ can be labeled so that they have properties (i) and (ii) of Sec. II B.

Proof: By (2.4a), (3.4), (3.16a), (3.16d), and $V(u+1,0)=V(u,0)$ $(u\in\mathbb{R})$, one sees that

$$\alpha(E)+\delta(E)=2-A_2/2E+O(E^{-2}), \tag{3.20}$$

whence by (2.4a) and (3.20),

$$0<\alpha(E)+\delta(E)<2, \tag{3.21}$$

for large enough E. By (3.21) and $\det M_{\rm NR}(E)=1$, we see that the eigenvalues of $M_{\rm NR}(E)$ at each such E are nonreal, lie on the unit circle, and are equal to

$$\lambda(E)=\tfrac{1}{2}\{\alpha(E)+\delta(E)+i\sqrt{4-[\alpha(E)+\delta(E)]^2}\} \tag{3.22}$$

and its complex conjugate. One readily proves that $\lambda(E)$ in (3.22) has the remaining properties (i) and (ii) when E is sufficiently large. ■

Notice that (2.4a), (3.20), and (3.22) imply that

$$\begin{aligned}\operatorname{Re}\lambda(E)&=1-A_2/4E+O(E^{-2}),\\ \operatorname{Im}\lambda(E)&=(A_2/2E)^{1/2}+O(E^{-3/2}).\end{aligned} \tag{3.23}$$

C. Asymptotic formulas for the quadratic and cubic nonlinearities of $\mathscr{P}_{\rm NR}(E)$ and proof of (2.12a)

Asymptotic formulas for the derivatives (2.10) of second and third orders in the limit $E\to\infty$ will be needed to derive Eq. (2.12a). We obtained these asymptotic formulas by arguments of similar kind, but more complicated than those used to prove (3.16).

The main step in deriving these formulas is to obtain appropriate estimates of the derivatives

$$\frac{\partial^n X(u;0,E)}{\partial\xi^m\,\partial\eta^{n-m}},\quad \frac{\partial^n P(u;0,E)}{\partial\xi^m\,\partial\eta^{n-m}},\qquad m=1,\ldots,n,\quad \rho\leqslant u\leqslant\rho+1, \tag{3.24}$$

of orders $n=2,3$ in this limit. Actually, such estimates can be found for derivatives (3.24) of any order n by the procedures sketched below.

Differentiating (3.13) and using (2.3), (3.4), and (3.18), one gets for $u\in[\rho,\rho+1]$, $E>E_2$

$$DX(u;0,E)=D\xi+\int_\rho^u\frac{1}{P_1(u';0,E)}DP(u';0,E)\,du'+\cdots, \tag{3.25a}$$

$$DP(u;0,E)=D\eta+\int_\rho^u\frac{1}{P_1(u';0,E)}\frac{\partial f_2(u',0)}{\partial x_2}\times DX(u';0,E)\,du'+\cdots, \tag{3.25b}$$

where $D=\partial^p/\partial\xi^q\,\partial\eta^{p-q}$ for any given $0\leqslant q\leqslant p$, $p\geqslant1$, and $+\cdots$ denotes a sum of integrals over $[\rho,u]$ in whose integrands only derivatives of orders less than p are present.

Using (3.25) in conjunction with an approach similar to that by which Eqs. (3.19) were proved and a simple inductive argument, each derivative in the first (second) line of (3.24) can be estimated as being $O(E^{-c(n,m)/2})(O(E^{-d(n,m)/2}))$ in the limit $E\to\infty$, uniformly in u on $[\rho,\rho+1]$. Here $c(n,m)$, $d(n,m)$ are appropriate non-negative integers. Indeed, a systematic and straightforward inductive procedure can be easily devised for obtaining estimates of this type for derivatives (3.24) of any order n, given similar estimates of lower-order derivatives (3.24).

By such crude uniform estimates of the derivatives (3.24) with $n=1,2,3$, together with procedures analogous to Step (2) of the proof of Lemma 3.2, one arrives at the results stated in Lemma 3.4.

Lemma 3.4: Using the notation (2.4c), one has the following.

(1) The quadratic nonlinearities of $\mathscr{P}_{\rm NR}(E)$ have the asymptotic behavior

$$\frac{\partial^2 X(\rho+1;0,E)}{\partial\xi^{2-s}\,\partial\eta^s}=O(E^{-1-s/2}),\quad s=0,1,2, \tag{3.26a}$$

$$\frac{\partial^2 P(\rho+1;0,E)}{\partial \xi^2} = -A_3(2E)^{-1/2} + O(E^{-3/2}), \qquad (3.26b)$$

$$\frac{\partial^2 P(\rho+1;0,E)}{\partial \xi^{2-s}\,\partial \eta^s} = O(E^{-1/2-s/2}), \quad s=1,2. \qquad (3.26c)$$

(2) The cubic nonlinearities of $\mathscr{P}_{\rm NR}(E)$ behave as

$$\frac{\partial^3 X(\rho+1;0,E)}{\partial \xi^{3-s}\,\partial \eta^s} = O(E^{-1-s/2}), \quad s=0,1,2, \qquad (3.27a)$$

$$\frac{\partial^3 X(\rho+1;0,E)}{\partial \eta^3} = O(E^{-3/2}), \qquad (3.27b)$$

$$\frac{\partial^3 P(\rho+1;0,E)}{\partial \xi^3} = -A_4(2E)^{-1/2} + O(E^{-3/2}), \qquad (3.27c)$$

$$\frac{\partial^3 P(\rho+1;0,E)}{\partial \xi^{3-s}\,\partial \eta^s} = O(E^{-1/2-s/2}), \quad s=1,2,3. \qquad (3.27d)$$

We now digress to discuss more fully the reduction of $\mathscr{P}_{\rm NR}(E)$ to Birkhoff normal form and related matters. The results of this discussion will be needed in this and the next subsection.

Consider any $E > E_4$, where $E_4 \geqslant E_3$ is a sufficiently large constant. By Remark (1) of Sec. III B, Lemma 3.3, and the general theory,[22] $\mathscr{P}_{\rm NR}(E)$ is conjugate to a canonical mapping $\mathscr{Q}_E$ in Birkhoff normal form

$$\mathscr{Q}_E = \psi_E^{-1} \circ \mathscr{P}_{\rm NR}(E) \circ \psi_E. \qquad (3.28)$$

At each such E, ψ_E, $(x,y) \mapsto (x',y') = (\psi_1(x,y,E), \psi_2(x,y,E))$ is a diffeomorphism from an open ball $\mathscr{U}_E = \{(x,y) \in \mathbb{R}^2 : (x^2+y^2)^{1/2} < \rho(E)\}$ with $\rho(E) > 0$ onto a neighborhood of $0 \in \mathbb{R}^2$. Indeed, ψ_1, ψ_2 are analytic functions of x, y, E for $(x,y) \in \mathscr{U}_E$, $E > E_4$ and $x = \chi_1(x',y',E)$, $y = \chi_2(x',y',E)$, with χ_1, χ_2 analytic in x', y', E for $(x',y') \in \psi_E(\mathscr{U}_E)$, $E > E_4$. If $(x,y) \in \mathscr{U}_E$, $E > E_4$, then

$$\mathscr{Q}_E(x,y) = (x',y'),$$

where, writing $z = x + iy$ and $z' = x' + iy'$,

$$z' = \lambda(E) z[1 + i\mu(E)|z|^2] + O(|z|^4), \quad z \to 0. \qquad (3.29)$$

Of course, here and in the remainder of this section, $\lambda(E)$ is as in (3.22), and hence has the asymptotic behavior (3.23).

The function μ in (3.29) is real valued and analytic for $E > E_4$ and is given at each such E by the formula[23]

$$\mu = -i\left[i\,\mathrm{Im}(\bar\lambda p_2^{(3)}) + \left(\frac{\lambda^3+1}{\lambda^3-1}\right) \times |p_0^{(2)}|^2 + 3\left(\frac{\lambda+1}{\lambda-1}\right)|p_2^{(2)}|^2 \right], \qquad (3.30)$$

where we have written $\lambda = \lambda(E)$ for short and have omitted the dependence of the other pertinent symbols on E. Here

$$p_l^{(n)} = p_l^{(n)}(E) = [(n-l)!\,l!]^{-1} \left.\frac{\partial^n G(u,\bar u,E)}{\partial u^l\,\partial \bar u^{n-l}}\right|_{u=0}, \qquad (3.31)$$

where $G(u,\bar u,E)$ is defined by the equation

$$\begin{bmatrix} G(u,\bar u,E) \\ \overline{G(u,\bar u,E)} \end{bmatrix} = C(E)^{-1}\begin{pmatrix} \xi' \\ \eta' \end{pmatrix}, \qquad (3.32a)$$

for $(\xi,\eta) \in \mathscr{U}, E > E_4$. Here $(\xi',\eta') = \mathscr{P}_{\rm NR}(E)(\xi,\eta)$ and the complex numbers $u, \bar u$ are related to the real numbers ξ, η by

$$\begin{pmatrix} u \\ \bar u \end{pmatrix} = C(E)^{-1}\begin{pmatrix} \xi \\ \eta \end{pmatrix}, \qquad (3.32b)$$

where $C(E)$ is a nonsingular 2×2 matrix of the form

$$C(E) = \begin{pmatrix} c_1(E) & \overline{c_1(E)} \\ c_2(E) & \overline{c_2(E)} \end{pmatrix}, \quad \det C(E) = -i, \qquad (3.33)$$

which diagonalizes $M_{\rm NR}(E)$

$$C(E)^{-1} M_{\rm NR}(E) C(E) = \begin{pmatrix} \lambda(E) & 0 \\ 0 & \overline{\lambda(E)} \end{pmatrix}. \qquad (3.34)$$

For $E > E_4$, $C(E)$ exists and c_1, c_2 can (and will) be chosen to be analytic at each such E. This follows by invoking, in particular, Lemma 3.3, the reality of $\alpha(E), \beta(E), \gamma(E), \delta(E)$ over the latter range of E, and the fact that over that range the function β is analytic and $\beta(E) > 0$.

By (2.9), (3.16), (3.23), (3.33), and (3.34),

$$|c_1(E)| = 2^{-1/2} A_2^{-1/4} + O(E^{-1}), \qquad |c_2(E)| = 2^{-1/2} A_2^{1/4} + O(E^{-1}). \qquad (3.35)$$

These two formulas will be useful in proving the more explicit version of (2.12a) given by the next lemma.

Lemma 3.5: Using the notation (2.4c),

$$\mu(E) = 2^{-7/2} A_2^{-2}(A_2 A_4 - \tfrac{5}{3} A_3^2) E^{-1/2} + O(E^{-1}), \qquad (3.36)$$

for the NR model.

Remarks: (1) By (2.4b) and (3.36), there exists a constant $E_5 \geqslant E_4$ such that

$$\mu(E) \neq 0, \quad E > E_5, \qquad (3.37)$$

in the case $H = H_{\rm NR}$ under discussion.

(2) In particular, by (3.37), Remark (1) of Sec. III B (fourth paragraph) and the discussion in the paragraph containing (3.28) we may apply the Birkhoff–Lewis theorem[24] to conclude that for each $\epsilon > 0$, $E > E_5$, the mapping $\mathscr{P}_{\rm NR}(E)$ has a periodic point in each punctured disk $0 < \xi^2 + \eta^2 < \epsilon^2$ whose period n tends to infinity as $\epsilon \to 0$ [i.e., $(\mathscr{P}_{\rm NR}(E))^n$ has a fixed point in each such disk for such a positive integer n having this latter property]. For reasons of the same kind, the counterpart $\mathscr{P}_{\rm R}(E)$ of $\mathscr{P}_{\rm NR}(E)$ in the case $H = H_{\rm R}$ has a similar property at sufficiently high energy.

Proof of Lemma 3.5: Using (3.31)–(3.34), we express the functions $p_l^{(n)}$ occurring in (3.30) as linear combinations of the derivatives (3.26) or (3.27), with coefficients depending on $c_1(E), c_2(E)$. Combining this result with (3.23), (3.26), (3.27), and (3.35), the desired formula (3.36) emerges. ■

D. Proof of Theorem 1 for $H = H_{\rm NR}$

Choose any $\tilde E > E_5$ and any compact interval $J \subset (E_5, \infty)$ having $\tilde E$ in its interior. If $E \in J$, then the mapping $\mathscr{Q}_E$ in (3.28) satisfies all the conditions of Theorem A1 of the Appendix (specializing ν, M_ν, and $[a,b]$ to E, a suitable restriction of $\mathscr{Q}_E$, and J, respectively). This follows by (3.37) and the facts that for each $E > E_5 \geqslant E_4$ the mapping $\mathscr{Q}_E$ is area preserving and that all the other properties stated in the paragraph containing (3.28) hold. Notice in particular that the analyticity properties of z' in (3.29) as a function of $z, \bar z, E$ and a finite-covering argument entail that $\mathscr{Q}_E$ is defined on an E-independent neighborhood of $0 \in \mathbb{R}^2$ for $E \in J$.

Applying Theorem A1 to $\mathscr{Q}_E$, and in view of (3.28) and the analyticity properties of $\mathscr{P}_{\rm NR}(E)$, ψ_E, ψ_E^{-1}, one sees that for each $\epsilon > 0$, $E \in J$ the mapping $\mathscr{P}_{\rm NR}(E)$ has an invariant curve of the form (A1) lying in the punctured disk $0 < x^2 + y^2 < \epsilon^2$. Whence by a familiar elementary argument which uses the continuous dependence of this invariant curve on E, one concludes there exist positive constants $c = c(\epsilon,\widetilde{E})$, $d = d(\epsilon,\widetilde{E})$ such that

$$\|(\mathscr{P}_{\rm NR}(E))^n(x,y)\| < \epsilon, \quad n = 1,2,..., \qquad \text{if } x^2 + y^2 < c^2 \text{ and } |E - \widetilde{E}| < d, \tag{3.38}$$

where $\|\cdot\|$ is the usual Euclidean norm in $\mathbb{R}^2$. By this stability result and a standard elementary argument, the orbital stability assertions of Theorem 1 follow for the case $H = H_{\rm NR}$. This argument uses the property of $(\mathscr{P}_{\rm NR}(E))^n$ in Remark (2) of Sec. III B and the fact that $|X(u;\zeta,E)|$, $|P(u;\zeta,E)| < C(\xi^2 + \eta^2)^{1/2}$ for $u\in[\rho,\rho+1]$, $\zeta\in\mathscr{U}$, $E \in J$, C being independent of u, ζ, E. ■

IV. PROOF OF THEOREM 1 FOR THE R MODEL

The results of this section can be proved by arguments of the same type as those in Sec. III, and hence we will at most give brief indications of their proofs. After presenting some auxiliary results in Sec. IV A, the relevant asymptotic formulas needed to prove Theorem 1 for the case $H = H_{\rm R}$ will be stated in Sec. IV B.

Throughout this section, $x_i(t;\xi,E)$, $p_i(t;\xi,E)(i = 1,2)$ will denote a solution of (2.5) as defined in Sec. II A, but specialized to $H = H_{\rm R}$. We remind the reader that conditions (I)–(III) are assumed to hold in this section.

A. Auxiliary results

Let ζ,E be such that the condition

$$E > \sqrt{\eta^2 + 1} + V(\rho,\xi), \tag{4.1}$$

necessary for the inequality $p_1(0;\zeta,E) > 0$ in (2.6) to hold for $H = H_{\rm R}$, is obeyed. Then by familiar results, such a solution $x_i(t;\zeta,E)$, $p_i(t;\zeta,E)(i = 1,2)$ exists and is unique at each t in some maximum open interval $I'(\zeta,E)\subset\mathbb{R}$ containing the point $t = 0$. At each such ζ,E, the functions $x_i(t;\zeta,E)$ satisfy the integral equations

$$x_i(t;\zeta,E) = \xi_i + \int_0^t \frac{1}{E - V(x(s;\zeta,E))} \times\left[\eta_i + \int_0^s f_i(x(r;\zeta,E))dr\right] ds, \quad i = 1,2, \tag{4.2}$$

if $t\in I'(\zeta,E)$. In (4.2), we have used the notation $x(s;\zeta,E) = (x_1(s;\zeta,E), x_2(s;\zeta,E))$ as well as the notations (3.3) and (3.4), except that now η_1 is defined by

$$\eta_1 = \eta_1(\zeta,E) = \sqrt{[E - V(\rho,\xi)]^2 - \eta^2 - 1}. \tag{4.3}$$

Notice that (4.1) holds under the conditions of the following lemma.

Lemma 4.1: Choose positive constants a,b,b',c, with a,b as in Lemma 3.1 and $c > b' + V_1$, where V_1 was defined in (3.6) [see also (3.5)]. In addition, let E_1' be a sufficiently large constant satisfying

$$E_1' > V_1 + \sqrt{b^2 + 1}$$

in particular. Then at each $(\zeta,(t,E))\in\mathscr{U}\times\mathscr{V}'\subset\mathbb{R}^4$ the solution $x_i(t;\zeta,E)$, $p_i(t;\zeta,E)(i = 1,2)$ considered in this section exists, is unique, and is such that each of these four functions is analytic in t, ξ,η,E. Here $\mathscr{U}$ is as in (3.7) and

$$\mathscr{V}' = \{(t,E)\in\mathbb{R}^2: E > E_1', |t| < t_0'\},$$

where t_0' is the unique positive root of

$$\tfrac{1}{2}Nt^2 + (3c + 2V_1)t - \nu(E_1' - V_1)(a_0 - a) = 0,$$

with N as in (3.6) and with $\nu\in(0,1)$ a constant close enough to unity. This solution has the property

$$(x_1(t;\zeta,E), x_2(t;\zeta,E))\in\mathscr{R}\cap\mathbb{R}\subset\mathscr{S}, \quad \text{if } (\zeta,(t,E))\in\mathscr{U}\times\mathscr{V}',$$

where $\mathscr{R}$ and $\mathscr{S}$ are as in (3.5) and (2.2), respectively.

Remark: Notice that t_0' is independent of E for $E > E_1'$ and that $t_0' = O((E_1')^{1/2})$, $E_1'\to\infty$.

Proof of Lemma 4.1: Similar to that of Lemma 3.1. ■

Reasons analogous to those mentioned in Sec. III B in a similar connection show that for every $\zeta\in\mathscr{U}$ and $E > E_2' > E_1'$, where E_2' is a sufficiently large constant, the function $t\mapsto u$ with

$$u = x_1(t;\zeta,E)$$

[where, of course, $x_1(t;\zeta,E)$ is understood in the sense $H = H_{\rm R}$] maps $(-t_0',t_0')$ bijectively onto an open interval $J'(\zeta,E)\subset\mathbb{R}$ containing $[\rho,\rho+1]$. They also show that $t = h(u,\zeta,E)$ for $u\in J'(\zeta,E)$ at each such ζ,E, where at these values of its arguments h is analytic in u,ξ,η,E, and strictly increasing in u.

Hence arguments of the same type as the relevant ones of Sec. III A prove that there exists a unique time $\tau(\zeta,E)$ at which the solution curve of system (2.5) for $H = H_{\rm R}$, emerging at $t = 0$ from a point $(x_1 = \rho, x_2 = \xi, p_2 = \eta)$ in the intersection of the surface of section $x_1 = \rho$ with the energy surface $H_{\rm R} = E$, intersects the surface of section $x = \rho + 1$. That is, Eq. (3.10), understood in the context $H = H_{\rm R}$, has such a unique solution for $\zeta\in\mathscr{U}$, $E > E_2'$. We mention in passing that this solution has the property

$$\tau(\zeta,E) = 1 + O(E^{-2}),$$

for $E\to\infty$, uniformly in ζ on $\mathscr{U}$, a result which will not be used in this paper.

The diffeomorphism $t\mapsto u$ defined in this subsection allows us to construct the integral Eqs. (4.5a) and (4.5b) below, which are the relativistic counterparts of (3.13a) and (3.13b), respectively. Our procedure for estimating the derivatives (2.10) in the case $H = H_{\rm R}$ is of the same type as that expounded in Sec. III A in the case $H = H_{\rm NR}$.

Define the functions X', P', P_1' as follows in terms of the present solution $x_i(t;\zeta,E)$, $p_i(t;\zeta,E)(i = 1,2)$:

$$X'(u;\zeta,E) = x_2(t;\zeta,E), \tag{4.4a}$$

$$P'(u;\zeta,E) = p_2(t;\zeta,E), \tag{4.4b}$$

$$P_1'(u;\zeta,E) = p_1(t;\zeta,E) = \sqrt{[E - V(u,X'(u;\zeta,E))]^2 - P'^2(u;\zeta,E) - 1}, \tag{4.4c}$$

for $u \in J'(\zeta,E)$, $\zeta \in \mathscr{U}$, $E < E_2$. That X', P', P'_1 are analytic in u,ξ,η,E at these values follows for reasons analogous to those adduced to prove the corresponding property of X,P,P_1 in Sec. III A.

The derivation of the desired integral equations obeyed by X' and P' is similar to that of (3.13); namely, these equations, analogous to (3.12), follow immediately by writing the relevant equations of type (2.5) for $H = H_{\rm R}$ in terms of the pertinent new dependent and independent variables and taking account of the appropriate initial conditions. These integral equations are

$$X'(u;\zeta,E) = \xi + \int_\rho^u \frac{P'(u';\zeta,E)}{P'_1(u';\zeta,E)}\,du', \tag{4.5a}$$

$$P'(u;\zeta,E) = \eta + \int_\rho^u \frac{[E - V(u';X'(u';\zeta,E))]}{P'_1(u';\zeta,E)} \times f_2(u',X'(u';\zeta,E))\,du', \tag{4.5b}$$

and obtain at each u,ζ,E specified in the previous paragraph.

B. Asymptotic estimates for $\mathscr{P}_{\rm R}(E)$ and proof of Theorem 1 for $H = H_{\rm R}$

The contents of this subsection are analogous to those of Secs. III B–III D.

We proceed to define a map $\mathscr{P}_{\rm R}(E)$ which is a precise version of the map $\mathscr{P}(E)$ of Sec. II B appropriate to the case $H = H_{\rm R}$. Analogously to Eqs. (3.14), we have in this case

$$X'(\rho+1;\zeta,E) = x_2(\tau(\zeta,E);\zeta,E),$$

$$P'(\rho+1;\zeta,E) = p_2(\tau(\zeta,E);\zeta,E),$$

for $\zeta \in \mathscr{U}$, $E > E'_2 \geqslant E'_1$, where $\mathscr{U}$ E'_1, E'_2 are as in (3.7), Lemma 4.1, and the first sentence of the paragraph immediately after the proof Lemma 4.1. Thus, $\mathscr{P}_{\rm R}(E)$ can be defined as follows.

Definition: For each $E > E'_2$, $\mathscr{P}_{\rm R}(E)$ is the mapping with domain $\mathscr{U} \subset \mathbb{R}^2$ which sends every $(\xi,\eta) \in \mathscr{U}$ into

$$(\xi',\eta') = (X'(\rho+1;\zeta,E), P'(\rho+1;\zeta,E)) \in \mathbb{R}^2.$$

Remark: By Lemmas 4.1 and 4.2, and the previously cited theorem in the book by Arnold and Avez,[21] $\mathscr{P}_{\rm R}(E)$ is well defined for $E > E'_2$ and has the properties ascribed to $\mathscr{P}_{\rm NR}(E)$ in Remarks (1) and (2) of Sec. III B (fourth and fifth paragraphs), provided that E_2, $H_{\rm NR}$, $\mathscr{P}_{\rm NR}(E)$ are replaced by E'_2, $H_{\rm R}$, $\mathscr{P}_{\rm R}(E)$, respectively.

The next two lemmas concern the linear part of $\mathscr{P}_{\rm R}(E)$, specified by the functions $\alpha,\beta,\gamma,\delta$ which are given by

$$\alpha(E) = \frac{\partial X'(\rho+1;0,E)}{\partial \xi}, \quad \beta(E) = \frac{\partial P'(\rho+1;0,E)}{\partial \xi}, \tag{4.6}$$

$$\gamma(E) = \frac{\partial P'(\rho+1;0,E)}{\partial \xi}, \quad \delta(E) = \frac{\partial P'(\rho+1;0,E)}{\partial \eta},$$

and are analytic in E for $E > E_2$ by the analyticity properties of X' and P'.

Lemma 4.2: One has for the functions $\alpha,\beta,\gamma,\delta$ in (4.6)

$$\alpha(E) = 1 + \frac{1}{E}\int_\rho^{1+\rho} (1+\rho-u)\,\frac{\partial f_2(u,0)}{\partial x_2}\,du + O(E^{-2}), \tag{4.7a}$$

$$\beta(E) = 1/E + O(E^{-2}), \tag{4.7b}$$

$$\gamma(E) = \int_0^1 \frac{\partial f_2(u,0)}{\partial x_2}\,du + O(E^{-1}), \tag{4.7c}$$

$$\delta(E) = 1 + \frac{1}{E}\int_\rho^{1+\rho} (u-\rho)\,\frac{\partial f_2(u,0)}{\partial x_2}\,du + O(E^{-2}). \tag{4.7d}$$

Proof: It is similar to that of Lemma 3.2. We will sketch the proof of (4.7a) and (4.7c) very briefly; the proofs of (4.7b) and (4.7d) are analogous.

(1) The equation

$$\frac{\partial X'(u;0,E)}{\partial \xi} = 1 + O(E^{-1}) \tag{4.8}$$

holds for $E \to \infty$, uniformly with respect to u on $[\rho,(\rho+1)]$. To prove this, one invokes the fact that

$$P'_1(u;\zeta,E) = E + O(1),$$

for $E \to \infty$ in this same uniform sense, as follows by arguments similar to those adduced to establish (3.9′). One also uses (4.5), the fact that (3.18) is true for $u \in \mathbb{R}$, $E > E'_2$ if X,P are replaced by X',P', respectively, and either a standard Gronwall estimate or the more pedestrian approach used to derive (3.19).

(2) By the fact that (4.8) holds in the above uniform sense and an approach patterned on Step (2) of the proof of Lemma 3.2, we easily arrive at (4.7a) and (4.7c). ■

Let $M_{\rm R}(E)$ be the matrix (2.9), with elements defined by (4.6). We have the following lemma.

Lemma 4.3: The eigenvalues of $M_{\rm R}(E)$ can be labeled to have properties (i) and (ii) for $E > E'_3$, where $E'_3 \geqslant E'_2$ is a sufficiently large constant.

Proof: Similar to that of Lemma 3.3. ■

In this section, λ will denote a function such that $\lambda(E)$, $\overline{\lambda(E)}$ are eigenvalues of $M_{\rm R}(E)$ having properties (i) and (ii) at sufficiently high E, and which in addition is such that $\operatorname{Im}\lambda(E) > 0$ for large enough E. Arguments analogous to those adduced to derive (3.23) yield the following asymptotic formulas for the present $\lambda(E)$:

$$\begin{aligned} \operatorname{Re}\lambda(E) &= 1 - A_2/2E + O(E^{-2}),\\ \operatorname{Im}\lambda(E) &= (A_2/E)^{1/2} + O(E^{-3/2}). \end{aligned} \tag{4.9}$$

Asymptotic estimates for the derivatives

$$\frac{\partial^n X'(u;0,E)}{\partial \xi^m\,\partial \eta^{n-m}}, \quad \frac{\partial^n P'(u;0,E)}{\partial \xi^m\,\partial \eta^{n-m}}, \qquad m = 1,\ldots,n, \quad \rho \leqslant u \leqslant \rho+1, \tag{4.10}$$

can be derived for any $n \geqslant 1$ by a systematic inductive procedure analogous to that by which estimates were obtained for the derivatives (3.24). In this way we find that each derivative in the first (second) line of (4.10) is $O(E^{-p(n,m)})$ $(O(^{-q(n,m)}))$ as $E \to \infty$, uniformly in u on $[\rho,\rho+1]$, where $p(n,m)$, $q(n,m)$ are appropriate non-negative integers.

Such rough uniform estimates of the derivatives (4.10) for $n = 1,2,3$, together with arguments patterned on Step (2) of the proof of Lemma 3.2 lead to the following results for the relevant nonlinear portions of $\mathscr{P}_{\rm R}(E)$.

Lemma 4.4: Using the notation (2.4c), one has for the case $H = H_{\rm R}$ the following.

(1) The quadratic nonlinearities of $\mathscr{P}_{\rm R}(E)$ behave as

$$\frac{\partial^2 X'(\rho+1;0,E)}{\partial\xi^{2-s}\,\partial\eta^s} = O(E^{-1-s}), \quad s = 0,1,2, \tag{4.11a}$$

$$\frac{\partial^2 P'(\rho+1;0,E)}{\partial\xi^2} = -A_3 + O(E^{-1}), \tag{4.11b}$$

$$\frac{\partial^2 P'(\rho+1;0,E)}{\partial\xi^{2-s}\,\partial\eta^s} = O(E^{-s}), \quad s = 1,2. \tag{4.11c}$$

(2) The cubic nonlinearities of $\mathscr{P}_{\rm R}(E)$ behave as

$$\frac{\partial^3 X'(\rho+1;0,E)}{\partial\xi^{3-s}\,\partial\eta^s} = O(E^{-1-s}), \quad s = 0,1,2, \tag{4.12a}$$

$$\frac{\partial^3 X'(\rho+1;0,E)}{\partial\eta^3} = O(E^{-3}), \tag{4.12b}$$

$$\frac{\partial^3 P'(\rho+1;0,E)}{\partial\xi^3} = -A_4 + O(E^{-1}), \tag{4.12c}$$

$$\frac{\partial^3 P'(\rho+1;0,E)}{\partial\xi^{3-s}\,\partial\eta^s} = O(E^{-s}), \quad s = 1,2,3. \tag{4.12d}$$

We summarized the relevant aspects of the reduction of $\mathscr{P}_{\rm NR}(E)$ to Birkhoff normal form in the respective paragraphs containing Eqs.(3.28) and (3.30). The definitions and other statements made there apply after obvious changes to the corresponding reduction of $\mathscr{P}_{\rm R}(E)$. One merely has to replace $\mathscr{P}_{\rm NR}(E)$, $M_{\rm NR}(E)$, E_3, E_4 by $\mathscr{P}_{\rm R}(E)$, $M_{\rm R}(E)$, E_3', E_4', respectively, where $E_4' \geqslant E_3'$ is a sufficiently large constant and E_3' is as in Lemma 4.3, and to interpret $\mathscr{Q}_E$, ψ_E, $\mathscr{U}_E$, λ, μ, and the other mathematical symbols occurring in the latter two paragraphs within the context and definitions of the present section. Thus, in the present relativistic context, the functions c_1, c_2 are analytic functions of E for high enough E which satisfy Eqs. (3.33) and (3.34), with $M_{\rm NR}(E)$ replaced by $M_{\rm R}(E)$ and with $\lambda(E)$ as defined in this section. Arguments analogous to those used to derive (3.35) yield for the present c_1, c_2

$$\begin{aligned} |c_1(E)| &= 2^{-1/2}A_2^{-1/4}E^{-1/4} + O(E^{-5/4}),\\ |c_2(E)| &= 2^{-1/2}A_2^{1/4}E^{1/4} + O(E^{-3/4}). \end{aligned} \tag{4.13}$$

Consider now the first twist coefficient $\mu(E)$ in the Birkhoff normal form (3.29) pertaining to $\mathscr{P}_{\rm R}(E)$. More precisely, for large enough E this function μ is analytic, real valued, and defined by the version of (3.30) appropriate to the case $H = H_{\rm R}$, in the sense of the last paragraph. The next lemma gives a more explicit version of (2.12b) for this μ.

Lemma 4.5: We have

$$\mu(E) = \tfrac{1}{8}A_2^{-2}(A_2A_4 - \tfrac{5}{3}A_3^2)E^{-1} + O(E^{-3/2}),$$

in terms of the notation (2.4c).

Proof: Taking into account (4.9) and (4.11)–(4.13), as well as the paragraph containing (4.13), the present lemma follows by arguments of the type of those used to prove Lemma 3.5. ■

Proof of Theorem 1 for $H = H_{\rm R}$: One uses Lemma 4.5, Theorem A1, and arguments which are almost a verbatim repetition of those stated in Sec. III D. ■

ACKNOWLEDGMENTS

It is a pleasure to thank Professor J. Ellison for a number of useful discussions. I am grateful to him and his student, H. S. Dumas, for having brought typographical errors in the text to my attention.

APPENDIX: EXISTENCE AND ANALYTICITY PROPERTIES OF INVARIANT CURVES OF ANALYTIC AREA-PRESERVING MAPS IN THE PLANE

In this Appendix, we will state the version of Moser's twist theorem used to prove Theorem 1.

Consider a family $\{M_\nu, \nu\in[a,b]\}$ of mappings, where $[a,b]\subset\mathbb{R}$. Each M_ν in this family maps every (x,y) lying in a ν-independent neighborhood $W\subset\mathbb{R}^2$ of the origin into $(x_1,y_1)\in\mathbb{R}^2$ according to

$$z_1 = \gamma_1(\nu)z[1 + i\gamma_2(\nu)|z|^2] + S(z,\bar z,\nu),$$

where $z = x + iy$, $z_1 = x_1 + iy_1$. For $\nu\in[a,b]$, γ_1 and γ_2 are analytic functions of ν with the properties

$$|\gamma_1(\nu)| = 1, \quad 0\neq\gamma_2(\nu)\in\mathbb{R},$$

and S is analytic in $z,\bar z,\nu$ for $((x,y),\nu)\in W\times[a,b]$ and such that

$$S(z,\bar z,\nu) = O(|z|^4), \quad z\to 0,$$

at each $\nu\in[a,b]$. Moreover, at all such ν every circle in W centered at the origin intersects its image under M_ν.

Theorem A1 (Moser's Twist Theorem): Under the assumptions on the family $\{M_\nu, \nu\in[a,b]\}$ in the preceding paragraph, there exists for each $\epsilon > 0$ and $\nu\in[a,b]$ an invariant closed curve of M_ν of the form

$$x = p(\xi,\nu), \quad y = q(\xi,\nu), \quad \xi\in\mathbb{R}, \tag{A1}$$

lying in the intersection of W with the punctured disk $0 < x^2 + y^2 < \epsilon^2$ in the plane. Here p,q are real-valued analytic functions of ξ,ν for $(\xi,\nu)\in\mathbb{R}\times[a,b]$ which are 2π-periodic in ξ over $\mathbb{R}$ at each $\nu\in[a,b]$.

Proof: Except for the analyticity of the invariant curves (A1) in ν, this theorem is essentially a special case of well-known results.[25] This analyticity property follows, in particular, from the fact that these curves were constructed by a variant[26] of the rapidly convergent iteration method of Kolmogorov, which converges uniformly in an appropriate sense. ■

Remarks: (1) The continuous dependence of the above invariant curves on ν entailed by Theorem A1 is important for proving orbital stability of periodic orbits of suitable Hamiltonian dynamical systems of two degrees of freedom.[27] This continuity property is used in the proof of the orbital stability assertions of Theorem 1 given in Sec. III D.

(2) It is well known[28] that for each $\nu\in[a,b]$ there is a Cantor set $\mathscr{F}_\nu$ of invariant curves of M_ν of the form (A1), such that the ratio of the plane Lebesgue measure of the set of points in each punctured disk $0 < x^2 + y^2 < \epsilon^2$ which lies on a curve of $\mathscr{F}_\nu$ to the measure $\pi\epsilon^2$ of this disk approaches unity as $\epsilon\to 0$. This result will not be used here.

[1]D. S. Gemell, Rev. Mod. Phys. **46**, 129 (1974).
[2]C. Lehman, *Interaction of Radiation and Elementary Defect Production* (North-Holland, Amsterdam, 1977).
[3]J. Stark, Phys. Z. **13**, 973 (1912).
[4]M. T. Robinson and O. S. Oen, Phys. Rev. **132**, 2385 (1963).
[5]A summary of earlier results of the author for nonrelativistic channeling models in two and three space dimensions was given by A. W. Sáenz, Phys. Lett. A **93**, 337 (1983). The following errata are present in that paper: (a) in p. 338, left column, third line of last paragraph, replace "spectra" by "space"; (b) in p. 339, right column, replace second ∞ symbol in Eq. (10) by b; and (c) in p. 339, right column, second line after Eq. (10), replace "(11)" by "(10)."
[6]All statements in this paper asserting the analyticity of some real- or complex-valued function with respect to a certain set of real or complex variables should be understood as analyticity in all these variables *jointly*.
[7]C. L. Siegel and J. K. Moser, *Lectures on Celestial Mechanics* (Springer, Berlin, 1971), pp. 127 and 128.
[8]J. K. Moser, Nachr. Akad. Wiss. Göttingen, Math. Phys. Kl. II, 1 (1962).
[9]H. Rüssmann, Nachr. Akad. Wiss. Göttingen, Math. Phys. Kl. II, 67 (1970).
[10]S. Sternberg, *Celestial Mechanics, Part II* (Benjamin, New York, 1969), Chap. III, Sec. 11, and bibliography.
[11]J. K. Moser, *Stable and Random Motions in Dynamical Systems, with Special Emphasis on Celestial Mechanics, Ann. Math. Studies No. 77* (Princeton U.P., Princeton, NJ, 1973), Chap. II, Sec. 4a, and bibliography.
[12]R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin/Cummings, Reading, MA, 1978), Sec. 8.3 and bibliography.
[13]Important mathematical work on invariant curves of twist maps has been done recently by S. Aubry, M. R. Herman, A. Katok, J. N. Mather, H. Rüssmann, and others. See, e.g., the review articles by A. Katok, "Dynamical Systems and Chaos," in *Lecture Notes in Physics*, Vol. 179, edited by L. Garrido (Springer, Berlin, 1983), p. 47; and the review article by S. Aubry, Physica D **7**, 240 (1983). See also the bibliography in R. S. Mackay, Physica D **7**, 283 (1983).
[14]A. Nagl (Department of Physics, Catholic University, Washington, D.C.) and the present author are making a numerical surface-of-section study of regular and chaotic motions of fast particles in two-dimensional crystals, in the context of a special case of the NR model considered in this paper.
[15]In January 1983, J. Ellison (Department of Mathematics and Statistics, University of New Mexico, Albuquerque) informed me of unpublished work, done collaboratively with J. Su and C. Seal, on regular and chaotic motions occurring for a continuum model of axial channeling. This appears to be the earliest numerical investigation of such motions by the surface-of-section method.
[16]A version of this generalized theorem is announced in my paper "Rigorous Results on Channeling Stability in Crystals via Canonical Maps," in *Proceedings of the Thirteenth International Colloquium on Group Theoretical Methods in Physics*, edited by W. W. Zachary (World Scientific, Singapore, 1985) (in press).
[17]The canonical (= symplectic) maps referred to in this paper are obvious two-dimensional local versions of the corresponding maps defined, e.g., in V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, Berlin, 1978), p. 239. Such canonical maps in the plane preserve *oriented* areas.
[18]See, e.g., W. Klingenberg, *Lectures on Closed Geodesics* (Springer, Berlin, 1977), Lemma 3.3.2, pp. 101, 102. Needless to say, the critical comments of R. Bott, Bull. Am. Math. Soc. **7**, 331 (1982) (see especially p. 347) about this book do not refer to portions of the book mentioned in this paper.
[19]P. Swinnerton-Dyer, Proc. London Math. Soc. **34**, 385 (1977).
[20]For the general theory, see, e.g., E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies* (Dover, New York, 1944), 4th ed., Chap. XII, Sec. 141, and A. Wintner, *The Analytical Foundations of Celestial Mechanics* (Princeton U.P., Princeton, NJ, 1941), especially Secs. 180–182.
[21]See, e.g., V. I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics* (Benjamin, New York, 1968), Theorem A 31.2, p. 231. Notice that the discussion in Siegel and Moser,[7] Sec. 22, ending in the sentence containing their Eq. (5), proves that the Poincaré map associated with a periodic orbit of a sufficiently smooth Hamiltonian system of two degrees of freedom is area preserving, but not that it is orientation preserving, and hence canonical in the present sense.
[22]See, e.g., Siegel and Moser,[7] Sec. 23 and Klingenberg,[18] Sec. 3.3, pp. 100–103.
[23]R. C. Churchill, M. Kummer, and D. L. Rod, J. Diff. Eq. **49**, 359 (1983), Sec. 5, and especially Theorem 5.3.
[24]See, e.g., Klingenberg,[18] Theorem 3.3.A.1, p. 116.
[25]Siegel and Moser,[7] Secs. 32, 33.
[26]Siegel and Moser,[7] Sec. 32.
[27]Siegel and Moser,[7] pp. 243, 248.
[28]See, e.g., Siegel and Moser,[7] p. 245. See also Theorem 12.2, p. 125 of Part II of Sternberg's celestial mechanics notes.[10]

Second-order corrected Hadamard formulas

L. N. Epele, H. Fanchiotti, and C. A. García Canal
Laboratorio de Física Teórica, Departamento de Física, Universidad Nacional de La Plata, C. C. 67-La Plata 1900, Argentina

(Received 24 January 1983; accepted for publication 6 December 1984)

The second-order correction to the Hadamard formulas for the Green's function, harmonic measures, and period matrix of a two-dimensional domain is obtained in the context of the domain-variational theory.

I. INTRODUCTION

The main problem of the domain-variational theory[1] is given the domain functions of a multiply connected domain, compute the functions corresponding to a new domain obtained by varying the original one. The usefulness of this method in pure and applied physic problems is self-evident. For example, if we know how to solve the Laplace equation for the case of boundaries of a given geometry, the corresponding domain functions can be used as an input in the domain-variational-theory calculation whenever asymmetries or different geometries are present.

The first-order corrections to the Green's function, the harmonic measures, and the corresponding period matrix of a given two-dimension domain are known as Hadamard formulas.[2] The purpose of this paper is to present compact and consistent expressions for the second correction to these domain functions. In this way we have insight on the convergence of the so-defined iterative procedure to compute higher-order domain-variational-theory contributions.

II. FORMAL DEVELOPMENTS

Our first objective is to study the behavior of a Green's function defined in a domain D when this domain is subject to slight variations to become the new one D^*. From now on we will indicate with an asterisk all the quantities referred to in the modified domain D^*. The original domain, that could be multiply connected, is assumed, as usual, to be bounded by closed analytic curves. As a result of this assumption, the Green's function of D, $g(z,\zeta)$, is harmonic on all the points of the curve Γ, the boundary of D. Now D^* is bounded by Γ^*, obtained from Γ through a slight deformation: at every point $z(s)$ of Γ we construct the normal distance $\delta n(s)$ between $z(s)$ and the intersection $z^*(s)$ of this normal with Γ^*. Conventionally, $\delta n(s)$ will be taken positive at points at which Γ is pushed outward and negative otherwise. The condition of small deformations is that $\delta n(s) = \epsilon f(s)$ is a bounded function and ϵ is a positive number as small as to guarantee the uniqueness of $z^*(s)$.

By using the standard Green's formula, the fact that $g^*(z,\zeta) - g(z,\zeta)$ is harmonic in D^*, and introducing the notation

$$D_D[u(z=x+iy),v(z)] = \iint_D \left[\frac{\partial u(z)}{\partial x}\frac{\partial v(z)}{\partial x} + \frac{\partial u(z)}{\partial y}\frac{\partial v(z)}{\partial y}\right] dx\,dy\,, \qquad (1)$$

it can be easily proved that

$$g^*(z,\zeta) - g(z,\zeta) = (1/2\pi)D_{D-D^*}[g(\eta,\zeta), g(\eta,z)] - (1/2\pi)D_{D^*}[g^*(\eta,\zeta) - g(\eta,\zeta),\ g^*(\eta,z) - g(\eta,z)], \qquad (2)$$

which is the starting point of the above-mentioned iterative procedure.

Let us consider the first term of the right-hand side of Eq. (2). Introducing local coordinates (n,s), where n is measured along the normal to Γ and s along this curve, that term can be written as

$$A = \frac{1}{2\pi} D_{D-D^*}[g(\eta,\zeta), g(\eta,z)] = -\frac{1}{2\pi}\int_\Gamma ds \int_{\delta n(s)}^0 dn\, J(n,s)\left[\frac{\partial g(\eta,\zeta)}{\partial x}\frac{\partial g(\eta,z)}{\partial x} + \frac{\partial g(\eta,\zeta)}{\partial y}\frac{\partial g(\eta,z)}{\partial y}\right], \qquad (3)$$

where

$$J(n,s) = 1 + n/R(s)\,, \qquad (4)$$

where $R(s)$, being the radius of curvature of Γ at the point s, is the Jacobian related to the local coordinates.

The integrand in Eq. (3) can be referred to the border value ($n=0$) of the functions involved (hereafter indicated by ^{0}). In so doing one is keeping contributions up to first order in n because it is enough in order to collect the terms up to second order in the variation $n(s)$. In this way the result is

$$A = -\frac{1}{2\pi}\int_\Gamma ds_\eta \int_{\delta n(s)} dn_\eta \left[1 + \frac{n}{R(s)}\right]\left\{\frac{\partial g^0(\eta,\zeta)}{\partial n_\eta}\frac{\partial g^0(\eta,z)}{\partial n_\eta} + \frac{\partial}{\partial n_\eta}\left[\frac{\partial g(\eta,\zeta)}{\partial n_\eta}\frac{\partial g(\eta,z)}{\partial n_\eta}\right]^0 n\right\} \qquad (5)$$

and performing the integration in n one obtains

$$A = \frac{1}{2\pi}\int_\Gamma ds_\eta \frac{\partial g^0(\eta,\zeta)}{\partial n_\eta}\frac{\partial g^0(\eta,z)}{\partial n_\eta}\delta n(s) + \frac{1}{4\pi}\int_\Gamma ds_\eta \left[\frac{1}{R(s)} + \frac{\partial}{\partial n_\eta}\right] \times \left[\frac{\partial g(\eta,\zeta)}{\partial n_\eta}\frac{\partial g(\eta,z)}{\partial n_\eta}\right]^0 [\delta n(s)]^2\,. \qquad (6)$$

In order to obtain the contribution coming from the

 0022-2488/85/081936-02$02.50

second term of Eq. (2) we recast in it the order $\delta n(s)$ of A above to obtain

$$B = -\frac{1}{2\pi} D_D \cdot [g^*(\eta, \zeta) - g(\eta, \zeta), g^*(\eta, z) - g(\eta, z)]$$
$$= \frac{1}{(2\pi)^2} \int_\Gamma ds_\tau \, \delta n(s_\tau) \frac{\partial g^0(\tau, \zeta)}{\partial n_\tau}$$
$$\times \int_\Gamma ds_\mu \, \delta n(s_\mu) \frac{\partial g^0(\mu, z)}{\partial \eta_\mu} \frac{\partial^2 g^0(\tau, \mu)}{\partial n_\tau \, \partial n_\mu}, \tag{7}$$

where we have used

$$D_D \cdot [g(\tau,\eta), g(\mu,\eta)] = -2\pi g(\tau,\mu) + O(\delta n(s)), \tag{8}$$

valid whenever τ and μ are exterior to D^*.

Going back to Eq. (2) with A and B given by Eqs. (6) and (7), respectively, we have the desired first- and second-order corrections to the Green's function. Notice that the first-order contribution in $n(s)$ comes from A and is precisely the well-known Hadamard's formula.

Having arrived at the results (6) and (7) it is simple to obtain the corresponding corrections, up to second order in $\delta n(s)$, for the harmonic measures and the period matrix of the domain.

The harmonic measure $\omega_\nu(\zeta)$ of the domain D bounded by the closed curves Γ_i $(i = 1,...,n)$ (notice $\Gamma = \Gamma_1 + \cdots + \Gamma_n$) is harmonic in D and has the values 1 on Γ_ν and 0 on Γ_λ $(\lambda \neq \nu)$, respectively, and can be written as

$$\omega_\nu(\zeta) = -\frac{1}{2\pi} \int_{\Gamma_\nu} \frac{\partial g^0(z, \zeta)}{\partial m_z} ds_z . \tag{9}$$

Then, the variation of $\omega_\nu(\zeta)$ when D is changed to D^* in the way that was defined above, gives the result

$$\omega_\nu^*(\zeta) - \omega_\nu(\zeta) = \frac{1}{2\pi} \int_\Gamma ds_\eta \frac{\partial g^0(\eta, \zeta)}{\partial n_\eta} \cdot \frac{\partial \omega_\nu^0(\eta)}{\partial n_\eta} \delta n(s_\eta)$$
$$+ \frac{1}{4\pi} \int_\Gamma ds_\eta \left[\frac{1}{R(s)} + \frac{\partial}{\partial n_\eta} \right]$$
$$\times \left[\frac{\partial g(\eta, \zeta)}{\partial n_\eta} \frac{\partial \omega_\nu(\eta)}{\partial n_\eta} \right]^0 [\delta n(s)]^2$$
$$+ \frac{1}{(2\pi)^2} \int_\Gamma ds_\tau \, \delta n(s_\tau) \frac{\partial g^0(\tau, \zeta)}{\partial n_\tau}$$
$$\times \int_\Gamma ds_\mu \, \delta n(s_\mu) \frac{\partial \omega_\nu^0(\mu)}{\partial n_\mu} \cdot \frac{\partial^2 g^0(\tau, \mu)}{\partial n_\tau \, \partial n_\mu}. \tag{10}$$

Finally, the period matrix, whose elements are the periods of $\omega_\nu(\zeta)$ with respect to the boundary Γ_λ, i.e.,

$$P_{\lambda\nu} = \int_{\Gamma_\lambda} \frac{\partial \omega_\nu^0(\zeta)}{\partial n_\zeta} ds_\zeta , \tag{11}$$

suffers the corresponding variation given by

$$P_{\lambda\nu}^* - P_{\lambda\nu} = -\int_\Gamma ds_\eta \frac{\partial \omega_\lambda^0(\eta)}{\partial n_\eta} \frac{\partial \omega_\nu^0(\eta)}{\partial n_\eta} \delta n(s_\eta)$$
$$- \frac{1}{2} \int_\Gamma ds_\eta \left[\frac{1}{R(s)} + \frac{\partial}{\partial n_\eta} \right]$$
$$\times \left[\frac{\partial \omega_\lambda(\eta)}{\partial n_\eta} \frac{\partial \omega_\nu(\eta)}{\partial n_\eta} \right]^0 [\delta n(s_\eta)]^2$$
$$- \frac{1}{2\pi} \int_\Gamma ds_\tau \, \delta n(s_\tau) \frac{\partial \omega_\lambda^0(\tau)}{\partial n_\tau}$$
$$\times \int_\Gamma ds_\mu \, \delta n(s_\mu) \frac{\partial \omega_\nu^0(\mu)}{\partial n_\mu} \frac{\partial^2 g^0(\tau, \mu)}{\partial n_\tau \, \partial n_\mu}. \tag{12}$$

III. FINAL COMMENTS

Our main results are Eqs. (10) and (12). They can be easily handled for a variety of geometries. In particular, they look extremely simple for the case of an annulus. Nevertheless, they can be used in problems related with any connectedness. Some of the most direct applications were the calculation of characteristic impedances of coaxial lines bounded by N-regular polygons,[3] the study of transmission-line conductors of various cross sections,[4] and the implementation of a $1/N$ expansion in Dirichlet problems.[5] Moreover, applications to the study of torsion of bars of different cross sections are in progress. These examples show the wide physical interest of the formulas.

ACKNOWLEDGMENTS

We thank H. Vucetich for enjoyable discussions.

Partial support from Consejo Nacional de Investigaciones Cientificas y Têcnicas (CONICET) and Comisión de Investigaciones científicas (CIC) Provineiz de Buenos Aires is acknowledged.

[1]M. Schiffer and D. Spencer, *Functionals of Finite Riemann Surfaces* (Princeton U. P., Princeton, NJ, 1954).

[2]A. Nehari, *Conformal Mapping* (McGraw-Hill, New York, 1952).

[3]L. N. Epele, H. Fanchiotti, and C. A. García Canal, IEEE Proc. Lett. **72**, 223 (1984).

[4]L. N. Epele, H. Fanchiotti, C. A. García Canal, and H. Vucetich, Int. J. Electron. **56**, 111 (1984).

[5]L. N. Epele, H. Fanchiotti, and C. A. García Canal, J. Appl. Phys. **56**, 858 (1984).

On a particular solution of the equation of Ernst with n fields, parametrized by an arbitrary harmonic function

B. Léauté and G. Marcilhacy
Université Paris VI, Unité Associée au Centre National de la Recherche Scientifique No. 769, Institut Henri Poincaré, 11, rue Pierre et Marie Curie, 75231 Paris Cedex 05, France

(Received 28 November 1984; accepted for publication 18 January 1985)

We determine a particular solution, dependent upon an arbitrary harmonic function in cylindrical coordinates, of the system of n partial differential equations, which characterizes both the axially symmetric field solution of the Einstein $(n-1)$-Maxwell equations and one class of axially symmetric static self-dual $SU(n+1)$ Yang–Mills fields.

I. INTRODUCTION

The equivalence between symmetric gravitational fields and static axially symmetric self-dual SU(2) Yang–Mills fields is known. It was Witten[1] who was the first to show that the Ernst equation,[2] which essentially governs the former, determines also the latter in Yang's R gauge.[3] This analogy is still valid between stationary axially symmetric electrovacuum fields and a particular class of static axially symmetric self-dual SU(3) Yang–Mills fields,[4] the basic system of equations being then the Ernst system, suitable, this time, for the case with electromagnetic field.[5]

The analogy mentioned above can be pursued still further. In a recent paper,[6] Gürses has in fact shown this between the stationary axially symmetric fields verifying the Einstein $(n-1)$-Maxwell equations and a special class of the static axially symmetric self-dual $SU(n+1)$ Yang–Mills fields.

In earlier papers[7] we gave particular transcendental solutions for the first two cases. The purpose of this paper is to give a particular solution corresponding to the third case. It is first necessary to note some aspects of Ref. 6 with a view to making clear the fundamental differential system that we shall study and solve.

II. BASIC EQUATIONS

The coupled Einstein–Abelian gauge field equations are given by

$$S_{\mu\nu} = \gamma_{ab}(F^{n}{}_{\mu\alpha}F_{\mu}{}^{b\alpha} - \tfrac{1}{4}g_{\mu\nu}F^{a}{}_{\alpha\beta}F^{b\alpha\beta}), \tag{1}$$

$$F^{a\mu\nu}{}_{;\nu} = 0, \tag{2}$$

$$F^{a}{}_{\mu\nu} = \partial_\mu A^{a}{}_{\nu} - \partial_\nu A^{a}{}_{\mu}, \tag{3}$$

where γ_{ab} is a diagonal matrix which can be taken as the Kronecker symbol δ_{ab} by a correct choice of basis to the gauge potentials $A^{a}{}_{\mu}$, $a, b,... = 1, 2,..., n-1$ $(n>0)$. A semicolon denotes covariant differentiation with respect to the Riemann connection.

The space-time is stationary and axially symmetric, and its metric can be written in the form

$$ds^2 = f(dt + \omega\, d\varphi)^2 - f^{-1}[e^{2\gamma}(d\rho^2 + dz^2) + \rho^2\, d\varphi^2], \tag{4}$$

where t, ρ, z, φ are the local coordinates and the functions f, ω, γ depend only on ρ and z. It is assumed that the gauge potential one-form A^{a} has two components

$$A^{a} \equiv A^{a}{}_{\mu}\, dx^\mu = A^{a}{}_{t}\, dt + A^{a}{}_{\varphi}\, d\varphi, \tag{5}$$

where $A^{a}{}_{t}$ and $A^{a}{}_{\varphi}$ are the $A^{a}{}_{\mu}$ in the direction of time (t) and the azimuthal angle (φ) coordinates. They also are assumed to depend only on ρ and z. The Einstein (1) and the gauge (2) field equations are explicitly written by taking into account the assumptions made about the metric (4) and about the gauge fields (5). The study of the system obtained, conducted in a manner similar to Ernst in the case of Einstein–Maxwell,[5] leads to the introduction of the n complex scalar functions ϵ and Φ^{a} fulfilling the system of equations

$$f\nabla^2\epsilon = (\nabla\epsilon + 2\Phi^{b*}\nabla\Phi^{b})\cdot\nabla\epsilon, \tag{6}$$

$$f\nabla^2\Phi^{a} = (\nabla\epsilon + 2\Phi^{b*}\nabla\Phi^{b})\cdot\nabla\Phi^{a}, \tag{7}$$

where ∇, ∇^2 denote, respectively, the gradient and Laplacian operators in cylindrical coordinates (ρ, z) related to the flat tridimensional metric; the symbol $*$ denotes the complex conjugation. The summation on the repeated indices is applied to the numeration indices $a, b,...$ of the fields Φ^{a}.

The functions ϵ and Φ^{a} are defined by

$$\epsilon = f - \Phi^{a}\Phi^{a*} + i\psi, \tag{8}$$

$$\Phi^{a} = A^{a}{}_{i} + iB^{a}, \tag{9}$$

with

$$\nabla B^{a} = -\rho^{-1}f\,\mathbf{n}\wedge(\nabla A^{a}{}_{\varphi} - \omega\nabla A^{a}{}_{i}), \tag{10}$$

$$\nabla\psi = -[\rho^{-1}f^2\mathbf{n}\wedge\nabla\omega + 2\,\mathrm{Im}(\Phi^{a*}\nabla\Phi^{a})], \tag{11}$$

where $\mathbf{n}$ is the unit vector along the azimuthal direction φ and Im ( ) denotes the imaginary part.

Gürses then shows by using the theory of the harmonic mappings of the Riemann manifolds that the system [Eqs. (6) and (7)], of which we have seen the origin, also characterizes a special class of axially symmetric static self-dual $SU(n+1)$ Yang–Mills fields.

We can change the formulation of the system [Eqs. (6) and (7)] by putting

$$\epsilon = (\xi + 1)/(\xi + 1) \quad \text{and } \Phi^{a} = \chi^{a}/(\xi + 1), \tag{12}$$

which yields

$$\Lambda\nabla^2\xi = 2(\xi^*\nabla\xi + \chi^{b*}\nabla\chi^{b})\cdot\nabla\xi, \tag{13}$$

$$\Lambda\nabla^2\chi^{a} = 2(\xi^*\nabla\xi + \chi^{b*}\nabla\chi^{b})\cdot\nabla\chi^{a}, \tag{14}$$

with

$$\Lambda \equiv \xi\xi^* + \chi^{b}\chi^{b*} - 1.$$

Now indicating by ζ^{a} the n scalar complex functions

 0022-2488/85/081938-04$02.50

$$\zeta^1 = \xi \quad \text{and} \; \zeta^{a+1} = \chi^a, \quad \text{when} \; a > 1,$$

we arrive finally at the following compact formulation:

$$\Lambda \nabla^2 \zeta^a = 2(\zeta^{b*} \nabla \zeta^b) \cdot \nabla \zeta^a, \tag{15}$$

with $\Lambda \equiv \zeta^b \zeta^{b*} - 1$ and $a \in [1,...,n]$. We could refer to Eq. (15) as an n-field generalization of the Ernst equation.

III. PARTICULAR SOLUTION PARAMETRIZED BY A HARMONIC FUNCTION

We propose to determine a particular class of solutions ζ^a of the system (15) dependent upon an arbitrary harmonic function $v(\rho,z)$. Let

$$\zeta^a = \zeta^a[v(\rho,z)], \quad \text{with} \; \nabla^2 v = 0. \tag{16}$$

With this assumption the system (15) may be written

$$\Lambda \zeta^{a''} = 2\zeta^{b*} \zeta^{b'} \zeta^{a'}, \tag{17}$$

where $\zeta^{a'} \equiv d\zeta^a/dv$.

The first stage of the integration of this system will be to find two families of first integrals.

Multiplying, in a contracted manner, the two members of (17) by ζ^{a*}, and taking the imaginary part of the result, we find

$$\Lambda (\zeta^{a''} \zeta^{a*} - \zeta^{a*''} \zeta^a) = 2(\zeta^{b'} \zeta^{b*} + \zeta^b \zeta^{b*'})(\zeta^{a'} \zeta^{a*} - \zeta^a \zeta^{a*'}),$$

from which we deduce the first integral

$$\zeta^{a'} \zeta^{a*} - \zeta^a \zeta^{a*'} = 2ia\Lambda^2, \tag{18}$$

where a is a first real constant of integration.

Multiplying again the two members of (17) by ζ^{a*} but *without summation* of the repeated index a—which we shall subsequently underline to point out the fact—and taking the imaginary part of the result we get

$$\begin{aligned} &\Lambda (\zeta^{\underline{a}''} \zeta^{\underline{a}*} - \zeta^{\underline{a}} \zeta^{\underline{a}*''}) \\ &\quad = 2(\zeta^{b*} \zeta^{b'} \zeta^{\underline{a}'} \zeta^{\underline{a}*} - \zeta^b \zeta^{b*'} \zeta^{\underline{a}} \zeta^{\underline{a}*'}). \end{aligned}$$

Rewriting the right member of this relation and taking into account the first integral (18), it follows that

$$\begin{aligned} &\Lambda (\zeta^{\underline{a}''} \zeta^{\underline{a}*} - \zeta^{\underline{a}} \zeta^{\underline{a}*''}) \\ &\quad = 2[\zeta^b \zeta^{b*} (\zeta^{\underline{a}'} \zeta^{\underline{a}*} - \zeta^{\underline{a}} \zeta^{\underline{a}*'}) + 2ia\Lambda^2 \zeta^{\underline{a}'} \zeta^{\underline{a}*}]. \end{aligned}$$

With the help of complex conjugation the formulation of the right member can again be changed, and in so doing we arrive at the expression

$$\begin{aligned} \Lambda (\zeta^{\underline{a}''} \zeta^{\underline{a}*} - \zeta^{\underline{a}} \zeta^{\underline{a}*''}) &= (\zeta^b \zeta^{b*})' (\zeta^{\underline{a}'} \zeta^{\underline{a}*} - \zeta^{\underline{a}} \zeta^{\underline{a}*'}) \\ &\quad + 2ia\Lambda^2 (\zeta^{\underline{a}} \zeta^{\underline{a}*})', \end{aligned}$$

which can be easily integrated to yield

$$\zeta^{\underline{a}'} \zeta^{\underline{a}*} - \zeta^{\underline{a}} \zeta^{\underline{a}*'} = 2i\Lambda (a\zeta^{\underline{a}} \zeta^{\underline{a}*} + b_{\underline{a}}). \tag{19}$$

Hence we have obtained n first integrals, characterized by n real constants, noted $b_{\underline{a}}$. The summation on $\underline{a}$ of the first integrals leads to (18), and consequently the constants a and $b_{\underline{a}}$ satisfy

$$a + \sum_{a=1}^{n} b_a = 0. \tag{20}$$

Let us now pass on to the calculation of the second family of first integrals.

Multiplying, once more, the two members of (17) by $\zeta^{\underline{a}*'}$ and taking the real part of the result, it follows that

$$\Lambda (\zeta^{\underline{a}''} \zeta^{\underline{a}*'} + \zeta^{\underline{a}*''} \zeta^{\underline{a}'}) = 2(\zeta^{b*} \zeta^{b'} + \zeta^b \zeta^{b*'}) \zeta^{\underline{a}'} \zeta^{\underline{a}*'},$$

from which, immediately, by integration, we obtain the result

$$\zeta^{\underline{a}'} \zeta^{\underline{a}*'} = \alpha_a^2 \Lambda. \tag{21}$$

Thus, we obtain again n first integrals characterized by n real constants α_a^2.

The problem of integration of the system of second-order differential equations (17) is now reduced to that of two sets of n first-order equations (19) and (21). The second stage of the solution of (17) therefore consists of this integration which will now be explained.

In order to progress it is useful to put

$$\zeta^a = R_a e^{iS_a}. \tag{22}$$

(From now on, to avoid confusion, we shall no longer underline the nonsummed repeated indices. When summations appear, they are explicitly indicated.)

It follows that the first integrals (19) and (21) may be written as

$$R_a^2 S_a' = \Lambda (aR_a^2 + b_a), \tag{23}$$

$$(R_a')^2 + (R_a S_a')^2 = \alpha_a^2 \Lambda, \tag{24}$$

with $\Lambda \equiv \Sigma_{a=1}^n R_a^2 - 1$.

The insertion of the expression S_a' from (23) into (24) leads to

$$(R_a R_a')^2 = \Lambda \; [(\alpha_a^2 - 2ab_a) R_a^2 - a^2 R_a^4 - b_a^2]. \tag{25}$$

Obviously, with regard to the solvability of the problem, it is necessary that the $2n$ constants $\{\alpha_a, b_a\}$ satisfy the inequalities

$$\alpha_a^2 - 4ab_a > 0, \tag{26}$$

for each value of numeration index $a = 1,2,...,n$. Looking at the relation (25) for two values i and j ($i \neq j$) of the index a, we obtain, after eliminating the quantity Λ^2 between them,

$$\begin{aligned} &(R_i R_i')^2 / [(\alpha_i^2 - 2ab_i) R_i^2 - a^2 R_i^4 - b_i^2] \\ &\quad = (R_j R_j')^2 / [(\alpha_j^2 - 2ab_j) R_j^2 - a^2 R_j^4 - b_j^2]. \end{aligned} \tag{27}$$

Taking out the square roots of the two members of this equation and making the change of functions defined by

$$R_a^2 = (1/2a^2)[\alpha_a (\alpha_a^2 - 4ab_a)^{1/2} Z_a + \alpha_a^2 - 2ab_a], \tag{28}$$

we arrive thus at the equation

$$\frac{Z_i'}{(1 - Z_i^2)^{1/2}} = \epsilon_{ij} \frac{Z_j'}{(1 - Z_j^2)^{1/2}}, \quad \text{with} \; \epsilon_{ij} \equiv \pm 1. \tag{29}$$

The integration is easily carried out and gives the result

$$Z_i = \epsilon_{ij} [Z_j (1 - K_{ij}^2)^{1/2} + (1 - Z_j^2)^{1/2} K_{ij}]. \tag{30}$$

We have thus obtained $n(n-1)/2$ relations, between the modulus R_a of the fields ζ^a, involving $n(n-1)/2$ real constants K_{ij} with $|K_{ij}| < 1$, taking into account (28). In particular it follows that every one of them can be expressed as a function of the first R_1; that is, by (28)

$$Z_i = \epsilon_i [Z_1 (1 - K_i^2)^{1/2} + (1 - Z_1^2)^{1/2} K_i], \tag{31}$$

with $\epsilon_{i1} \equiv \epsilon_i$ and $K_{i1} \equiv K_i$.

From this relation it results that Eq. (25), written with $a = 1$, can be reformulated in a manner which involves only the function R_1 (or Z_1); that gives

$$Z_1' = 2a\epsilon_1(1-Z_1^2)^{1/2}[A + B(1-Z_1^2)^{1/2} + CZ_1], \tag{32}$$

with the notations

$$\epsilon_1 = \pm 1, \quad A = \frac{1}{2a^2}\sum_{i=1}^{n}(\alpha_i^2 - 2ab_i) - 1,$$

$$B = \frac{1}{2a^2}\sum_{i=2}^{n}\epsilon_i\alpha_i(\alpha_i^2 - 4ab_i)^{1/2}, \tag{33}$$

$$C = \frac{1}{2a^2}\Big[\alpha_1(\alpha_1^2 - 4ab_1)^{1/2} + \sum_{i=2}^{n}\epsilon_i\alpha_i(\alpha_i^2 - 4ab_i)^{1/2}(1-K_i^2)^{1/2}\Big].$$

Equation (32) is integrated by putting $Z_1 = \sin\theta$ (Ref. 8). The expression of the result depends on the sign of $B^2 + C^2 - A^2$.

When $B^2 + C^2 - A^2 > 0$, $B \neq 0$ we obtain

$$Z_1 = 2\frac{(A-B)\{C - (B^2+C^2-A^2)^{1/2} + [C + (B^2+C^2-A^2)^{1/2}]e^{\mu}\}(e^{\mu}-1)}{(A-B)^2(e^{\mu}-1)^2 + \{C - (B^2+C^2-A^2)^{1/2} - [C + (B^2+C^2-A^2)^{1/2}]e^{\mu}\}}, \tag{34}$$

where $\mu \equiv 2\epsilon_1 a(B^2 + C^2 - A^2)^{1/2}(v - v_0)$.

When $B^2 + C^2 - A^2 < 0$, $B \neq 0$, we have

$$Z_1 = 2\frac{(A-B)[(A^2 - B^2 - C^2)^{1/2}tg\nu - C]}{(A-B)^2 + [(A^2 - B^2 - C^2)^{1/2}tg\nu - C]^2}, \tag{35}$$

where $\nu \equiv \epsilon_1 a(A^2 - B^2 - C^2)^{1/2}(v - v_0)$.

The form of the function $Z_1(v)$ thus depends on the relative values of constants A, B, C, i.e., according to (33) on those of the $2n$ integration constants $\{\alpha_a, b_a\}$ [restricted only by the n inequalities (26)] and also on the choice of the sign $+$ or $-$, symbolized by the indicator ϵ_a. Then v_0 notes the last integration constant.

Besides previous general forms of solutions there is, of course, the possibility of various particular cases; we have, for instance, the following.

(a) When $A = B$, $C \neq 0$,

$$Z_1 = 2C(e^{\mu} - A)/[C^2 + (e^{\mu} - A)^2], \tag{36}$$

where $\mu \equiv 2\epsilon_1 Ca(v - v_0)$.

(b) When $A^2 = B^2 + C^2$, $C \neq 0$,

$$Z_1 = -2(A-B)(1 + C\nu)/[(A-B)^2\nu + (1 + C\nu)^2], \tag{37}$$

where $\nu = 2\epsilon_1 a(v - v_0)$.

This integration, taking into account (28), fixes how R_1 depends on $v(\rho,z)$. In consequence of (28) and (31), all other moduli $R_a (a > 1)$ are also determined. Now it remains, in the last stage, to obtain the phases S_a. To do that we dispose of the first integrals (23), which immediately give by quadrature

$$S_a(v) = \int_{v_0}^{v} A\left(a + \frac{b_a}{R^2(u)}\right)du, \tag{38}$$

where $A \equiv \Sigma_{a=1}^{n} R_a^2(v) - 1$. The result researched is, in principle, formally obtained. With the help of formulas (23) and (25) we can write it more explicitly; we have the relation

$$\frac{dS_a}{dR_a^2} = \epsilon_a \frac{aR_a^2 + b_a}{2R_a^2[(\alpha_a^2 - 2ab_a)R_a^2 - a^2R_a^4 - b_a^2]^{1/2}}, \tag{39}$$

which shows that each phase S_a depends functionally only on modulus R_a^2. By a process of integration which we do not relate in detail here and which involves the intermediate function Z_α, defined by (28), we arrive finally at the following result:

$$S_a = \frac{\epsilon_a}{2}tg^{-1}\left\{\frac{(\alpha_a^2 - 2ab_a)[1 + (1-Z_a^2)^{1/2}] - (\alpha_a^2 - 4ab_a)Z_a^2 + \alpha_a(\alpha_a^2 - 4ab_a)^{1/2}(1-Z_a^2)^{1/2}Z_a}{Z_a[\alpha_a^2 - 2ab_a + (\alpha_a^2 - 4ab_a)(1-Z_a^2)^{1/2} + \alpha_a(\alpha_a^2 - 4ab_a)^{1/2}Z_a]}\right\} + \lambda_a, \tag{40}$$

$\{\lambda_a\}$ being an ultimate set of n constants of integration. The particular solution of the system (17) that we have investigated is now entirely known; all the sets $\{R_a, S_a\}$ giving the expressions of the n scalar complex fields ζ^a, parametrized by an arbitrary harmonic function $v(\rho,z)$, having been completely determined.

IV. CONCLUDING REMARKS

In conclusion we would like to make some comments on the above results.

(a) When there is only one scalar complex field $\zeta(v)$ we are in the vacuum gravitational case. Equation (15) is then the Ernst equation. The type of solution obtained, depending on an arbitrary harmonic function, is then the Papapetrou solution,[9] which contains as a particular case, corresponding to a specific choice of the harmonic function and to the parameters, the Newman–Unti–Tamburino metric.[10]

(b) When there are two scalar complex fields, we are in the electrovacuum case. The corresponding solution of the form envisaged here has been given by Halilsoy.[11] The similarity solutions which have been given by Kalippan and Lakshmanan[12] and by Fisher[13] were deduced by a similar process.

(c) The particular solutions considered in the foregoing paragraphs for $n = 1, 2$ are equally significant both from the point of view of the theory of general relativity and the theory of Yang–Mills. It is only the latter which is concerned for the n fields ζ^a where $n > 2$. It should be possible to obtain solutions ζ^a of the nonseparable type by the techniques[14] applicable to the equation of Ernst, because (15) constitutes a generalization for n fields of the Ernst equation.

[1]L. Witten, Phys. Rev. D **19**, 718 (1978).
[2]F. J. Ernst, Phys. Rev. **167**, 1175 (1968).
[3]C. N. Yang, Phys. Rev. Lett. **38**, 1317 (1977).
[4]M. Gürses and B. Xanthopoulos, Phys. Rev. **26**, 1912 (1982).
[5]F. J. Ernst, Phys. Rev. **168**, 1415 (1968).
[6]M. Gürses, Phys. Rev. **30**, 486 (1984).
[7]B. Léauté and G. Marcilhacy, Phys. Lett. A **93**, 393 (1983).
[8]W. Gröbner and N. Hofreiter, *Integraltafel* (Springer, Vienna, 1961).
[9]A. Papapetrou, Ann. Phys. **12**, 309 (1953).
[10]E. T. Newman, L. Tamburino, and T. J. Unti, J. Math. Phys. (NY) **4**, 915 (1963).
[11]M. Halilsoy, Lett. Nuovo Cimento **37**, 231 (1983).
[12]P. Kaliappan and M. Lakshmanan, J. Math. Phys. **22**, 2447 (1981).
[13]E. Fisher, J. Math. Phys. **23**, 1295 (1982).
[14]V. A. Belinski and V. E. Zakharov, Sov. Phys. JETP **50**, 1 (1979).

Some remarks on torsion in Kaluza–Klein unification

I. J. Muzinich[a)]
Department of Physics, Brookhaven National Laboratory, Upton, New York 11973

(Received 26 September 1984; accepted for publication 12 April 1985)

The role of manifolds endowed with a parallelizing torsion in Kaluza–Klein theories is examined. In particular the spin connection on such manifolds is demonstrated to be a single pure gauge almost everywhere on the manifold. This follows from the Frobenius integration theorem. As a consequence of this result the computation of the representation of massless fermions follows immediately and trivially. The spectrum of massless fermions on manifolds with a parallelizing torsion is contrasted with the analogous spectrum on manifolds with nontrivial topological configurations. While the remarks are primarily of pedagogic value much of the relevant mathematics is made intuitive.

I. INTRODUCTION

The old idea of Kaluza and Klein[1] on the unification of gravity with the other known gauge theories in greater than four dimensions has enjoyed a high degree of popularity in the past few years.[2,3] However, this popularity has not been matched with a complementary degree of success. The formulation of a realistic unified theory has suffered from at least four major problems[3]: (i) massless fermions (masses shielded from the Planck mass), (ii) chirally realistic fermion multiplets (complex representations of the internal symmetry group), (iii) the cosmological constant, and (iv) a well-behaved quantum theory. All of these difficulties are not necessarily distinct, and I shall comment on them during the course of this article. However, the main emphasis will be addressed to the issue of massless fermions.

First, the following general results on nonsupersymmetric Kaluza–Klein theories with fermions are recalled. Consider a generic Kaluza–Klein theory on a manifold M in $4+n$ dimensions, which is reduced to $M^4 \otimes C$ by a process called spontaneous compactification in the literature. Here C is a compact manifold and M^4 is the usual four-dimensional Minkowski space. The celebrated theorem of Lichnerowicz[4] eliminates massless fermions on any manifold C with positive scalar curvature. This includes all compact Einstein manifolds with isometry group, a compact Lie group, which is directly relevant to most candidate gauge theories. In order to circumvent the Lichnerowicz theorem one has to consider (i) Rarita–Schwinger fermions and the incumbent theories of supergravity,[3] (ii) manifolds with a parallelizing torsion,[5,6] and (iii) that other possibilities that have been advocated include composite models[7] and noncompact internal manifolds.[8]

The first possibility limits us to the maximum 11-dimensional theory allowed for $N=8$ supergravity.[3] While massless fermions are possible in such a scenario, the internal manifold C is of too small a dimension to incorporate a realistic physical model. In addition the massless fermions lie in real representations of the isometry group of the internal manifold (S^7 for 11 dimensions). Although there do not exist chiral fermions in odd dimensions, the argument is very general and makes use of the Atiyah–Hirzebruch theorem, which is applicable for all homogeneous spaces. This general result was recently communicated in this form by Witten.[9]

An analogous conclusion essentially prevails for the second possibility with a parallelizing torsion. Here massless fermions occur only in the Clifford representation of the tangent space group of the internal manifold. We present in this note a streamlined and direct proof of this statement for those manifolds that are parallelizable. If the tangent bundle for the manifold has a direct product structure $T^n(C) = C \otimes R^n$, then the manifold C is parallelizable.[10] These include all semisimple Lie group manifolds, exceptional spheres S^3 [SU(2) group manifold], S^7, and particular product manifolds $S^1 \otimes S^n$ (nonmaximally symmetric spaces).[11,12] The latter manifolds are not parallelizable in the classic Cartan–Schouten sense.[5] The proof makes use of the fact that the spin connection is a pure gauge almost everywhere in these cases and can be eliminated from the Dirac operator on C, which is nothing but the mass operator on M^4. The pure gauge property of the connection follows from the Frobenius integration theorem of classical mathematics. The vanishing of the curvature two-form (parellelizability) provides an integrability condition for the spin connection.

The phenomenological implications of all of the results are that models based upon complex representations of the chiral fermions are indeed difficult to obtain from a higher-dimensional Kaluza–Klein theory. These include fermions of the standard model and its lowest-rank grand unified extension SU(5) (see Ref. 13). In contrast models based upon orthogonal grand unification, i.e., SO(10), are possible from compactification of a 14-dimensional or greater Kaluza–Klein theory for example. The massless fermions would necessarily belong to the spinor representation of the relevant gauge group, a desirable feature for orthogonal grand unification.[11,13]

On the other hand the explicit introduction of fundamental (nongravitational) gauge fields[9,14] does allow for complex representations for chiral fermions. However, I feel that all of these models including those with torsion have to be motivated, either by the demonstration of a solution to Einstein's equations in the extra dimensions or in some other plausible dynamical context. At the present time such a motivation is lacking. A recent attempt in this direction has been made by Weinberg.[15]

[a)] Leave of absence address (after 1 September 1985): Laboratory of Nuclear Studies, Cornell University, Ithaca, New York 14853.

 0022-2488/85/081942-06$02.50

II. REVIEW OF THE KALUZA–KLEIN APPROACH

The standard mathematical introduction to the Kaluza–Klein framework is by now well known[2,3]; it will be briefly summarized here. The starting point is the usual Einstein–Hilbert action in $4+n$ dimensions:

$$S(g)=\int dx\,\sqrt{g}\,R\,. \tag{2.1}$$

This theory is assumed to have a ground state of the form $M^4\otimes C$, where C is a compact manifold with isometry group G. The physical spectrum is determined by small oscillations around this ground state. The metric which exhibits the usual massless modes, i.e., gravitons, gauge bosons, and possibly Brans–Dicke scalars, is of the form

$$g_{AB}(x,h)=\left(\begin{array}{c|c} g_{\mu\nu}(x)+A^{\;a}_{\mu}(x)\xi^{\;i}_{a}(h)A^{\;b}_{\nu}(x)\xi_{bi}(h) & A^{\;a}_{\mu}(x)\xi^{\;i}_{a}(h) \\ \hline A^{\;a}_{\nu}(x)\xi^{\;i}_{a}(h) & g_{ij}(h)\end{array}\right). \tag{2.2}$$

Here $A_\mu(x)$ are gauge potentials, $\xi^{\;i}_{a}(h)$ $(h\in G)$ are a complete set of Killing vectors for G, and $g_{\mu\nu}(x)$ and $g_{ij}(h)$ are metrics for the four-dimensional space and C, respectively. Of course such an ansatz for the metric g_{AB} with the ground state $M_4\otimes C$ is a nontrivial assumption. One would rather obtain Eq. (2.2) as an output from a solution of Einstein's equations, for example. With carefully chosen matter fields and a cosmological constant, the $M_4\otimes C$ ground state can be achieved.[2,3] The compatibility of the Kaluza–Klein ansatz [Eq. (2.2)] with the field equations is nontrivial in general.[3]

Fermions are added to the action [Eq. (2.1)] in the usual minimal way:

$$S(g)\to S(g)+\int\bar\psi(-i\not{D})\psi\sqrt{g}\,dx\,. \tag{2.3}$$

The fermion fields are then expanded into harmonic functions on the group manifold, $\mathscr{D}^\tau(h)$, with spectral measure $\rho(\tau)$,

$$\psi(x,h)=\int d\rho(\tau)\mathscr{D}^\tau(h)\psi^\tau(x)\,. \tag{2.4}$$

Upon expansion of the action in Eq. (2.3), use of the harmonic expansion [Eq. (2.4)], and integration over the internal coordinates h, we obtain the Einstein action on M^4, the Yang–Mills action and the gauged interactions with the fermions. The mass operator for the fermions on M_4 is simply

$$M=-i\not{D}_{\rm int}\,, \tag{2.5}$$

where $\not{D}_{\rm int}$ is the Dirac operator on the internal manifold

$$\not{D}_{\rm int}=\Gamma^a(\partial_a+\tfrac{1}{2}\omega_{abc}\sigma^{bc})\,. \tag{2.6}$$

The notations are the following: Γ^a are members of the Dirac algebra for the internal manifold

$$\{\Gamma^a,\Gamma^b\}=2\delta^{ab}\,, \tag{2.7a}$$

$$\sigma^{bc}=(1/4i)(\Gamma^b,\Gamma^c)\,, \tag{2.7b}$$

relative to a fixed orthonormal tangent basis $a=(5,\ldots,4+n)$. (See Sec. III for the definition of ∂_a.) The metric of the internal manifold is Euclidean flat $\delta^{ab}=(1,\ldots,1)$. The matrices σ^{bc} generate SO(N) in the spinor representation, and ω_{abc} is the spin connection on C.

The mass operator [Eq. (2.6)] commutes with a complete set of operators that generate the group G and anticommutes with a generalized γ_5 matrix. If C is a compact homogeneous space G/H (H the maximal subgroup of G) endowed with the standard torsion-free connection, then $-i\not{D}_{\rm int}$ has no massless eigenvalues (Lichnerowicz[4]). All fermion masses are $O(M_{\rm KK}=M_{\rm Planck}/\text{gauge coupling})$. From Eqs. (2.6) and (2.5),

$$M^2=(i\not{D}_{\rm int})^2=-\tfrac{1}{2}((\not{D},\not{D})+\{\not{D},\not{D}\})=-\Delta+R/4\,. \tag{2.8}$$

The first term in Eq. (2.8) is a generalized Laplacian

$$-\Delta=-(\partial_a+\tfrac{1}{2}\omega_{abc}\sigma^{bc})^2\geqslant 0\,,$$

for a compact manifold of Euclidean signature. The second term is the scalar curvature of the internal manifold. Therefore, if $R>0$, $\phi\in\mathscr{L}_2(G/H)$, then

$$\langle\phi,M^2\phi\rangle\geqslant\langle\phi,R\phi\rangle>0\,.$$

III. KINEMATIC DESCRIPTION OF TORSION

In this section some elementary notions of Riemannian geometry are recalled[16,17] for reasons of pedagogic completeness. The coveilbein and veilbein bases are introduced in the language of differential forms and derivatives:

$$e^a=e^a_i\,dx^i\,,\quad E_a=E^{\;i}_{a}\frac{\partial}{\partial x^i}=\partial_a\,. \tag{3.1}$$

An abbreviated notation is used in the following, where $E_a=\partial_a$. The Riemannian metric on the manifold C has the standard decomposition

$$g_{ij}=e^a_ie^a_j\,. \tag{3.2}$$

The bases $E^{\;i}_{a}$ and e^a_i of Eq. (3.2) are in general anholonomic with structure coefficients c^a_{bc}

$$de^a=-\tfrac{1}{2}c^a_{bc}e^b\wedge e^c\,, \tag{3.3}$$

$$(E_b,E_c)=c^a_{bc}E_a\,. \tag{3.4}$$

The Cartan structure relations are

$$T^a=de^a+\omega^a_b\wedge e^b\,, \tag{3.5a}$$

$$R^{\;a}_{b}=d\omega^a_b+\omega^a_c\wedge\omega^c_b\,, \tag{3.5b}$$

where T and R are the torsion and curvature two-forms, respectively, and ω is the connection one-form.

In the language of the covariant derivative one has

$$\nabla_ce^a=-\omega^a_{bc}e^b\,. \tag{3.6}$$

The Riemannian connection is defined by two conditions:

(i) $\nabla g=0$,

(ii) $T^a=0$.

Condition (i) is simply the covariant constancy of the metric which gives

$$\omega_{ab}=-\omega_{ba}\,.$$

Condition (ii) is the requirement of a torsion-free connection or symmetry. With the usual veilbein decomposition $T^a=\frac{1}{2}T^a_{bc}e^b\wedge e^c$, it follows that

$$T^a_{bc}=\omega^a_{bc}-\omega^a_{cb}-c^a_{bc}=0\,. \tag{3.7}$$

In the presence of torsion, Eq. (3.7) is modified. Expanding

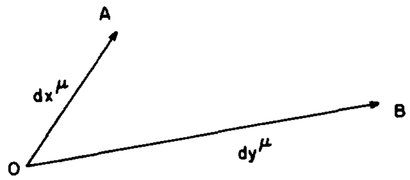


FIG. 1. Triangle AOB.

Eq. (3.6) in terms of the usual tensor components $\nabla_i e^a_j = (\partial/\partial x_i)e^a_j - \Gamma^a_{ij}$, the relation between the usual affine connection Γ^a_{bc} and torsion is simply

$$\Gamma^a_{[i,j]} = \omega^a_{[i,j]} - c^a_{[i,j]} = T^a_{ij}\,, \tag{3.8}$$

where [,] means antisymmetric part.

What does it mean to have a manifold with torsion? The notion of torsion has a simple geometric interpretation[18] that is easy to describe. First, recall that the infinitesimal parallel displacement of a vector a along a curve γ is simply

$$da^\mu = -\Gamma^\mu_{\lambda\sigma} a^\lambda\, dx^\sigma\,. \tag{3.9}$$

Next consider the triangle AOB with infinitesimal sides denoted by dx^μ and dy^μ (see Fig. 1). Infinitesimally displace the vector dx^μ from O to B and dy^μ from O to A using Eq. (3.9). Two new infinitesimal vectors are obtained, $dx^\mu - \Gamma^\mu_{\lambda\sigma}\, dx^\lambda\, dy^\sigma$ and $dy^\mu - \Gamma^\mu_{\lambda\sigma}\, dy^\lambda\, dx^\sigma$. These vectors form the candidate parallelogram of Fig. 2. This figure will be closed and form a parallelogram if $OC - OD = 0$ or equivalently $(\Gamma^\mu_{\lambda\sigma} - \Gamma^\mu_{\sigma\lambda})dx^\lambda\, dy^\sigma = 0$, and hence $T^\mu_{\lambda\sigma} = \Gamma^\mu_{[\lambda,\sigma]} = 0$.

Another guise that torsion assumes is easily identified through the veilbein and coveilbein bases[17] of Eq. (3.3). Namely a vector, v, can be associated with the projection $E_a\langle e^a,\cdot\rangle$ and decomposition

$$v = E_a\langle e^a, v\rangle\,. \tag{3.10}$$

Next we realize that, based upon requirements of symmetry, $d(E_a\langle e^a,\cdot\rangle) = 0$, or

$$\begin{aligned} 0 = d(E_a e^a) &= dE_a e^a + E_a\, de^a \\ &= E_a(\omega^a_b \wedge e^b + de^a) \\ &= E_a T^a\,. \end{aligned} \tag{3.11}$$

Hence, the existence of nonzero torsion precludes a mutually orthogonal veilbein and coveilbein bases locally in such a manifold.

Some general remarks conclude this section on the kinematical aspects of torsion. Any theory of gravitation based upon the principles of equivalence and general covariance cannot have torsion as a possibility in the physical four dimensions. The affine connections do not have explicitly an antisymmetric part in order to be compatible with a local inertial frame structure for the particle equations of motion. In addition a totally antisymmetric (Cartan–Schouten) torsion does not modify the geodesic equation. However, in the extra non-space-time dimensions this total antisymmetry need not be true, in general.

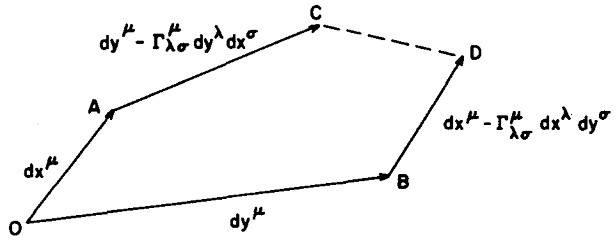


FIG. 2. Candidate parallelogram.

IV. PARALLELIZING TORSION AND PARALLELIZABLE MANIFOLDS

The notion of parallelizability originates from the classic work of Cartan and Schouten on semisimple Lie group manifolds.[5] Parallelizability simply means that the tangent space bundle is defined everywhere on the manifold and has a direct product structure. For example, the tangent bundle for SO(3), TSO(3) is simply $R^3 \otimes SO(3)$ (see Ref. 10). Besides semisimple Lie groups other parallelizable manifolds are the exceptional spheres S^1, S^3 [the group manifold SU(2)], and S^7. In addition, direct products $S^m \otimes S^n$, etc., are also parallelizable. Unfortunately, the totality of such manifolds are more rare than common. Even worse, the rare parallelizable manifold does not necessarily present itself in a physically interesting context for Kaluza–Klein.

First, the Lie group manifold is presented as a prototype parallelizable manifold. A Lie group is a group G that is a differentiable manifold and for which the natural operations are differential maps $G \times G \to G$ and $G \to G$. The tangent space $T(G)$ has a natural Lie algebra structure.[10] A left and right set of coveilbein bases $e^a_{L,R}$ exist everywhere on the manifold. Lie-algebra-valued differential forms are defined through the relations

$$g^{-1}dg = ie^a_L\lambda_a\,, \quad dg\,g^{-1} = ie^a_R\lambda_a\,. \tag{4.1}$$

If g is the fundamental representation of G, then the matrices λ_a are in the corresponding representation of the Lie algebra

$$[\lambda_a,\lambda_b] = if_{abc}\lambda_c\,, \tag{4.2}$$

with normalization $\mathrm{Tr}\,\lambda_a\lambda_b = 2\delta_{ab}$ and structure constants f_{abc}. The indices a, b, c, are in the adjoint representation. The Maurer–Cartan structure relations follow immediately from Eq. (4.1):

$$\begin{aligned} de^a_L\,\lambda_a &= -i\,dg^{-1}\wedge dg = -ig^{-1}dg\wedge g^{-1}dg \\ &= -(i/2)[\lambda_b,\lambda_c]e^b_L\wedge e^a_L \\ &= \tfrac{1}{2}f_{abc}\lambda_a e^b_L\wedge e^c_L\,. \end{aligned}$$

Hence,

$$de^a_{\substack{L\\R}} = \pm\tfrac{1}{2}f_{abc}\left(e^b_{\substack{L\\R}}\wedge e^c_{\substack{L\\R}}\right). \tag{4.3}$$

The anholonomic structure coefficients, see Eq. (3.4), are $c^a_{bc} = -f^a_{bc}$.

Next the torsion-free connection and curvature are deduced from the Cartan structure relations Eq. (3.5a), $T = de + \omega\wedge e$, and $R = d\omega + \omega\wedge\omega$. One easily obtains from Eq. (4.3) the results

$$\omega_{ab} = \tfrac{1}{2}f_{abc}e^c_L \tag{4.4a}$$

and

$$R_{ab} = \tfrac{1}{4}f_{abc}f_{cef}\,(e^e_L\wedge e^f_L)\,. \tag{4.4b}$$

Since a semisimple Lie group manifold is parallelizable, the following question arises: Is there a non-Riemannian connection with the property $R = d\omega + \omega \wedge \omega = 0$ that flattens the manifold? The answer is clearly yes. A hint of this connection comes directly from Eq. (3.5b); $R = d\omega + \omega \wedge \omega$ plays the role of an integrability condition on the manifold. The Frobenius integration theorem always ensures the existence of a system of integrating factors. The statement of the theorem is as follows.

Let $f^1,\ldots,f^r$ be a complete set of linearly independent one-forms in R^n, $n = r + s$. And suppose these one-forms satisfy

$$df^i = \sum_{j=1} \theta^i_j \wedge f^j \quad (i = 1,\ldots,r),$$

then there are functions F^i_j and g^i such that

$$f^i = \sum_{j=1}^{r} F^i_j \, dg^j .$$

The hypothesis is certainly a necessary one, the sufficiency is established from considerations in the theory of ordinary differential equations. The proof can be found in the book by Flanders.[19]

Applying the Frobenius integration theorem to Eq. (3.5b) there exists everywhere on G a connection one-form (see the example p. 101 in the book of Flanders[19])

$$_L\omega^a_b = (\phi^{-1})^a_c \, d\phi^c_b . \tag{4.5}$$

The matrix ϕ in Eq. (4.5) lies in the adjoint representation of the group G, and, using Eq. (4.1), one easily obtains the following results, noting that $(\lambda^{\text{adjoint}}_a)_{bc} = if_{abc}$:

$$_R\omega_{ab} = 0 , \quad _L\omega_{ab} = f_{abc} e^a_L , \quad T_{abc} = -f_{abc} . \tag{4.6}$$

Similarly, one can use the right covariant basis in Eq. (4.1) and obtain, using $\phi^a_b e^b_L = e^a_R$, the results

$$_L\omega_{ab} = 0 , \quad _R\omega_{ab} = f_{abc} e^c_R , \quad T_{abc} = +f_{abc} . \tag{4.7}$$

The distinction between the right and left covariant veilbein basis is familiar from the corresponding space-fixed and body-fixed coordinate systems in the motion of rigid bodies.[10]

The conclusion that the connection is pure gauge does not mean that the manifold is flat. Equation (3.5a), $T = de + \omega \wedge e$, does not imply that there exists a basis where e is an exact differential and $d^2 = 0$ as in flat space. The torsion is a local obstruction to flatness.

The same conclusions essentially hold for the parallelizable manifolds $S^m \otimes S^n$. Recently Turok and Zee[11] have pointed out the relevance of such manifolds in the incorporation of the notion of orthogonal grand unification, i.e., SO(10) into the Kaluza–Klein framework. This picture has the attractive feature that massless fermions must lie in the spinor representation of G as opposed to a contrived reducible representation such as the $\bar{5} + 10$ of the familiar SU(5).

The construction of the connection and parallelizing torsion for manifolds of the generic type $S^1 \otimes S^n$ has been given by Turok and Zee.[11] Again the second Cartan structure relation (3.5b) with $R = d\omega + \omega \wedge \omega = 0$ is an integrability condition for the parallelizing connection on $S^1 \otimes S^n$. The corresponding gauge function ϕ analogous to Eq. (4.5) can be constructed via the path ordered exponentiation $\phi = P \exp \int_{x_0}^{x} \omega$ on almost all of R^{n+1}, $R^{n+1} - \{0\}$. The general procedure is again a consequence of the theorem of Frobenius. The more important consideration is that such results exist. In the situation where $R = 0$, the gauge function is actually independent of path and well behaved on almost all of R^{n+1}. It is also worth articulating that manifolds of the type $S^n \otimes S^m$ are not parallelizable in the same sense as Lie group manifolds. The (Cartan–Schouten) torsion tensor is not completely antisymmetric as in Eqs. (4.6) and (4.7).

Finally, the torsion and parallelizability discussed here are not dynamically motivated. A dynamical origin of the parallelizing torsion perhaps lies in a consistent quantum treatment of the problem. A common difficulty of the Kaluza–Klein approach also involves the cosmological constant, which appears as a matter term added by hand to the Einstein equations. However, theories based upon a parallelizing torsion have a vanishing cosmological constant, a desirable feature.

V. MASSLESS FERMIONS

Once a simple expression for the spin connection on parallelizable manifolds is achieved, the determination of the representation for massless fermions is immediate. An intuitive picture is obtained by realizing that the mass operator [Eqs. (2.5) and (2.6)]

$$M = -i\Gamma^a(\partial_a + S_a)$$

consists of two pieces: $-i\,\partial_a$, a generalized orbital angular momentum, and $S_a = \frac{1}{2}\omega_{abc}\sigma^{bc}$, a generalized spin angular momentum. Massless fermions lie in representations with the orbital angular momentum when $-i\,\partial_a$ is equal to zero.

If ω_{abc} is the quasi-Riemannian spin connection which generates the tangent space group, then the following algebra is easy to obtain from Eq. (2.7):

$$[S_a, S_b] = if_{abc} S_c , \quad [S_a, \Gamma_b] = if_{abc} \Gamma_c . \tag{5.1}$$

The algebra of Eq. (5.1) spanned by S_a, $a = (1,\ldots,N)$, is the $2^{K/s}$-dimmensional Clifford representation. For Lie groups, $K = (N, N-1)$ and $\omega_{abc} = f_{abc}$ (N = even, odd), where N is the dimensionality of the adjoint representation. The matrices S_a provide the mapping from the adjoint representation to the spin $O(N(N-1)/2)$ representation.

If the connection is pure gauge for the parallelized case, the mass operator simplifies to

$$M = -i\Gamma^a(\partial_a + S_a) = \Gamma^a \mathscr{S}^{-1}(-i\,\partial_a)\mathscr{S} , \tag{5.2}$$

where S satisfies

$$\mathscr{S}^{-1} d\mathscr{S} = S_a e^a . \tag{5.3}$$

Equation (5.3) is simply Eq. (4.5) in the disguise of the spinor representation for the connection and $\phi \equiv \mathscr{S}$. After use of Eq. (5.1) it is obvious that the Dirac equation in the extra dimensions $M\psi = 0$ is equivalent to

$$\Gamma^a(-i\,\partial_a)\psi_{\mathscr{S}} = 0 , \tag{5.4}$$

where $\psi = \mathscr{S}\psi_{\mathscr{S}}$. It is clear that a class of zero-mode solutions to

$$M\psi = 0$$

is of the form $\psi = \mathscr{S}\psi_{\mathscr{S}}$, where $\psi_{\mathscr{S}}$ is a constant or in the singlet representation in the harmonic expansion Eq. (2.4).

Such zero modes are in the spinor or Clifford representation for the spinor ψ.

A less trivial question is the following: Are there other zero mode solutions? That this is not the case can be made plausible by the following argument. Only a finite number of irreducible representations occur in the harmonic expansion of $\psi_{\mathscr{S}}$ on the group manifold in Eq. (5.4), $\psi_{\mathscr{S}}$ is a polynomial. In addition to the constant $\psi_{\mathscr{S}}$ (already considered) there are higher polynomials in the coordinates on the group manifold. However, these terms must vanish from the completeness of the veilbein basis and the nonsingular nature of the Dirac matrices. In summary the zero modes correspond to a generalized orbital angular momentum equal to zero.

The above argument demonstrates that the fermions couple to the *unique spinor representation of the Lie group*. This representation, which is irreducible or irreducible repeated mod 2, has been analyzed in great detail by Wu and Zee.[5] Unfortunately, such representations are not useful for physics. In contrast, for parallelizable manifolds of the generic type $S^n \otimes S^m$, the spinor representations may be relevant to physics. The example of $S^1 \otimes S^9$, whose isometry group is SO(10) with the spinor representation $16 \oplus \overline{16}$, is important for orthogonal grand unification. However, a serious point has been raised on the Hermiticity of the Dirac operator in this case, since the torsion is not completely antisymmetric.[20] While this is not a difficulty for semisimple group manifolds, the general resolution of this point and the relevant mass operator are still under investigation.

VI. SUMMARY AND CONCLUSIONS

Although the final results of the torsion strategy for Kaluza–Klein theory are still far from satisfactory, it is still tempting to explore such possibilities in any case. The possibilities for group manifolds as the compactified space C do not occur in a way that is interesting for Kaluza–Klein. The manifolds do not have the phenomenologically attractive gauge group $SU_c(3) \otimes SU(2) \otimes U_1 \subset SU(5)$ as possible isometry groups. Needless to say, the massless fermion representations do not look anything like those of the standard model or its lowest-rank grand unified extensions with the maximal breaking of discrete symmetries (parity). The multiplets found are reducible spinor representations of a dimension that are totally unattractive.

The other strategy outlined by Turok, Zee,[11] and possibly others[10] can at least accompany orthogonal grand unification whose lowest-rank possibility contains SO(10) with fermions in the reducible $16 \oplus \overline{16}$ (irreducible under γ_5) representation. Mirror fermions are a necessary appendage in such a model. While the masses of the mirror fermions are less than M_{KK}, perhaps their masses can be pushed higher by more conventional, albeit ugly, symmetry-breaking schemes.

A few remarks about nontrivial topological configurations are also in order at the conclusion of this paper. It is well known from the work on classical field configurations, i.e., instantons, and monopoles, and the related index theory that chirally asymmetric fermion zero modes are abundant on all G bundles. The general index theorem on the twisted spin complex, where both the signature characteristic $\mathrm{tr}\, R \wedge R + \cdots$ and Chern characteristic $\mathrm{tr}\, F \wedge F + \cdots$ enter, gives a result for a compact manifold of the form

$$\text{index} = \int_M A(M) \wedge \mathrm{ch}\, V. \tag{6.1}$$

In Eq. (6.1) A is the A roof genus which generates the signature complex and ch V is the usual Chern form. Equations like (6.1) are the generalized Atiyah–Singer[16] index theorem extended to a Dirac operator including an external gauge potential. Such external gauge potentials defeat the Lichnerowicz theorem; the gauge potential is not part of the metric in extra dimensions. And, indeed, chirally asymmetric zero fermion modes occur and have been proposed by many authors.[9,11]

One might be tempted in the Kaluza–Klein spirit to look for solutions of the Einstein equations in higher dimensions or metrics with nontrivial signature characteristic polynomials in $\mathrm{tr}\, R \wedge R \cdots$ itself. The matter gauge fields are not part of R in the higher dimensions, and nontrivial index theory exists in $4n$ dimensions. In this manner Lichnerowicz may be circumvented. If such configurations are obtained from the Einstein equations $R_{AB} - \frac{1}{2} g_{AB} R = 0$, without a cosmological constant, the spaces will necessarily be Einstein manifolds with $R = 0$ (Ricci flat). Unfortunately, nontrivial Ricci flat compact manifolds do not have infinitesimal symmetries (Killing vectors). This follows from Yano's integral formula in differential geometry.[21] Within the strict Kaluza–Klein geometric approach this strategy does not appear to be promising. However, such an argument is not a relevant criticism for the super-string approach.

ACKNOWLEDGMENTS

I would like to thank the organizers of the Aspen Summer Institute of Physics for hospitality during the summer of 1984. Many important conversations with the participants there helped solidify some of the ideas in this paper.

This paper has been supported in part by the U.S. Department of Energy under Contract No. DE-AC02-76CH00016.

[1]Th. Kaluza, Sitzungsber. Preuss. Akad. Wiss Berlin K1, 966 (1921); O. Klein, Z. Phys. **37**, 895 (1926).
[2]J. Scherk and J. H. Schwarcz, Phys. Lett. B **57**, 463 (1975); Y. M. Cho and P. G. O. Freund, Phys. Rev. D **12**, 1711 (1975); E. Cremer and J. Scherk, Nucl. Phys. B **108**, 408 (1976); P. G. O. Freund and M. Rubin, Phys. Lett. B **97**, 233 (1980); E. Witten, Nucl. Phys. B **186**, 412 (1981); A. Salam and J. Strathdee, Ann. Phys. **141**, 316 (1982); T. Appelquist and A. Chodos, Phys. Rev. Lett. **50**, 141 (1983); Phys. Rev. D **28**, 772 (1983).
[3]M. J. Duff and D. Toms, *Unification of Fundamental Interactions II*, edited by S. Ferrara and J. Ellis (Plenum, New York, 1982), Vol II. For the consistency with the field equations, see M. J. Duff, B. E. W. Nilsson, C. N. Pope, and N. P. Warner, Phys. Lett. B **149**, 90 (1984).
[4]A. Lichnerowicz, in *Proceedings of the 1963 Les Houches Summer School, Relativity, Groups and Topology* (Gordon and Breach, New York, 1963).
[5]E. Cartan and J. A. Schouten, Proc. Kon. Ned. Akad. Amsterdam **29**, 803 (1926); C. Destri, C. A. Orzalesi, and P. Rossi, Ann. Phys. **147**, 321 (1983); M. J. Duff and C. A. Orzalesi, Phys. Lett. **122**, 37 (1983); Y. S. Wu and A. Zee, J. Math. Phys. **25**, 2696 (1984).
[6]X. Z. Wu, Phys. Rev. D **29**, 2769 (1984).
[7]M. Duff, B. E. W. Nilsson, and C. N. Pope, Nucl. Phys. B **214**, 491 (1983).
[8]M. Gell-Mann and B. Zwiebach, University of California, Berkeley preprint U(B-PTH-84/20), 1984, C. Wetterich, University of Bern preprints BUTP-84/5, BUTP-84/6, both 1984.

[9]E. Witten, "Fermion Quantum Numbers in Kaluza–Klein Theories," Princeton University preprint, 1983.
[10]A. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978).
[11]N. Turok and A. Zee, Institute for Theoretical Physics, University of California, Santa Barbara preprint, 1984.
[12]N. Steenrod, *The Topology of Fibre Bundles* (Princeton U. P., Princeton, 1951).
[13]P. Langacker, Phys. Rep. **72**, 195 (1981).
[14]S. Randjbar-Daemi, A. Salam, and J. Strathdee, Phys. Lett. B **132**, 56 (1983); Nucl. Phys. B **214**, 491 (1983); G. F. Chapline and B. Grossman, Phys. Lett. B **135**, 109 (1984).
[15]S. Weinberg, Phys. Lett. B **138**, 47 (1984).
[16]T. Eguchi, P. Gilkey, and A. Hanson, Phys. Rep. **66**, 214 (1980).
[17]C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[18]S. Sen and C. Nash, *Topology and Geometry for Physicists* (Academic, New York, 1983).
[19]H. Flanders, *Differential Forms* (Academic, New York, 1963).
[20]C. Wetterich, University of Bern preprint BUTP-84/5, 1984.
[21]K. Yano, *Integral Formulas in Differential Geometry* (Marcel Dekker, New York, 1970).

Concepts of conditional expectations in quantum theory

Masanao Ozawa
Department of Mathematics, College of General Education, Nagoya University, Chikusa-ku, Nagoya 464, Japan

(Received 17 July 1984; accepted for publication 29 March 1985)

A concept of conditional expectations in quantum theory is established with interrelations to previously introduced concepts of the Cycon–Hellwig conditional expectations and *a posteriori* states, which are analogous to the existing interrelations in the classical probability theory among conditional expectations related to random variables, those related to σ subalgebras and conditional probability distributions. These three concepts are shown to have satisfactory statistical interpretation in the quantum measuring processes. For the above purpose, we introduce an integration with respect to functions with values in the states of operator algebras and positive operator valued measures such that the resulting indefinite integrals are completely positive map valued measures. Eventually, it is proved that in the von Neumann algebraic formulation, the Cycon–Hellwig conditional expectations always exist as completely positive map valued measures.

I. INTRODUCTION

The statistical interpretation of quantum mechanics introduces a probability theory concerning states and observables[1] analogous to probability measures and random variables in the classical probability theory.[2] However, the problem of introduction of the concept of conditional expectations in quantum theory has not been settled on a firm basis. The aim of this paper is to settle this problem from the aspects of physical interpretation and mathematical formalism together with the results obtained in our preceding papers.[3–5]

The noncommutative conditional expectations are first introduced by Umegaki[6] in the theory of operator algebras inspired by Moy's operator theoretic characterization of conditional expectations in the classical probability theory.[7] A simple characterization of his conditional expectations is obtained by Tomiyama[8] as projections of norm 1. It is proved by Nakamura–Umegaki[9] that Umegaki's conditional expectations describe the state changes caused by measurements of observables with a simple discrete spectrum satisfying the repeatability hypothesis due to von Neumann.[1] However, Umegaki's conditional expectation exists if and only if the observable has a discrete spectrum (c.f. Refs. 10–13 and Ref. 14, Theorem 4.4.2).

For the purpose of introducing joint probability distributions of noncommuting observables and conditional expectations of continuous observables, Davies–Lewis[15] introduces the mathematical concept of instruments as transformation valued measures, which are mathematical models of state changes caused by measurements. In their proposed interpretation, the analogs of conditional expectations are supposed to be the dual objects of instruments. Cycon and Hellwig[16] pointed out, however, that such an interpretation is misleading even in the classical probability theory and they introduce the new conditional expectations. In fact, measurements of classical ensembles cause no state changes but any statistical measurements yield the conditional expectation. Thus we must discriminate between state changes and conditional expectations in quantum mechanics.

In our previous paper,[4] we introduce the mathematical concept of measuring processes which generalizes von Neumann's analysis of measuring processes of discrete observables (see Ref. 1, Chap. VI) to continuous observables. The measuring processes provide us with a natural interpretation of the concept of instruments that states changes caused by a given measuring process naturally correspond to a CP instrument. In our continued work,[5] we introduce the concept of *a posteriori* states which is an analog of regular conditional probability distributions in the classical probability theory, and related to the Bayes principle in the statistical interpretation of measuring processes. Using the *a posteriori* states we can easily manipulate the conditional expectation of an observable conditioned by the outcomes of the preceding measurement.

Our interpretation of conditional expectations is as follows. Let ρ be a state of a quantum system. A measurement of an observable in the system causes the state change $\rho \rightarrow \rho'$ of the whole ensemble and the conditional state change $\rho \rightarrow \rho_x$ of the subensemble distinguished by the outcome x of the measurement. The state ρ_x is called the *a posteriori* state with respect to the *a priori* state ρ in Ref. 5. The well-known reduction of wave packets is the state change of the latter type. We can illustrate with the analogous situation of measurements of random variables in the classical statistical ensemble. In this case, the state change of the whole ensemble does not occur, i.e., $\rho = \rho'$. But even in the classical case the information of an outcome of a measurement changes the probability law of the system from *a priori* probability to *a posteriori* probability by the Bayes principle. Then conditional expectations are simply the expectation values with respect to *a posteriori* probabilities. Analogously, we can interprete conditional expectations as the expectation values with respect to *a posteriori* states.

In the conventional measurements of observables with simple discrete spectrum, this concept of conditional expectations coincides with the dual transformation of the total state change. But they differ in general. It will be shown that our conditional expectation is equivalent to the Cycon–Hellwig conditional expectations, which eventually admit the above interpretation.

 0022-2488/85/081948-08$02.50

In the classical probability theory, we have three types of concepts of conditional expectations: conditional expectations with respect to σ subfields, conditional expectations with respect to random variables, and regular conditional probability distributions. Thus we can expect that there are three types of analogs of conditional expectations in quantum mechanics corresponding, respectively. However, it is a quite confusing fact that the corresponding conditional expectations will have very different mathematical forms: the first type is expressed by observables related to the Cycon–Hellwig conditional expectations, the second type is expressed by measurable functions on a value space that will be introduced in this paper, and the last type is expressed by a family of states related to *a posteriori* states. In this paper, we shall give interrelations among these three types of analogs of conditional expectations.

In Sec. II, we introduce a concept of conditional expectations in the von Neumann algebraic formulation of the operational quantum probability theory of Davies and Lewis and show their existence and basic properties. In Secs. III and IV, we show the interrelations among the conditional expectations introduced in Sec. II, the Cycon–Hellwig conditional expectations and *a posteriori* states, by giving the formulas transferring one to another. As a consequence, it is proved that the Cycon–Hellwig conditional expectation always exists as a dual object of some completely positive instrument. In Sec. V, we discuss the statistical interpretation of these three concepts of conditional expectations in the quantum measuring processes. In the Appendix, we provide an integration theory with respect to families of states on von Neumann algebras and positive operator valued measures such that the resulting indefinite integrals are completely positive map valued measures, which is used in Sec. IV.

II. CONDITIONAL EXPECTATIONS IN QUANTUM PROBABILITY THEORY

In this section we shall study a concept of conditional expectations within the framework of operational quantum probability theory due to Davies–Lewis.[15] The conditional expectation treated in this section is a generalization of a conditional expectation with respect to a random variable in the classical probability theory (cf. Ref. 17, Definition 4.12).

Our setting for operational quantum probability theory consists of a von Neumann algebra $\mathscr{M}$ and a Borel space $(\Lambda,\mathscr{B}(\Lambda))$. Denote by $\mathscr{M}_*$ the predual of $\mathscr{M}$, and by $\mathscr{L}^+(\mathscr{M}_*)$ the space of all positive linear tranformations on $\mathscr{M}_*$. Denote by $\langle\cdot,\cdot\rangle$ the duality pairing between $\mathscr{M}_*$ (or $\mathscr{M}^*$) and $\mathscr{M}$. A *semiobservable* X in $\mathscr{M}$ is a positive operator valued measure X: $\mathscr{B}(\Lambda)\to\mathscr{M}$ such that $X(\Lambda)=1$. An *observable* in $\mathscr{M}$ is a semiobservable which is projection valued. A map $\mathscr{I}$: $\mathscr{B}(\Lambda)\to\mathscr{L}^+(\mathscr{M}_*)$ is called an *instrument* for $(\mathscr{M},\Lambda)$, if it satisfies the following conditions.

(I1) For each ρ in $\mathscr{M}_*$, $\langle\mathscr{I}(\Lambda)\rho,1\rangle=\langle\rho,1\rangle$.

(I2) For each disjoint sequence $\{B_i\}$ in $\mathscr{B}(\Lambda)$,

$$\mathscr{I}(\cup_i B_i)=\sum_i \mathscr{I}(B_i),$$

where the sum is convergent in the strong operator topology of $\mathscr{L}^+(\mathscr{M}_*)$.

An instrument $\mathscr{I}$ is called a *CP instrument* if it further enjoys the following condition.

(I3) For each B in $\mathscr{B}(\Lambda)$, $\mathscr{I}(B)$ is completely positive.

Let $\mathscr{I}$ be an instrument. Then we can associate a semiobservable X in $\mathscr{M}$ such that

$$X(dx)=\mathscr{I}(dx)^*1.$$

In this case, $\mathscr{I}$ is called an instrument for $(\mathscr{M},\Lambda,X)$. Now we fix a normal state ρ on $\mathscr{M}$. Then we can associate a probability distribution μ_X of X on $(\Lambda,\mathscr{B}(\Lambda))$ such that

$$\mu_X(dx)=\|\mathscr{I}(dx)\rho\|=\langle\rho,X(dx)\rangle.$$

We call μ_X the *probability distribution* associated with $(\mathscr{I},\rho)$.

In the rest of this section, let $\mathscr{I}$ be an instrument for $(\mathscr{M},\Lambda,X)$ and ρ a normal state on $\mathscr{M}$.

Definition 2.1: For any a in $\mathscr{M}$, a *conditional expectation* $\{E(a|x);x\in\Lambda\}$ of a with respect to $(\mathscr{I},\rho)$ is a complex-valued $\mathscr{B}(\Lambda)$-measurable function $x\to E(a|x)$ such that

$$\int_B E(a|x)\mu_X(dx)=\langle\mathscr{I}(B)\rho,a\rangle, \tag{2.1}$$

for all B in $\mathscr{B}(\Lambda)$.

In what follows, we shall prove the existence and basic properties of conditional expectations.

Theorem 2.2: For any a in $\mathscr{M}$, there is a conditional expectation $\{E(a|x);x\in\Lambda\}$ of a with respect to $(\mathscr{I},\rho)$ unique up to μ_X-almost everywhere.

Proof: For any a in $\mathscr{M}$ and B in $\mathscr{B}(\Lambda)$, we have

$$|\langle\mathscr{I}(B)\rho,a\rangle|\leqslant\|a\|\ \|\mathscr{I}(B)\rho\|=\|a\|\mu_X(B).$$

Thus the function $B\to\langle\mathscr{I}(B)\rho,a\rangle$ is a finite signed measure on $(\Lambda,\mathscr{B}(\Lambda))$ absolutely continuous with respect to μ_X. Let $E(a|x)$ be a Radon–Nikodym derivative $\langle\mathscr{I}(dx)\rho,a\rangle/\mu_X(dx)$. Then it is easy to see that Eq. (2.1) holds for all B in $\mathscr{B}(\Lambda)$. Thus the required conditional expectation exists. The uniqueness follows immediately from Eq. (2.1). Q.E.D.

Theorem 2.3: Any conditional expectation $\{E(a|x);x\in\Lambda\}$ of a in $\mathscr{M}$ with respect to $(\mathscr{I},\rho)$ has the following properties.

(1) $E(1|x)=1$, μ_x-a.e.

(2) For any α,β in $\mathbf{C}$ and a, b in $\mathscr{M}$,

$E(\alpha a+\beta b|x)=\alpha E(a|x)+\beta E(b|x)$, μ_X-a.e.

(3) For any $a\geqslant 0$ in $\mathscr{M}$, $E(a|x)\geqslant 0$, μ_X-a.e.

(4) For any a in $\mathscr{M}$, $|E(a|x)|\leqslant\|a\|$, μ_X-a.e.

(5) For any net $\{a_\alpha\}$ convergent to a in $\mathscr{M}$ in the σ-weak topology, we have

$$\int_\Lambda E(a_\alpha|x)f(x)\mu_X(dx)\to\int_\Lambda E(a|x)f(x)\mu_X(dx),$$

for all f in $\mathscr{L}^1(\Lambda,\mu_X)$.

Proof: Properties (1)–(3) are obvious from the definition. To prove (4), let $a\in\mathscr{M}$. As in the proof of Theorem 2.2, we have

$$\left|\int_B E(a|x)\mu_X(dx)\right|=|\langle\mathscr{I}(B),a\rangle|=\|a\|\mu_X(B),$$

for all B in $\mathscr{B}(\Lambda)$. It follows tht $|E(a|x)|\leqslant\|a\|$, μ_X-a.e.

For (5), suppose that $a_\alpha\to a$ in the σ-weak topology.

Then $\langle \mathscr{I}(B)\rho, a_\alpha\rangle \to \langle \mathscr{I}(B)\rho, a\rangle$ and hence

$$\int_B E(a_\alpha|x)\mu_X(dx) \to \int_B E(a|x)\mu_X(dx),$$

for all B in $\mathscr{B}(\Lambda)$. Since $\{E(a_\alpha|\cdot)\}$ is a bounded net in $L^\infty(\Lambda,\mu_X)$, it follows that it is convergent in the $\sigma(L^\infty, L^1)$ topology. Thus we have

$$\int_\Lambda E(a_\alpha|x)f(x)\mu_X(dx) \to \int_\Lambda E(x|x)f(x)\mu_X(dx),$$

for all f in $L^1(\Lambda,\mu_X)$. Q.E.D.

Let $\mathbf{E}: \mathscr{M} \to L^\infty(\Lambda,\mu_X)$ be the map defined by

$$\mathbf{E}(a) = \text{the } \mu_X\text{-a.e. equivalence class of } E(a|\cdot), \tag{2.2}$$

for all a in $\mathscr{M}$. Since the function $x \to E(a|x)$ is essentially bounded and unique up to μ_X-almost everywhere, $\mathbf{E}$ is well defined. We call this $\mathbf{E}$ the *conditional expectation operator* with respect to $(\mathscr{I}, \rho)$.

Theorem 2.4: The conditional expectation operator $\mathbf{E}: \mathscr{M} \to L^\infty(\Lambda,\mu_X)$ with respect to $(\mathscr{I}, \rho)$ is a positive normal unit preserving linear map.

Proof: Immediate from Theorem 2.3. Q.E.D.

III. CYCON–HELLWIG CONDITIONAL EXPECTATIONS

In Ref. 16, Cycon and Hellwig pointed out that the concept of a conditional expectation is not the dual object of a given instrument and introduced their generalized conditional expectation as a generalization of conditional expectations with respect to σ subfields in classical probability theory (see Ref. 17, Definition 4.13). In this section, we shall establish a relation between the Cycon–Hellwig conditional expectations and the conditional expectations introduced in Sec. II. The relation is analogous with the existing relation between the conditional expectations with respect to σ subfields and the conditional expectations with respect to random variables in the classical probability theory. At the same time we shall show that the Cycon–Hellwig conditional expectations always exist.

Let $\mathscr{M}$ be a von Neumann algebra and let $(\Lambda, \mathscr{B}(\Lambda))$ be a Borel space. In literatures (see Ref. 18, for instance), the dual object of an instrument is called an expectation. To avoid the unnecessary conceptual confusions we adopt the following terminology. Let $\mathscr{L}^+(\mathscr{M})$ be the space of positive linear maps on $\mathscr{M}$. An $\mathscr{L}^+(\mathscr{M})$-valued map $\mathscr{E}$ on $\mathscr{B}(\Lambda)$ is called a *dual instrument* if there is an instrument $\mathscr{I}$ for $(\mathscr{M}, \Lambda)$ such that $\langle \rho, \mathscr{E}(B)a\rangle = \langle \mathscr{I}(B)\rho, a\rangle$ for all a in $\mathscr{M}$, ρ in $\mathscr{M}_*$, B in $\mathscr{B}(\Lambda)$. Then an $\mathscr{L}^+(\mathscr{M})$-valued map $\mathscr{E}$ on $\mathscr{B}(\Lambda)$ is a dual instrument if and only if it satisfies the following conditions.

(D1) For each B in $\mathscr{B}(\Lambda)$, $\mathscr{E}(B)$ is a normal positive linear map on $\mathscr{M}$.

(D2) $\mathscr{E}(\Lambda)1 = 1$.

(D3) For each countable family $\{B_i\}$ of pairwise disjoint sets in $\mathscr{B}(\Lambda)$, with $B = \cup_i B_i$,

$$\langle \rho, \mathscr{E}(B)a\rangle = \sum_i \langle \rho, \mathscr{E}(B_i)\rangle,$$

for all a in $\mathscr{M}$, ρ in $\mathscr{M}_*$ (cf. Ref. 18).

Let $\mathscr{I}$ be an instrument for $(\mathscr{M}, \Lambda, X)$ and ρ be a fixed normal state on $\mathscr{M}$. A dual instrument $\mathscr{E}$ is called a *Cycon–Hellwig conditional expectation* with respect to $(\mathscr{I}, \rho)$, if it satisfies the following conditions.

(E1) For each a in $\mathscr{M}$ and B in $\mathscr{B}(\Lambda)$, $\mathscr{E}(B)a$ is in the σ-weak closure of the linear span of $\{X(B); B \in \mathscr{B}(\Lambda)\}$.

(E2) For each a in $\mathscr{M}$ and B in $\mathscr{B}(\Lambda)$,

$$\langle \mathscr{I}(B)\rho, a\rangle = \langle \rho, \mathscr{E}(B)a\rangle.$$

In Ref. 16, condition (E1) is called the averaging requirement and (E2) the mean value requirement, which are characteristic for the conditional expectations with respect to σ subfields in the classical probability theory. However, the definition of Cycon and Hellwig is weaker than ours in the sense that their conditional expectation does not satisfy condition (D1) so that it is not a dual instrument in general (see Definition 3.1 of Ref. 16 and see also Ref. 19, p. 156).

To deal with the Cycon–Hellwig conditional expectations for normal states that are not faithful on the range of X, the following lemma on Lebesgue-type decompositions of semiobservables is crucial.

Lemma 3.1: Let $X: \mathscr{B}(\Lambda) \to \mathscr{M}$ be a semiobservable and ρ be a normal state on $\mathscr{M}$. Then there are two positive operator valued measures $X_a: \mathscr{B}(\Lambda) \to \mathscr{M}$ and $X_s: \mathscr{B}(\Lambda) \to \mathscr{M}$ satisfying the following conditions.

(1) $X(B) = X_a(B) + X_s(B)$, for all B in $\mathscr{B}(\Lambda)$.

(2) For every B in $\mathscr{B}(\Lambda)$, $\langle \rho, X(B)\rangle = 0$ if and only if $X_a(B) = 0$.

(3) The ranges of X_a and X_s are both contained in the σ-weak closure of the linear span of the range of X.

(4) For every B in $\mathscr{B}(\Lambda)$, $\langle \rho, X(B)\rangle = \langle \rho, X_a(B)\rangle$ and $\langle \rho, X_s(B)\rangle = 0$.

Proof: Let $\mu_X(dx) = \langle \rho, X(dx)\rangle$. Let $\mathscr{H}$ be a Hilbert space on which $\mathscr{M}$ acts. By taking a suitable representation, we may assume a vector η in $\mathscr{H}$ such that $\mu_X(B) = \langle \eta|X(B)|\eta\rangle$ for all B in $\mathscr{B}(\Lambda)$. Let $\mathscr{K}$ be a Hilbert space containing $\mathscr{H}$ and $E: \mathscr{B}(\Lambda) \to \mathscr{L}(\mathscr{K})$ be a minimal dilation of X, so that E is a spectral measure and that $X(B) = PE(B)P$ for all B in $\mathscr{B}(\Lambda)$, where P is the projection onto $\mathscr{H}$. Let C be the projection whose range is $\{\xi \in \mathscr{K}; \langle \xi|E(dx)|\xi\rangle < < \mu_X(dx)\}$. Then by Ref. 20 (Sec. 66, Theorem 1), $C \in E(\mathscr{B}(\Lambda))''$. By Ref. 20 (Sec. 66, Theorem 2), the vector η is in the range of C. We define X_a and X_s as follows:

$$X_a(B) = PE(B)CP \quad \text{and} \quad X_s(B) = PE(B)(1-C)P,$$

for all B in $\mathscr{B}(\Lambda)$. Then condition (1) holds obviously. If $\mu_X(B) = 0$ then $E(B)C = 0$ so that $X_a(B) = 0$. Conversely, if $X_a(B) = 0$ then $E(B)CP = 0$ so that $E(B)CaP = 0$ for every a in $E(\mathscr{B}(\Lambda))''$, and hence by the minimality of the dilation, $E(B)C = 0$, whence $\mu_X(B) = 0$. Thus condition (2) holds. Since C is in the weak closure of the range of E, it is easy to see that condition (3) holds. Condition (4) follows from the relations below:

$$\begin{aligned}\langle \rho, X(B)\rangle &= \mu_X(B) = \langle \eta|X(B)|\eta\rangle = \langle \eta|E(B)C|\eta\rangle \\ &= \langle \eta|X_a(B)|\eta\rangle = \langle \rho, X_a(B)\rangle,\end{aligned}$$

for all B in $\mathscr{B}(\Lambda)$. Q.E.D.

We shall call the decomposition $X = X_a + X_s$ obtained above the *Lebesgue decomposition* of X with respect to ρ, and call X_a the *absolutely continuous part* and X_s the *singular part* with respect to ρ. For the integration with respect to positive operator valued measures we shall refer to Berberian.[21]

In the rest of this section, let $\mathscr{I}$ be an instrument for $(\mathscr{M},\Lambda,X)$ and ρ a normal state on $\mathscr{M}$. Let $\{E(a|x); x\in\Lambda\}$ be a conditional expectation with respect to $(\mathscr{I},\rho)$.

Theorem 3.2: Let $X = X_a + X_s$ be the Lebesgue decomposition of X and ρ_0 be an arbitrary normal state on $\mathscr{M}$. Then the relation

$$\mathscr{E}(B)a = \int_B E(a|x)X_a(dx) + \rho_0(a)X_s(B), \tag{3.1}$$

for all a in $\mathscr{M}$ and B in $\mathscr{B}(\Lambda)$, defines a Cycon–Hellwig conditional expectation with respect to $(\mathscr{I},\rho)$.

Since the conditional expectation $\{E(a|x); x\in\Lambda\}$ always exists by Theorem 2.2, the following corollary immediately follows.

Corollary 3.3: Cycon–Hellwig conditional expectations with respect to $(\mathscr{I},\rho)$ always exist.

Proof of Theorem 3.2: By Theorem 2.2, the conditional expectation $E(a|x)$ is unique up to μ_X-almost everywhere, so that it follows from Lemma 3.1(2) that the right-hand side of Eq. (3.1) depends only on B and a. Thus Eq. (3.1) determines a unique element $\mathscr{E}(B)a$ in $\mathscr{M}$. Obviously, $\mathscr{E}(B)$ is a positive linear map on $\mathscr{M}$. To show that $\mathscr{E}$ is a dual instrument, let $\{a_\alpha\}$ be a net in $\mathscr{M}$ convergent to a in $\mathscr{M}$ in the σ-weak topology. Let $\varphi\in\mathscr{M}_*$ and $B\in\mathscr{B}(\Lambda)$. By Lemma 3.1(2), $\langle\varphi,X_a(dx)\rangle << \mu_X(dx)$, so that by Theorem 2.3(5), we have

$$\int_B E(a_\alpha|x)\langle\varphi,X_a(dx)\rangle \to \int_B E(a|x)\langle\varphi,X_a(dx)\rangle.$$

It follows that $\langle\varphi,\mathscr{E}(B)a_\alpha\rangle$ converges to $\langle\varphi,\mathscr{E}(B)a\rangle$. Thus condition (D1) holds. It is straightforward to verify conditions (D2) and (D3), and hence $\mathscr{E}$ is a dual instrument. By Lemma 3.1(1), it is easy to see that condition (E1) holds. For any a in $\mathscr{M}$ and B in $\mathscr{B}(\Lambda)$, we have

$$\begin{aligned}\langle\mathscr{I}(B)\rho,a\rangle &= \int_B E(a|x)\langle\rho,X(dx)\rangle \\ &= \int_B E(a|x)\langle\rho,X_a(dx)\rangle + \rho_0(a)\langle\rho,X_s(B)\rangle \\ &= \langle\rho,\mathscr{E}(B)a\rangle,\end{aligned}$$

where the second equality follows from Lemma 3.1(4). Hence condition (E2) holds. Therefore, we have proved that Eq. (3.1) defines a Cycon–Hellwig conditional expectation $\mathscr{E}$ with respect to $(\mathscr{I},\rho)$. Q.E.D.

IV. *A POSTERIORI* STATES AND CONDITIONAL EXPECTATIONS

In Ref. 5, we introduced a concept of *a posteriori* states from an analysis of statistical interpretation of measuring processes of quantum observables. The relation between this concept and the conditional expectation introduced in Sec. II is easily seen from their definitions (cf. Ref. 5, Definition 4.1). Now we shall rephrase the definition of *a posteriori* states in a more convenient way.

Let $\mathscr{I}$ be an instrument for $(\mathscr{M},\Lambda,X)$ and let ρ be a normal state on $\mathscr{M}$. Let μ_X be the probability distribution associated with $(\mathscr{I},\rho)$, i.e., $\mu_X(dx) = \langle\rho,X(dx)\rangle$. Denote by $\{E(a|x); x\in\Lambda\}$ a conditional expectation of a in $\mathscr{M}$ with respect to $(\mathscr{I},\rho)$.

Definition 4.1: A family $\{\rho_x; x\in\Lambda\}$ of normal states on $\mathscr{M}$ is called a family of *a posteriori states* with respect to $(\mathscr{I},\rho)$ if for any a in $\mathscr{M}$ we have

$$\langle\rho_x,a\rangle = E(a|x), \quad \mu_X\text{-almost everywhere.}$$

In Ref. 5, some conditions for existence of families of *a posteriori* states are obtained. In particular, for the standard formulation of quantum mechanics, they always exist. But in the general framework of operational quantum probability theory they do not always exist. We have shown in Ref. 5 (Theorem 5.1) that a weakly repeatable instrument $\mathscr{I}$ is discrete if, for a faithful normal state ρ, there is a family of *a posteriori* states with respect to $(\mathscr{I},\rho)$. This settles the conjecture of Davies and Lewis[15] completely. Now we shall introduce a more convenient notion for our purpose.

Definition 4.2: A family $\{\rho_x; x\in\Lambda\}$ of (not necessarily normal) states on $\mathscr{M}$ is called a *disintegration* with respect to $(\mathscr{I},\rho)$ if for any a in $\mathscr{M}$ we have

$$\langle\rho_x,a\rangle = E(a|x), \quad \mu_X\text{-almost everywhere.}$$

Thus families of *a posteriori* states are disintegrations such that all ρ_x are normal states.

A disintegration $\{\rho_x; x\in\Lambda\}$ is called *proper* if for any a in $\mathscr{M}$ with $a\geqslant 0$, $\langle\rho_x,a\rangle = 0$ for all x in Λ, whenever $\langle\mathscr{I}(\Lambda)\rho,a\rangle = 0$. Denote by $M^\infty(\Lambda,\mu_X)$ the space of all bounded complex-valued μ_X-measurable functions on Λ.

Theorem 4.3: Proper disintegrations $\{\rho_x; x\in\Lambda\}$ with respect to $(\mathscr{I},\rho)$ always exist. If $\{\rho'_x; x\in\Lambda\}$ is another disintegration with respect to $(\mathscr{I},\rho)$, then $\langle\rho_x,a\rangle = \langle\rho'_x,a\rangle$, μ_X-almost everywhere for all a in $\mathscr{M}$.

Proof: Let $\mathbf{E}:\mathscr{M}\to L^\infty(\Lambda,\mu_X)$ be the conditional expectation operator with respect to $(\mathscr{I},\rho)$ and let $\Phi: L^\infty(\Lambda,\mu_X)\mapsto M^\infty(\Lambda,\mu_X)$ be a lifting on $L^\infty(\Lambda,\mu_X)$ (see Ref. 22). For any x in Λ, let $\rho_x\in\mathscr{M}^*$ be such that $\langle\rho_x,a\rangle = \Phi(\mathbf{E}(a))(x)$ for all a in $\mathscr{M}$. Then it is easy to see that ρ_x is a state on $\mathscr{M}$ for any x in Λ and that $\langle\rho_x,a\rangle = E(a|x)$, μ_X-a.e. for all a in $\mathscr{M}$. Thus $\{\rho_x; x\in\Lambda\}$ is a disintegration of ρ with respect to I. To prove that it is proper, let $a\geqslant 0$ in $\mathscr{M}$ be such that $\langle\mathscr{I}(\Lambda)\rho,a\rangle = 0$. Then we have

$$\int_\Lambda E(a|x)\mu_X(dx) = \langle\mathscr{I}(\Lambda)\rho,a\rangle = 0,$$

whence $\mathbf{E}(a) = 0$. Thus by the property of the lifting, we have $\langle\rho_x,a\rangle = 0$ for all x in Λ. The assertion on the uniqueness up to μ_X-a.e. follows immediately from the corresponding property of the conditional expectation. Q.E.D.

In Ref. 16, the nuclear instruments are introduced. In the following, we shall consider the nuclear instruments in the von Neumann algebraic formulation. In this case, it is shown that every nuclear instrument is completely positive; see Ref. 4 for the physical meaning of complete positivity of instruments. For the details of the terminology and the necessary integration theory, we refer the reader to the Appendix.

Theorem 4.4: Let $X:\mathscr{B}(\Lambda)\mapsto\mathscr{M}$ be a semiobservable. Then for any X-integrable family $\{\rho_x; x\in\Lambda\}$ of (not necessarily normal) states on $\mathscr{M}$, there is a CP instrument $\mathscr{I}$ for $(\mathscr{M},\Lambda,X)$ such that

$$\langle\mathscr{I}(B)\rho,a\rangle = \int_B \langle\rho_x,a\rangle\langle\rho,X(dx)\rangle, \tag{4.1}$$

for any a in $\mathscr{M}$ and ρ in $\mathscr{M}_*$.

Proof: By Theorem A.6, Eq. (4.1) defines a CP $(\mathscr{M}_*)$-valued measure $\mathscr{I}$. Hence conditions (I2) and (I3) in Sec. II hold. Condition (I1) follows from

$$\langle \mathscr{I}(\Lambda)\rho,1\rangle = \int_\Lambda \langle \rho_x,1\rangle\langle \rho,X(dx)\rangle$$
$$= \langle \rho,X(\Lambda)\rangle = \langle \rho,1\rangle,$$

for any ρ in $\mathscr{M}_*$ and B in $\mathscr{B}(\Lambda)$. Thus $\mathscr{I}$ is a CP instrument. The relation $\mathscr{I}(dx)^*1 = X(dx)$ is obvious from Eq. (4.1). Q.E.D.

We shall write $\mathscr{I}(B) = \int_B \rho_x \otimes X(dx)$ for the CP instrument $\mathscr{I}$ defined by Eq. (4.1), and call it a *nuclear instrument* obtained from $\{\rho_x; x\in\Lambda\}$ and X.

Corollary 4.5: For any semiobservable $X:\mathscr{B}(\Lambda)\rightarrow\mathscr{M}$ and for any X-integrable family $\{\rho_x; x\in\Lambda\}$ of states on $\mathscr{M}$, there is a CP instrument $\mathscr{I}$ for $(\mathscr{M},\Lambda,X)$ such that for any normal state ρ on $\mathscr{M}$ the family $\{\rho_x; x\in\Lambda\}$ is a disintegration with respect to $(\mathscr{I},\rho)$.

Proof: The nuclear instrument $\mathscr{I}$ obtained from $\{\rho_x; x\in\Lambda\}$ and X has obviously the required properties. Q.E.D.

Now we shall give relations between *a posteriori* states and Cycon–Hellwig conditional expectations.

In the rest of this section, let $\mathscr{I}$ be an instrument of $(\mathscr{M},\Lambda,X)$ and ρ a normal state on $\mathscr{M}$.

Theorem 4.6: If there is an X-integrable disintegration $\{\rho_x; x\in\Lambda\}$ with respect to $(\mathscr{I},\rho)$, then the dual object $\mathscr{E}$ of the nuclear instrument $\mathscr{E}_*$, obtained from $\{\rho_x; x\in\Lambda\}$ and X, is a Cycon–Hellwig conditional expectation with respect to $(\mathscr{I},\rho)$, i.e., $\mathscr{E}(B) = \mathscr{E}_*(B)^*$ and $\mathscr{E}_*(B) = \int_B \rho_x \otimes X(dx)$.

Proof: The assertion follows from the fact tht $\langle \rho_x,a\rangle = E(a|x)$, μ_X-a.e. and from Theorem 4.4. Q.E.D.

Remark: The assumptions in Theorem 4.6 are satisfied in the following cases.

(1) ρ is faithful on the range of X.

(2) $\mathscr{M} = \mathscr{A}^{**}$ for some C^* algebra $\mathscr{A}$ such that $\mathscr{A}^*$ is separable.

(3) $\mathscr{M} = \mathscr{L}(\mathscr{H})$ for some separable Hilbert space $\mathscr{H}$.

(4) $\mathscr{M} = \mathscr{L}(\mathscr{H})$ for some Hilbert space $\mathscr{H}$ and $\mathscr{I}$ is completely positive.

In the case (1), we have $X_a = X$ and hence $\{\rho_x; x\in\Lambda\}$ is X integrable from Proposition A.5. In the cases (2)–(4), strongly $\mathscr{B}(\Lambda)$-measurable families of *a posteriori* states exist from Theorems 4.3 and 4.5 in Ref. 5 and by Proposition A.3 they are X integrable.

For the general case, the construction of Cycon–Hellwig conditional expectations is somewhat complicated.

Theorem 4.7: Let $X = X_a + X_s$ be the Lebesgue decomposition of X and $\{\rho_x; x\in\Lambda\}$ be a disintegration with respect to $(\mathscr{I},\rho)$. Then the dual object of the CP instrument of the form

$$\mathscr{E}_*(B) = \int_B \rho_x \otimes X_a(dx) + \rho_0 \otimes X_s(B), \tag{4.2}$$

for all B in $\mathscr{B}(\Lambda)$, where ρ_0 is an arbitrary normal state on $\mathscr{M}$, is a Cycon–Hellwig conditional expectation with respect to $(\mathscr{I},\rho)$.

Proof: By Theorem A.6, Eq. (4.2) defines a CP$(\mathscr{M}_*)$-valued measure $\mathscr{E}_*$. For all normal states φ on $\mathscr{M}$, we have

$$\langle \mathscr{E}_*(\Lambda)\varphi,1\rangle = \int_\Lambda \langle \rho_x,1\rangle\langle \varphi,X_a(dx)\rangle + \rho_0(1)\langle \varphi,X_s(\Lambda)\rangle$$
$$= \langle \varphi,X_a(\Lambda) + X_s(\Lambda)\rangle$$
$$= 1.$$

Thus $\mathscr{E}_*$ is a CP instrument. Now it is immediate from Theorem 3.2 and Definition 4.2 that the dual object of $\mathscr{E}_*$ is a Cycon–Hellwig conditional expectation with respect to $(\mathscr{I},\rho)$. Q.E.D.

V. STATISTICAL INTERPRETATION OF CONDITIONAL EXPECTATIONS

In the classical probability theory, conditional expectations arise to handle relationships between dependent random variables (cf. Ref. 17, p. 6). Let X and Y be two random variables on a probability measure space $(\Omega,\mathscr{F},P)$. Then Ω may be interpreted as a statistical ensemble and X and Y may be physical observables in the classical statistical mechanics. The measurement of the value of X affects the expectation of Y. If we know that the value of X is in a Borel set B, in symbols $X\in B$, this information means that the probabilistic event $X\in B$ occurs or that the subensemble $\{\omega\in\Omega; X(\omega)\in B\}$ is observed. Thus we can predict the expectation of the result of the successive measurement of Y as the conditional expectation $E(Y|X\in B)$ of Y given $X\in B$, where

$$E(Y|X\in B) = \left(\frac{1}{\mathrm{Prob}(X\in B)}\right)\int_R y\,\mathrm{Prob}(Y\in dy, X\in B). \tag{5.1}$$

If the information of the measurement of X is such that the value of X is x, in symbols $X = x$, then the conditional expectation $E(Y|X = x)$ of Y given $X = x$ is

$$E(Y|X = x) = \left(\frac{1}{\mathrm{Prob}(X\in dx)}\right)\int_R y\,\mathrm{Prob}(Y\in dy, X\in dx), \tag{5.2}$$

where the derivative concerning dx is interpreted as the Radon–Nikodym derivative. In both cases, the conditional expectations are defined only by means of the joint probability distribution $\mathrm{Prob}(Y\in dy, X\in dx)$ of X and Y.

The similar situation arises in statistical interpretation of quantum measurements. Let X be an observable of a quantum system described by a Hilbert space $\mathscr{H}$. In the following, we shall write $X = \int_{\mathbf{R}} xX(dx)$, where $X(dx)$ is the spectral measure corresponding to the self-adjoint operator X. Suppose that a measurement of X is described by the following quantum measuring process (cf. Refs. 4 and 5 for a detailed discussion). Let $\widetilde{X}$ be the observable in the apparatus, described by a Hilbert space $\mathscr{K}$ to show the value of X. The measurement is carried out by the interaction between those two quantum systems during a finite time interval. Let U be the unitary operator on $\mathscr{H}\otimes\mathscr{K}$ of the time evolution of the composite system and σ the preparted state of the apparatus. Let ρ be the initial state of the observed system. We actually measure the value of $\widetilde{X}$ after the interaction as an indirect measurement of the value of X in the initial state, so that we may require that the probability distributions of these two observables in the corresponding states coincide, i.e.,

$$\mathrm{Tr}[\rho X(dx)] = \mathrm{Tr}[U(\rho\otimes\sigma)U^*(1\otimes\tilde{X}(dx))]. \tag{5.3}$$

Let Y be an arbitrary bounded observable in the observed system. Now we suppose that we measure the value of Y immediately after the first measurement. Since $\tilde{X}$ and Y are compatible, we can calculate the joint probability distribution of the values of $\tilde{X}$ and Y,

$$\mathrm{Prob}(Y\in dy,\ \tilde{X}\in dx) = \mathrm{Tr}[U(\rho\otimes\sigma)U^*(Y(dy)\otimes\tilde{X}(dx))]. \tag{5.4}$$

Then by applying Eq. (5.1) to the joint probability distribution obtained in Eq. (5.4) the conditional expectation $E(Y|\tilde{X}\in B)$ of Y given $\tilde{X}\in B$ is obtained by

$$\begin{aligned} E(Y|\tilde{X}\in B) &= \left(\frac{1}{\mathrm{Prob}(\tilde{X}\in B)}\right)\int_B y\,\mathrm{Prob}(Y\in dy,\ \tilde{X}\in B) \\ &= \left(\frac{1}{\mathrm{Tr}[\rho X(B)]}\right)\int_B y\,\mathrm{Tr}[U(\rho\otimes\sigma)U^*(Y(dy)\otimes\tilde{X}(B))] \\ &= \left(\frac{1}{\mathrm{Tr}[\rho X(B)]}\right)\mathrm{Tr}[U(\rho\otimes\sigma)U^*(Y\otimes\tilde{X}(B))]. \end{aligned} \tag{5.5}$$

Similarly, applying Eq. (5.2), we have

$$E(Y|\tilde{X}=x) = (1/\mathrm{Tr}[\rho X(dx)]) \times\mathrm{Tr}[U(\rho\otimes\sigma)U^*(Y\otimes\tilde{X}(dx))]. \tag{5.6}$$

Now we can give a physical interpretation of the conditional expectation with respect to instruments and states given in Sec. II. As shown in Ref. 4 (Sec. 5), the above measuring process naturally corresponds to a CP instrument $\mathscr{I}$ defined by

$$\mathscr{I}(B)\rho = \mathrm{Tr}_{\mathscr{K}}[U(\rho\otimes\sigma)U^*(1\otimes\tilde{X}(B))], \tag{5.7}$$

for all ρ in $\mathscr{T}(\mathscr{H})$, B in $\mathscr{B}(\mathbf{R})$, where $\mathrm{Tr}_{\mathscr{K}}$ stands for the partial trace over $\mathscr{K}$, and $\mathscr{T}(\mathscr{H})$ stands for the space of trace class operators on $\mathscr{H}$. Then we have

$$\begin{aligned}\int_B E(Y|\tilde{X}=x)\mu_X(dx) &= \mathrm{Tr}[U(\rho\otimes\sigma)U^*(Y\otimes\tilde{X}(B))] \\ &= \langle\mathscr{I}(B)\rho,Y\rangle,\end{aligned} \tag{5.8}$$

for all B in $\mathscr{B}(\mathbf{R})$, where $\mu_X(dx) = \langle\mathscr{I}(dx)\rho,1\rangle = \mathrm{Tr}[\rho X(dx)]$. Thus by Definition 2.1 and Theorem 2.2, we have

$$E(Y|x) = E(Y|\tilde{X}=x),\quad \mu_X\text{-a.e.}, \tag{5.9}$$

where the left-hand side is the conditional expectation of Y with respect to $(\mathscr{I},\rho)$.

Let $\{\rho_x;\ x\in\mathbf{R}\}$ be a family of *a posteriori* states with respect to $(\mathscr{I},\rho)$. Then by Ref. 5, ρ_x is interpreted as the state at the instant after the measurement of X, of the subensemble in which the result of the measurement is x. The state change and the conditional expectation are related by

$$\mathrm{Tr}[\rho_x Y] = E(Y|\tilde{X}=x),\quad \mu_X\text{-a.e.} \tag{5.10}$$

[cf. Ref. 5, Eq. (3.7)]. Thus the measurement of X changes the initial state ρ into the *a posteriori* state ρ_x such that the conditional expectation $E(Y|\tilde{X}=x)$ of Y is the expectation $\mathrm{Tr}[\rho_x Y]$ of Y in the state ρ_x.

Let $\mathscr{E}$ be a Cycon–Hellwig conditional expectation with respect to $(\mathscr{I},\rho)$. Putting $E(Y|\tilde{X}=x) = \mathrm{Tr}[\rho_x Y]$, we have

$$\mathscr{E}(B)Y = \int_B E(Y|\tilde{X}=x)X(dx). \tag{5.11}$$

Let $f:\mathbf{R}\to\mathbf{R}$ be such that $f(x) = E(Y|\tilde{X}=x)$. Since X is an observable, we have by Ref. 4 (Proposition 4.3),

$$\mathscr{E}(B)Y = X(B)\mathscr{E}(\Lambda)Y = X(B)f(X) = f(X)X(B). \tag{5.12}$$

Thus by Eq. (5.8) we have the following tricky formula:

$$\mathrm{Tr}[U(\rho\otimes\sigma)U^*(Y\otimes g(\tilde{X}))] = \mathrm{Tr}[\rho f(X)g(X)], \tag{5.13}$$

for all Borel functions $g:\mathbf{R}\to\mathbf{R}$.

This is an analog of a trick of a conditional expectation in classical probability theory.

ACKNOWLEDGMENT

The author wishes to thank Professor H. Umegaki for his useful comments and encouragement.

APPENDIX: INTEGRATION OF FAMILIES OF STATES

In this appendix we shall give a minimum of integration theory for families of states with respect to positive operator valued measures such that the resulting indefinite integrals are completely positive map valued measures.

In what follows, $\mathscr{M}$ is a von Neumann algebra and $(\Lambda,\mathscr{B}(\Lambda))$ is a Borel space.

Let μ be a finite measure on $(\Lambda,\mathscr{B}(\Lambda))$. A family $\{\rho_x;\ x\in\Lambda\}$ of (not necessarily normal) states on $\mathscr{M}$ is called μ *measurable* if the function $x\to\langle\rho_x,a\rangle$ is μ measurable for all a in $\mathscr{M}$. A μ-measurable family $\{\rho_x;\ x\in\Lambda\}$ is called μ *integrable* if for any B in $\mathscr{B}(\Lambda)$, there is a *normal* linear functional ρ_B on $\mathscr{M}$ such that

$$\langle\rho_B,a\rangle = \int_B\langle\rho_x,a\rangle\mu(dx),$$

for all a in $\mathscr{M}$.

Proposition A.1: Let $\{\rho_x;\ x\in\Lambda\}$ be a μ-integrable family of states on $\mathscr{M}$. For any f in $L^1(\Lambda,\mu)$ there is a unique ρ_f in $\mathscr{M}_*$ such that

$$\langle\rho_f,a\rangle = \int_\Lambda f(x)\langle\rho_x,a\rangle\mu(dx), \tag{A1}$$

for all a in $\mathscr{M}$.

Proof: It is easy to see that Eq. (A1) defines a unique ρ_f in $\mathscr{M}^*$ such that $\|\rho_f\|\leqslant\|f\|$. Thus the correspondence $T:f\to\rho_f$ is a bounded linear transformation from $L^1(\Lambda,\mu)$ into $\mathscr{M}^*$. Since $T(\chi_B)\in\mathscr{M}_*$ for all B in $\mathscr{B}(\Lambda)$ by the μ integrability, the range of T is contained in $\mathscr{M}_*$ (cf. Ref. 23, Theorem 3.7.1, p. 77). Q.E.D.

The unique ρ_f satisfying Eq. (A1) will be denoted by $\int_\Lambda f(x)\rho_x\mu(dx)$ or, if $f(x) = \chi_B g(x)$, $\int_B g(x)\rho_x\mu(dx)$.

Let $X:\mathscr{B}(\Lambda)\mapsto\mathscr{M}$ be a positive operator valued measure. For any normal state φ, let μ_φ be the finite measure on $(\Lambda,\mathscr{B}(\Lambda))$ such that $\mu_\varphi(dx) = \langle\varphi,X(dx)\rangle$. A family $\{\rho_x;\ x\in\Lambda\}$ of states on $\mathscr{M}$ is called X *measurable* if it is μ_φ measurable for all normal states φ on $\mathscr{M}$, and it is called X *integrable* if it is μ_φ integrable for all normal states φ on $\mathscr{M}$.

Some conditions for the X integrability will be obtained in the following propositions.

We say that a family $\{\rho_x; x\in\Lambda\}$ is *$\mathscr{B}(\Lambda)$ measurable* if the function $x\to\langle\rho_x, a\rangle$ is $\mathscr{B}(\Lambda)$ measurable for all a in $\mathscr{M}$.

Proposition A.2: If $\mathscr{M}$ is σ finite, then every $\mathscr{B}(\Lambda)$-measurable family $\{\rho_x; x\in\Lambda\}$ of normal states on $\mathscr{M}$ is X integrable for any positive operator valued measure $X:\mathscr{B}(\Lambda)\mapsto\mathscr{M}$.

Proof: Let ρ be a normal state on $\mathscr{M}$. Let $B\in\mathscr{B}(\Lambda)$. Then the functional φ defined by

$$\varphi(a)=\int_B\langle\rho_x,a\rangle\mu_\rho(dx)$$

is obviously positive, linear, and bounded. To show that $\varphi\in\mathscr{M}_*$, we have only to prove that φ is countably additive, since $\mathscr{M}$ is σ finite. Let $\{e_n\}$ be a sequence of mutually orthogonal projections in $\mathscr{M}$. Then by the monotone convergence theorem, we have

$$\varphi\Big(\sum_n e_n\Big)=\int_B\Big\langle\rho_x,\sum_n e_n\Big\rangle\mu_\varphi(dx)$$

$$=\sum_n\int_B\langle\rho_x,e_n\rangle\mu_\varphi(dx),$$

$$=\sum_n\varphi(e_n).$$

Thus the family $\{\rho_x; x\in\Lambda\}$ is μ_φ integrable for all normal states φ on $\mathscr{M}$, so that it is X integrable. Q.E.D.

A family $\{\rho_x; x\in\Lambda\}$ of normal states on $\mathscr{M}$ is called *strongly $\mathscr{B}(\Lambda)$ measurable* if there is a sequence F_n of $\mathscr{B}(\Lambda)$-measurable $\mathscr{M}_*$-valued simple functions such that $\lim_n\|\rho_x-F_n(x)\|=0$, for every x in Λ.

Proposition A.3: Every strongly $\mathscr{B}(\Lambda)$-measurable family $\{\rho_x; x\in\Lambda\}$ of normal states is X integrable for any positive operator valued measures $X:\mathscr{B}(\Lambda)\mapsto\mathscr{M}$.

Proof: The function $x\to\|\rho_x\|$ is bounded and hence μ_φ integrable for all normal states φ on $\mathscr{M}$. Thus the assertion follows from the theory of Bochner integrals (cf. Ref. 23, Theorem 3.7.4, p. 80). Q.E.D.

Let $X:\mathscr{B}(\Lambda)\mapsto\mathscr{M}$ be a semiobservable and ρ be a normal state on $\mathscr{M}$. Denote by μ_X the associate probability distribution, i.e., $\mu_X(dx)=\langle\rho,X(dx)\rangle$. Then we have the following.

Proposition A.4: Any μ_X-integrable family $\{\rho_x; x\in\Lambda\}$ of states is X_a integrable, where X_a is the absolutely continuous part of X with respect to ρ.

Proof: Let φ be a normal state on $\mathscr{M}$ and $a\in\mathscr{M}$. By Lemma 3.1(2), $\langle\varphi,X_a(dx)\rangle<<\mu_X(dx)$ and hence the function $x\to\langle\rho_x,a\rangle$ is $\langle\varphi,X_a(dx)\rangle$ measurable. It follows that $\{\rho_x;x\in\Lambda\}$ is X_a measurable. Let $f(x)=\langle\varphi,X_a(dx)\rangle/\mu_X(dx)$. By Definitions 2.1 and 4.2 and Proposition A.1, we have

$$\int_B\langle\rho_x,a\rangle\langle\varphi,X_a(dx)\rangle=\int_B f(x)\langle\rho_x,a\rangle\mu_X(dx)$$

$$=\Big\langle\int_B f(x)\rho_x\mu_X(dx),a\Big\rangle,$$

so that $\{\rho_x; x\in\Lambda\}$ is $\langle\varphi,X_a(dx)\rangle$ integrable and therefore X_a integrable. Q.E.D.

Let $\mathscr{I}$ be an instrument for $(\mathscr{M},\Lambda,X)$ and ρ be a normal state on $\mathscr{M}$. Denote by X_a the absolutely continuous part of X with respect to ρ. Then we have the following.

Proposition A.5: Any disintegration $\{\rho_x; x\in\Lambda\}$ with respect to $(\mathscr{I},\rho)$ is X_a integrable.

Proof: Follows immediately from Proposition A.4, since $\{\rho_x; x\in\Lambda\}$ is μ_X integrable. Q.E.D.

Denote by $\mathrm{CP}(\mathscr{M}_*)$ the space of completely positive maps on $\mathscr{M}_*$. A set function $\mathscr{I}:\mathscr{B}(\Lambda)\mapsto\mathrm{CP}(\mathscr{M}_*)$ is called a *$\mathrm{CP}(\mathscr{M}_*)$-valued measure* if it is countably additive in the following sense; for each countable disjoint sequence $\{B_j\}$ in $\mathscr{B}(\Lambda)$, we have

$$\mathscr{I}(\cup_i B_i)=\sum_i\mathscr{I}(B_i),$$

in the strong operator topology.

Theorem A.6: Let $X:\mathscr{B}(\Lambda)\mapsto\mathscr{M}$ be a positive operator valued measure and $\{\rho_x; x\in\Lambda\}$ an X-integrable family of states. Then there is a unique $\mathrm{CP}(\mathscr{M}_*)$-valued measure $\mathscr{I}$ such that

$$\langle\mathscr{I}(B)\rho,a\rangle=\int_B\langle\rho_x,a\rangle\langle\rho,X(dx)\rangle. \tag{A2}$$

Proof: Let $\{\rho_x; x\in\Lambda\}$ be an X-integrable family of states on $\mathscr{M}$. Then for any B in $\mathscr{B}(\Lambda)$ and any ρ in $\mathscr{M}_*$, Eq. (A2) determines an element $\mathscr{I}(B)\rho$ in $\mathscr{M}_*$. It is easy to see that the mapping $\mathscr{I}(B):\rho\to\mathscr{I}(B)\rho$ is a positive linear map on $\mathscr{M}_*$. To show the complete positivity of $\mathscr{I}(B)$, let $a_1,\ldots,a_n$ and $b_1,\ldots,b_n$ be finite sequences in $\mathscr{M}$. We have to show that $\Sigma_{ij}b_j^*(\mathscr{I}(B)^*(a_j^*a_i))b_i\geqslant 0$. Let φ be a normal state on $\mathscr{M}$. Let $\mu_{ij}(dx)=\langle\varphi,b_iX(dx)b_j^*\rangle$ and $\mu=\Sigma_{ij}|\mu_{ij}|$, where $|\cdot|$ stands for the total variation. Then for any i,j we have $\mu_{ij}<<\mu$. Let g_{ij} be the Radon–Nikodym derivative $g_{ij}(x)=\mu_{ij}(dx)/\mu(dx)$. Let $f_{ij}(x)=\langle\rho_x,a_j^*a_i\rangle$. Then the $n\times n$ matrix $(f_{ij}(x))$ is positive definite for all x in Λ, and $(g_{ij}(x))$ is positive definite for μ-almost all x in Λ. Thus we have

$$\int_B\sum_{ij}f_{ij}(x)g_{ij}(x)\mu(dx)\geqslant 0.$$

Now the following computations show that $\mathscr{I}(B)$ is completely positive:

$$\Big\langle\varphi,\sum_{ij}b_j^*(\mathscr{I}(B)^*(a_j^*a_i))b_i\Big\rangle$$

$$=\sum_{ij}\langle b_i\varphi b_j^*,\mathscr{I}(B)^*(a_j^*a_i)\rangle$$

$$=\sum_{ij}\langle\mathscr{I}(B)(b_i\varphi b_j^*),a_j^*a_i\rangle$$

$$=\sum_{ij}\int_B\langle\rho_x,a_j^*a_i\rangle\langle b_i\varphi b_j^*,X(dx)\rangle$$

$$=\sum_{ij}\int_B\langle\rho_x,a_j^*a_i\rangle\langle\varphi,b_j^*X(dx)b_i\rangle$$

$$=\int_B\sum_{ij}f_{ij}(x)g_{ij}(x)\mu(dx)\geqslant 0.$$

To show the countable additivity of $B\to\mathscr{I}(B)$, let $\{B_i\}$ be a disjoint sequence in $\mathscr{B}(\Lambda)$. Let $B=\cup_i B_i$ and $B_n=\cup_{i=1}^n B_i$. For any positive φ in $\mathscr{M}_*$, we have $\mathscr{I}(B)\varphi\geqslant\mathscr{I}(B_n)\varphi$ and hence by Eq. (A2),

$$\|\mathscr{I}(B)\varphi - \mathscr{I}(B_n)\varphi\| = \langle \mathscr{I}(B)\varphi - \mathscr{I}(B_n)\varphi, 1\rangle$$
$$= \langle \varphi, X(B)\rangle - \langle \varphi, X(B_n)\rangle$$
$$\to 0, \quad \text{as } n \to \infty.$$

For any φ in $\mathscr{M}_*$, there are positive φ_i's in $\mathscr{M}_*$ and α_i's in $\mathbf{C}$ $(i = 1,...,4)$ such that $\varphi = \Sigma_i\, \alpha_i \varphi_i$. It follows that

$$\|\mathscr{I}(B)\varphi - \mathscr{I}(B_n)\varphi\| < \sum_i |\alpha_i| \, \|\mathscr{I}(B)\varphi_i - \mathscr{I}(B_n)\varphi_i\|$$
$$\to 0, \quad \text{as } n \to \infty.$$

Q.E.D.

We shall write $\mathscr{I}(B) = \int_B \rho_x \otimes X(dx)$ for the CP($\mathscr{M}_*$)-valued measure $\mathscr{I}$ defined by Eq. (A2).

Let ρ be a normal state on $\mathscr{M}$ and $X: \mathscr{B}(\Lambda) \to \mathscr{M}$ a positive operator valued measure. Then the constant function $x \to \rho_x = \rho$ is obviously X integrable. In this case, we shall write

$$\rho \otimes X(B) = \int_B \rho_x \otimes X(dx).$$

[1] J. von Neumann, *Mathematical Foundations of Quantum Mechanics*, translated by R. T. Beyer (Princeton U. P., Princeton, NJ, 1955).
[2] A. N. Kolmogorov, *Foundations of Probability Theory* (Chelsea, New York, 1950).
[3] M. Ozawa, "Conditional expectation and repeated measurements of continuous quantum observables," Lecture Notes Math. **1021**, 518 (1983).
[4] M. Ozawa, "Quantum measuring processes of continuous observables," J. Math. Phys. **25**, 79 (1984).
[5] M. Ozawa, "Conditional probability and a posteriori states in quantum mechanics," Publ. RIMS Kyoto Univ. (to appear).
[6] H. Umegaki, "Conditional expectation in an operator algebra. I; II; III; IV," Tôhoku Math. J. **6**, 177 (1954); **8**, 86 (1956); Kōdai Math. Sem. Rap. **11**, 51 (1959); **14**, 59 (1962).
[7] S-T. C. Moy, "Characterizations of conditional expectations as a transformation on function spaces," Pac. J. Math. **4**, 47 (1954).
[8] J. Tomiyama, "On the projection of norm one in W^*-algebra," Proc. Jpn. Acad. **33**, 608 (1957).
[9] M. Nakamura and H. Umegaki, "On von Neumann's theory of measurements in quantum statistics," Math. Japon. **7**, 151 (1962).
[10] J. Tomiyama, "On the projection of norm one in W^*-algebras. III," Tôhoku Math. J. **11**, 125 (1958).
[11] W. B. Arveson, "Analyticity in operator algebras," Am. J. Math. **89**, 578 (1967).
[12] A. de Korvin, "Complete sets of expectations on von Neumann algebras," Quart. J. Math. Oxford Ser. (2) **22**, 135 (1971).
[13] E. Størmer, "On projection maps of von Neumann algebras," Math. Scand. **30**, 46 (1972).
[14] E. B. Davies, *Quantum Theory of Open Systems* (Academic, London, 1976).
[15] E. B. Davies and J. T. Lewis, "An operational approach to quantum probability," Commun. Math. Phys. **17**, 239 (1970).
[16] H. Cycon and K-E. Hellwig, "Conditional expectations in generalized probability theory," J. Math. Phys. **18**, 1154 (1977).
[17] L. Breiman, *Probability* (Addison–Wesley, London, 1968).
[18] E. B. Davies, "On the repeated measurement of continuous observables in quantum mechanics," J. Funct. Anal. **6**, 318 (1970).
[19] S. P. Gudder, "Axiomatic operational quantum mechanics," Rep. Math. Phys. **16**, 147 (1979).
[20] P. R. Halmos, *Introduction to Hilbert Space and the Theory of Spectral Multiplicity* (Chelsea, New York, 1951).
[21] S. K. Berberian, *Notes on Spectral Theory* (Van Nostrand, Princeton, NJ, 1966).
[22] I. T. Tulcea and I. C. Tulcea, *Topics in the Theory of Lifting* (Springer, New York, 1969).
[23] E. Hille and P. S. Phillips, *Functional Analysis and Semigroups* (Am. Math. Soc., Providence, RI, 1957), 2nd ed.

On the one- and two-dimensional Toda lattices and the Painlevé property

J. D. Gibbon
Department of Mathematics, Imperial College, London SW7 2BZ, England
M. Tabor
Department of Applied Physics and Nuclear Engineering, Columbia University, New York, New York 10027

(Received 13 February 1985; accepted for publication 4 April 1985)

The Toda lattice and the two-dimensional Toda lattice (2-DTL) are shown to possess a type of "Painlevé property" that is based on the use of separate "singular manifolds" for each dependent variable. The isospectral problem for the 2-DTL found by both Mikhailov and by Fordy and Gibbons can be simply and logically derived from this analysis. Some remarks are made about the connection between our work and independent work of Kametaka and Airhault on the relationship between the Toda lattice and the second Painlevé transcendent.

I. INTRODUCTION

Studies using the idea of the Painlevé property to investigate the integrability of differential equations have so far been fairly successful. Partial differential equations (pde's) known to be integrable by the inverse scattering transform (IST) have been found to pass the Painlevé test, either in the ordinary differential equation (ode) form proposed by Ablowitz *et al.*[1] or the pde form proposed by Weiss *et al.*[2,3] For ode's such as the Henon–Heiles equations, which are only integrable for certain parameter values, the equations can be shown to pass the Painlevé test for precisely these values.[4] Nevertheless, in the case of systems of differential difference equations such as the Toda lattice

$$\ddot{Q}_n = \exp(Q_{n-1} - Q_n) - \exp(Q_n - Q_{n+1}), \tag{1.1}$$

problems arise in applying the Painlevé tests as they have currently been formulated. Equation (1.1), first suggested by Toda (see references in Ref. 5) as a nonlinear chain equation, has been shown to be integrable by Flaschka,[6] Manakov,[7] and Henon.[8] Bountis, Vivaldi, and Segur[9] have shown that the three-particle Toda lattice possesses the Painlevé property by expanding each of the dependent variables as a Laurent series about a common pole position t_0. As we will demonstrate, this approach does not lend itself to generalization to a system of N particles (N large), with or without periodic boundary conditions. In this paper, we partly resolve this problem and, furthermore, we demonstrate that our results are almost trivially extendable to the two-dimensional Toda lattice (2-DTL)

$$Q_{n,xt} = \exp(Q_{n-1} - Q_n) - \exp(Q_n - Q_{n+1}), \tag{1.2}$$

which, while no longer a set of differential difference equations, poses the same difficulties as (1.1). Equation (1.2) has been studied by various people such as Darboux,[10] Mikhailov,[11] and Fordy and Gibbons.[12] It was shown, in the two latter references, that these equations were completely integrable by inverse scattering and we shall show that their results fall out very easily once we have found a suitable formulation of a Painlevé-type property.

Some of the first solutions of (1.1) were identified by Toda[5] before it was solved by the IST. By transforming the Q_n variables such that

$$Q_n = S_{n-1} - S_n, \tag{1.3}$$

Eq. (1.1) becomes

$$\epsilon + \ddot{S}_n = \exp(S_{n+1} + S_{n-1} - 2S_n), \tag{1.4}$$

where, for pure soliton solutions, we need $\epsilon = 1$ (Toda's original form). Equation (1.4) is reducible to homogeneous form by the transformation

$$S_n = \log f_n, \tag{1.5}$$

thereby giving

$$\epsilon + \frac{\partial^2}{\partial t^2}\log f_n = \frac{f_{n+1} f_{n-1}}{f_n^2}. \tag{1.6}$$

The single-soliton solution ($\epsilon = 1$) is

$$f_n = \cosh(an - \omega t), \tag{1.7a}$$

$$\ddot{S}_n = \omega^2 \operatorname{sech}^2(an - \omega t), \tag{1.7b}$$

$$\omega = \sinh a. \tag{1.7c}$$

Toda[5] found a two-soliton solution via (1.6) that was generalized by Hirota.[13] In addition, there is a close connection between results for the Toda lattice and rational solutions of the second Painlevé transcendent that was noted by Kametaka.[14] These rational solutions are expressed in terms of polynomials introduced by Yablonskii[15] and Vorobiev.[16] Variables defined in terms of these polynomials are then seen to satisfy the Toda lattice expressed in a form similar to Flaschka's.[6] Some of Kametaka's results were previously derived by Airhault,[17] although in that work, the connection with the Toda lattice was not made.

II. THE FORMULATION OF A PAINLEVÉ-TYPE EXPANSION

The Toda lattice equations (1.1) can be cast in the form

$$\dot{p}_n = p_n(q_n - q_{n+1}), \tag{2.1a}$$

$$\dot{q}_n = p_{n-1} - p_n, \tag{2.1b}$$

where $q_n = \dot{Q}_n$ and $p_n = \exp(Q_n - Q_{n+1})$. While similar to Flaschka's[6] variables (b_n, a_n), the variables q_n and p_n are not quite the same. The relation between them is $q_n = -2b_n$ and $p_n = 4a_n^2$. To find the p_n and q_n variables in terms of Toda's function f_n, we note that we need to write

$$p_n = \frac{\partial^2}{\partial t^2}\log f_n. \tag{2.2a}$$

 0022-2488/85/081956-05$02.50

It follows from (2.1b) that

$$q_n = \frac{\partial}{\partial t} \log \frac{f_{n-1}}{f_n} \tag{2.2b}$$

and integration of (2.1a) now gives

$$\frac{f_{n+1} f_{n-1}}{f_n^2} = \frac{\partial^2}{\partial t^2} \log f_n + \epsilon, \tag{2.3}$$

which is exactly (1.5). Here, ϵ arises as a constant of integration in (2.3). We remark, however, that the ϵ term in (2.3) can be absorbed by a transformation $f_n \to f_n \exp(-\epsilon t^2/2)$ and so it can be left out or kept in Eq. (2.3) at will.

In studying the three-particle Toda lattice, Bountis, Vivaldi, and Segur[9] used Flaschka's variables (a_n, b_n) to find a local Laurent expansion for each dependent variable about a common pole position t_0. As an aside, it is possible to derive a simple polelike solution directly from (2.3) ($\epsilon = 0$). This is given by the following:

$$f_n(t) = g_n(t - t_0)^{-n^2}, \tag{2.4a}$$

provided the g_n satisfy

$$g_{n+1} g_{n-1} = n^2 g_n^2. \tag{2.4b}$$

Treating (2.4b) as a recursion relation with $g_0 = 1$ and $g_{-1} = 0$, there is a nonzero solution of (2.4b)

$$g_n = (n-1)^2 (n-2)^4 (n-3)^6 \ldots 2^{2(n-2)} g_1^n. \tag{2.4c}$$

This solution is exactly consistent with the standard "leading order analysis"[1–4] when applied to the Toda equations in the form ($r_n = \dot{S}_n$)

$$\ddot{r}_n = (\dot{r}_n + \epsilon)(r_{n+1} + r_{n-1} - 2r_n), \tag{2.5}$$

which is another form used by Toda[5] and which is the differentiated version of (1.4) [for (2.5) we find $r_n = -n^2(t - t_0)^{-1}$ at leading order]. Of course, (1.7a) and (2.4) are only two possible forms of solution and we need a method that will cover more than just these. The approach of Bountis, Vivaldi, and Segur rapidly becomes cumbersome as the number of particles increases, and is not conducive to constructing global solutions. Here we propose to introduce a set of "singular manifolds" $\{\phi_n(t)\}$ akin to those introduced in Ref. 2, which can be associated with *each* particle. This has the advantage of enabling us to look for a connection with the IST rather than just identifying single-valued local Laurent expansions about a common pole position.

Consequently, motivated by (2.2a), (2.2b), and the "truncated Painlevé expansions" used in Refs. 2 and 3, we postulate that solutions of the (p,q) equations (2.1a) and (2.1b) should take the form

$$p_n = p_n^{(0)}/\phi_n^2 + p_n^{(1)}/\phi_n + p_n^{(2)}, \tag{2.6a}$$

$$q_n = q_{n-1}^{(0)}/\phi_{n-1} + q_n^{(0)}/\phi_n + q_n^{(1)}, \tag{2.6b}$$

where we are demanding that *both* (p_n, q_n) and $(p_n^{(2)}, q_n^{(1)})$ are solutions of (2.1a) and (2.1b). From now on, we shall relabel the second solution set $(p_n^{(2)}, q_n^{(1)})$ as $(\bar{p}_n, \bar{q}_n)$.

From Eq. (2.1), we can equate powers of ϕ_n^{-3} since no terms of this type can occur elsewhere. This gives

$$q_n^{(0)} = \dot{\phi}_n. \tag{2.7}$$

Using this in (2.1b) and equating powers of ϕ_n^{-2} we find

$$p_n^{(0)} = -\dot{\phi}_n^2. \tag{2.8}$$

The remnant of (2.1b) immediately yields

$$p_n^{(1)} = \ddot{\phi}_n, \tag{2.9}$$

provided $p_n^{(2)}$ and $q_n^{(1)}$ satisfy (2.1a) and (2.1b). Use of (2.7)–(2.9) in the rest of (2.1a) yields, after a little rearrangement,

$$\frac{\partial^3}{\partial t^3} \log \phi_n = \left(\frac{\partial^2}{\partial t^2} \log \phi_n \right) \left(\frac{\dot{\bar{p}}_n}{\bar{p}_n} + \frac{\dot{R}}{R} \right) + \bar{p}_n \frac{\dot{R}}{R}, \tag{2.10}$$

where $R = \phi_{n+1}\phi_{n-1}/\phi_n^2$. Designating

$$L = \frac{\partial^2}{\partial t^2} \log \phi_n, \tag{2.11}$$

we find that (2.10) can be put in the form

$$\frac{\partial}{\partial t}\left(\frac{L}{\bar{p}_n} \right) = \frac{\dot{R}}{R}\left(1 + \frac{L}{\bar{p}_n} \right), \tag{2.12}$$

which immediately integrates to

$$\frac{\partial^2}{\partial t^2} \log \phi_n = \bar{p}_n \left[\frac{\phi_{n+1}\phi_{n-1}}{\phi_n^2} - 1 \right]. \tag{2.13}$$

Equation (2.13) is the basic equation from which we shall now work along with (2.6a) and (2.6b), which now become

$$p_n = \frac{\partial^2}{\partial t^2} \log \phi_n + \bar{p}_n, \tag{2.14a}$$

$$q_n = \frac{\partial}{\partial t} \log \frac{\phi_{n-1}}{\phi_n} + \bar{q}_n. \tag{2.14b}$$

Equations (2.13) and (2.14) constitute a Bäcklund transformation for the Toda lattice. We remark that the simple form in which we have derived this Bäcklund transformation is dependent on choosing suitable initial variables, which turned out to be the (p_n, q_n) variables. These are clearly more suitable than the Q_n variables of (1.1) since, in terms of the f_n, these Q_n take the form $Q_n = \log(f_{n-1}/f_n)$ and we appear to have a logarithmic singularity with which to cope. Using (2.1) instead removes this logarithm and enables us to then use the truncated expansion in the ϕ_n variables without any problems.

Results for the 2-DTL can now be obtained immediately. Equations (2.1a) and (2.1b) are replaced by

$$p_{n,x} = p_n(q_n - q_{n+1}), \tag{2.15a}$$

$$q_{n,t} = p_{n-1} - p_n, \tag{2.15b}$$

with $q_n = Q_{n,x}$ and $p_n = \exp(Q_n - Q_{n+1})$.

We obtain the equivalent of (2.13), which is

$$\frac{\partial^2}{\partial x \partial t} \log \phi_n = \bar{p}_n \left[\frac{\phi_{n+1}\phi_{n-1}}{\phi_n^2} - 1 \right], \tag{2.16}$$

and Eqs. (2.14a) and (2.14b) are replaced by

$$p_n = \frac{\partial^2}{\partial x \partial t} \log \phi_n + \bar{p}_n, \tag{2.17a}$$

$$q_n = \frac{\partial}{\partial x} \log \frac{\phi_{n-1}}{\phi_n} + \bar{q}_n. \tag{2.17b}$$

III. THE INVERSE SCATTERING TRANSFORM

In looking for linear isospectral operators, it is easier to work with the 2-DTL first since the x and t variables allow us to keep the two halves of the calculation separate. We shall begin from Eq. (2.16) written in the form

$$\phi_n\phi_{n,xt} - \phi_{n,x}\phi_{n,t} = \bar{p}_n(\phi_{n+1}\phi_{n-1} - \phi_n^2). \tag{3.1}$$

It was shown in Gibbon *et al.*[18] that the functions ϕ_n were related to the eigenfunctions of the KdV equation and other equations either as direct eigenfunctions for rational solutions or as the integral of squared eigenfunctions for other types of solutions. With this in mind we shall seek a *linear* problem of the form

$$\phi_{n,x} = \alpha_{n+1}\phi_n + \beta_n\phi_{n+1}, \tag{3.2a}$$

$$\phi_{n,t} = \gamma_n\phi_{n-1} + \delta_{n-1}\phi_n. \tag{3.2b}$$

When substituted into (3.1) these equations give

$$\begin{aligned}\phi_n^2(\alpha_{n+1,t} + \beta_n\gamma_{n+1}) - \phi_{n+1}\phi_{n-1}(\beta_n\gamma_n)\\ + \phi_n\phi_{n+1}(\beta_n\delta_n - \beta_{n,t} - \beta_n\delta_{n-1})\\ = \bar{p}_n(\phi_{n+1}\phi_{n-1} - \phi_n^2).\end{aligned} \tag{3.3}$$

Equating coefficients of ϕ_n^2, etc., we find

$$\alpha_{n+1,t} + \beta_n\gamma_{n+1} = -\bar{p}_n, \tag{3.4a}$$

$$\beta_n\gamma_n = -\bar{p}_n, \tag{3.4b}$$

$$\beta_{n,t} = \beta_n(\delta_n - \delta_{n-1}). \tag{3.4c}$$

We can also cross differentiate Eqs. (3.2a) and (3.2b) to find

$$\gamma_{n,x}/\gamma_n = \alpha_{n+1} - \alpha_n, \tag{3.5a}$$

$$\alpha_{n+1,t} + \beta_n\gamma_{n+1} = \gamma_n\beta_{n-1} + \delta_{n-1,x}. \tag{3.5b}$$

We now need to be able to choose the coefficients α, β, γ, and δ such that these equations are consistent with (2.15a) and (2.15b). This consistency comes about in choosing the $\beta_n = \text{const} = \lambda$, thereby giving from (3.5a)

$$\gamma_n = -\lambda^{-1}\bar{p}_n. \tag{3.6}$$

Equation (3.5a) then gives

$$\alpha_n = -\bar{q}_n. \tag{3.7}$$

Equation (3.4a) is now consistent with (2.15b) as is (3.5b) provided that the δ_n are constant. In fact, without loss of generality we can take all $\delta_n = 0$. The linear problem (3.2) now becomes

$$\phi_{n,x} = -\bar{q}_{n+1}\phi_n + \lambda\phi_{n+1}, \tag{3.8a}$$

$$\phi_{n,t} = -\lambda^{-1}\bar{p}_n\phi_{n-1}, \tag{3.8b}$$

and we see that the ϕ_n turn out to be eigenfunctions and λ the spectral parameter. Equations (3.8) are the same as the spectral problem derived by Fordy and Gibbons[12] except that they had a difference in sign in the original equations (their $\theta_n = -Q_n$).

For the Toda lattice itself, the x in (3.8a) is replaced by t and upon rearrangement, the spectral problem becomes

$$\phi_{n,t} = -\lambda^{-1}\phi_{n-1}\bar{p}_n, \tag{3.9a}$$

$$\bar{q}_{n+1}\phi_n = \lambda\phi_{n+1} + \lambda^{-1}\phi_{n-1}\bar{p}_n. \tag{3.9b}$$

This is not the same as Flaschka's form of the Lax pair for the Toda lattice,[6] but it can be transformed into it by a gauge transformation.

We now show that our result (3.1), which comes from a Painlevé-type analysis, is consistent with results which can be derived by Hirota's method.

For the 2-DTL, we know that the equation reduces to

$$\epsilon + \frac{\partial^2}{\partial x\,\partial t}f_n = \frac{f_{n+1}f_{n-1}}{f_n^2}, \tag{3.10}$$

which is the generalization of (2.3). Gibbon *et al.*[18] have shown that Hirota's method of reducing an equation down to a bilinear form is intimately related to the Painlevé method for pde's. They showed that if f is one solution of a bilinear equation and $\bar{f}$ is another solution, then the two are related by $f = \phi\bar{f}$. In this case we shall write $f_n = \phi_n\bar{f}_n$, where both f_n and $\bar{f}_n$ are solutions of (4.1). Equation (3.10) now becomes

$$\frac{\partial^2}{\partial x\,\partial t}\log\phi_n = \bar{p}_n\left(\frac{\phi_{n+1}\phi_{n-1}}{\phi_n^2} - 1\right), \tag{3.11}$$

which is exactly the equation derived from the Painlevé ansatz in Sec. II [see Eq. (2.1b)]. Hence we have, as expected, a consistency between results from Hirota's method and the Painlevé ansatz made in the (p,q) variables. Indeed, if we label f_n and $\bar{f}_n$ as adjacent solutions $f_n^{(i)}$ and $f_n^{(i-1)}$ in the set $\{f_n^{(i)}\}$ of solutions of (3.10) with corresponding $\phi_n^{(i-1)}$, then we find that

$$f_n^{(N)} = \prod_{i=0}^{N-1}\phi_n^{(i)}. \tag{3.12}$$

We can also use the relation $f_n = \phi_n\bar{f}_n$ to find a standard Bäcklund transformation for either the 2-DTL or the Toda lattice itself. Equation (3.8a), when translated into the f_n functions, becomes

$$\frac{\partial}{\partial x}\log\frac{\bar{f}_n}{f_{n-1}} = -\lambda\frac{\bar{f}_{n-1}f_n}{f_{n-1}\bar{f}_n}, \tag{3.13}$$

where we have dropped one down in the n ladder. From (3.13) we can subtract the same equation with $n \to n-1$, giving finally

$$\begin{aligned}(Q_{n-1} - \bar{Q}_n)_x = \lambda\{&\exp(\bar{Q}_{n-1} - Q_{n-1})\\ &- \exp(\bar{Q}_n - Q_n)\}.\end{aligned} \tag{3.14}$$

The same exercise applied to (3.8b) gives

$$\begin{aligned}(\bar{Q}_{n-1} - Q_{n-1})_t = \lambda^{-1}\{&\exp(Q_{n-2} - \bar{Q}_{n-1})\\ &- \exp(Q_{n-1} - \bar{Q}_n)\}.\end{aligned} \tag{3.15}$$

Equations (3.14) and (3.15) constitute a Bäcklund transformation between two solutions $Q_n = \log(f_{n-1}/f_n)$ and $\bar{Q}_n = \log(\bar{f}_{n-1}/\bar{f}_n)$. The equivalent result for the Toda lattice is found by replacing x by t in (3.14). Equations (3.14) and (3.15) are the Bäcklund transformations found by Fordy and Gibbons.[12]

IV. A DISCRETE SCHWARZIAN DERIVATIVE

The Schwarzian derivative[3] has arisen many times in Painlevé calculations. It is defined by

$$\{\phi;x\} \equiv \frac{\phi_{xxx}}{\phi_x} - \frac{3}{2}\left(\frac{\phi_{xx}}{\phi_x}\right)^2 \tag{4.1a}$$

$$= \left(\frac{\phi_{xx}}{\phi_x}\right)_x - \frac{1}{2}\left(\frac{\phi_{xx}}{\phi_x}\right)^2, \tag{4.1b}$$

and possesses the remarkable property that it remains invariant under fractional linear transformations (i.e., under the Möbius group) $\phi \to (a\phi + b)/(c\phi + d)$. Note that in the form given in (4.1b), we have a type of Miura transformation $v_x - \frac{1}{2}v^2$ with $v = \phi_{xx}/\phi_x$. It arises from the KdV equation by combining two equations together. Weiss et al.[2] showed that for two solutions of the KdV equation, u and $\bar{u}$, then

$$u=\bar{u}+\frac{\partial^2}{\partial x^2}\log\phi, \tag{4.2a}$$

$$\phi_x\phi_t+12\bar{u}\phi_x^2+4\phi_x\phi_{xxx}-3\phi_{xx}^2=0, \tag{4.2b}$$

$$\phi_{xt}+12\bar{u}\phi_{xx}+\phi_{xxxx}=0. \tag{4.2c}$$

Eliminating $\bar{u}$ from (4.2b) and (4.2c) and integrating once, we find

$$\phi_t/\phi_x+\{\phi;x\}=\lambda. \tag{4.3}$$

Clearly Eq. (4.3) is invariant under fractional linear transformations. The Schwarzian derivative has been used to determine various isospectral deformations.[3]

The question now arises as to whether we can obtain a similar result in the discrete case of the Toda lattice. To do this we need more than just Eq. (2.13). We take guidance from the well-known fact that the Toda lattice reduces to the Boussinesq equation in the continuum limit (see the Appendix) and that the discrete equivalent of the two equations arising out of the Painlevé analysis for this latter equation must match in the continuum limit. We are therefore required to split Eq. (3.1) in the following way to make it match with Eqs. (A4a) and (A4b):

$$\phi_{n,tt}=\bar{p}_n(\phi_{n+1}+\phi_{n-1}-2\phi_n), \tag{4.4a}$$

$$\phi_{n,t}^2=\bar{p}_n\{\phi_n(\phi_{n+1}+\phi_{n-1})-\phi_n^2-\phi_{n+1}\phi_{n-1}\}. \tag{4.4b}$$

We recover (2.13) by multiplying (4.4a) by ϕ_n and subtracting (4.4b). Division of these two equations now gives

$$\frac{\phi_{n,tt}}{\phi_{n,t}^2}=\frac{\phi_{n+1}+\phi_{n-1}-2\phi_n}{\phi_n(\phi_{n+1}+\phi_{n-1})-\phi_n^2-\phi_{n+1}\phi_{n-1}}. \tag{4.5}$$

After a lengthy calculation, we do indeed find that (4.5) is invariant under fractional linear transformations. Consequently it is the equivalent of (4.3) for the Toda lattice and we therefore conclude that the idea of the Schwarzian derivative can go over to discrete systems. [We speculate that this property will be connected with the group SL(N,R).]

V. SOME REMARKS ON PAINLEVÉ II AND THE WORK OF AIRHAULT AND KAMETAKA

One of the main aims of this paper has been to show how the Painlevé method for pde's can, with suitable modification, be used on the Toda lattice and the 2-DTL. This method, which takes the form of postulating a truncated Laurent expansion in the (p,q) variables, has the advantage over the ode method of yielding results which give an isospectral problem with very little extra work. Establishing the existence of a Laurent series around a common pole position for each dependent variable generally gives little further information. This idea can be illustrated by considering Painlevé-II (P-II) itself

$$\ddot{q}=2q^3+tq+\alpha. \tag{5.1}$$

As was originally established by Painlevé,[19] the existence of a Laurent expansion

$$q(t)=(t-t_0)^{-1}\sum_{j=0}^{\infty}a_j(t-t_0)^j \tag{5.2}$$

is the first step in showing that no movable critical points exist. This procedure does not tell us anything about any special solutions which might exist. One was determined by Gambier[20] by other means, as Airhault has pointed out.[17] This solution can be found quite logically by using the truncated Laurent series method which introduces the singular manifold. Since the leading order is -1, we let

$$q=q_0/\phi+q_1, \tag{5.3}$$

where we require both q and q_1 to be solutions of P-II. Substitution of (5.3) into (5.1) gives, at various orders of ϕ,

$$\phi^{-3}:\ q_0=\pm\phi_t, \tag{5.4a}$$

$$\phi^{-2}:\ q_1=\mp\tfrac{1}{2}(\phi_{tt}/\phi_t), \tag{5.4b}$$

$$\phi^{-1}:\ \phi_{ttt}/\phi_t-\tfrac{3}{2}(\phi_{tt}/\phi_t)^2=t, \tag{5.4c}$$

$$\phi^{0}\ :\ \ddot{q}_1=2q_1^3+tq_1+\alpha. \tag{5.4d}$$

Equation (5.4d) is clearly satisfied if we require q_1 to be a solution of P-II, but we also need a substitute (5.4b) into this to check for consistency. If we define the Schwarzian derivative in (5.4c) to be S, we find (5.4d) becomes

$$S_t-(\phi_{tt}/\phi_t)(S-t)=\mp 2\alpha. \tag{5.5}$$

Since $S=t$ we note that (5.5) is satisfied if and only if

$$\alpha=\mp\tfrac{1}{2}. \tag{5.6}$$

Furthermore, from (5.4b) and (5.4c), we find that

$$\dot{q}_1\pm q_1^2=\tfrac{1}{2}t, \tag{5.7}$$

and so the substitution $q_1=\pm\dot{\gamma}/\gamma$ gives Airy's equation

$$\ddot{\gamma}+\tfrac{1}{2}t\gamma=0. \tag{5.8}$$

This solution of P-II ($q_1=\dot{\gamma}/\gamma$ with $\alpha=-\frac{1}{2}$) is the one found by Gambier[20] and used by Airhault,[17] who went on to generalize this result to values of α for which $\alpha=n-\frac{1}{2}$ (n is an integer). We remark at this point that Eq. (5.4b) gives a solution for ϕ itself which is

$$\phi=\int\gamma^{-2}\,dt. \tag{5.9}$$

This is exactly the form found for ϕ in the KdV equation in similarity variables,[18] which can be found by taking (4.2b) and (4.2c) and going into similarity variable form with $z=x/t^{1/3}$. The resulting formula for ϕ is (5.9) with z replacing t. We conclude therefore that using the truncated Laurent type of expansion as in (5.2), which is successful for pde's, is also possible for ode's when special solutions are required. This approach was first used by Weiss[21] to solve the integrable cases of the Henon–Heiles system. There is, of course, the further solution $q=\phi_t/\phi+q_1$ with ϕ given by (5.9) and $q_1=\dot{\gamma}/\gamma$ for $\alpha=-\frac{1}{2}$. The more general result found by Airhault[17] has also been found by Weiss[22] using a technique similar to that used here.

Both Airhault[17] and Kametaka[14] discuss special solutions of P-II when $\alpha=n$ (n is an integer). Consider Eq. (1.6) with $\epsilon=0$ and take $f_n\to f_n\exp(-t^3/24)$. We now have

$$f_{n+1}f_{n-1}=f_n^2\left(\frac{\partial^2}{\partial t^2}\log f_n-\frac{t}{4}\right). \tag{5.10}$$

Starting with $f_0=1$, it is possible to calculate a series of polynomials for $f_n(t)$. The f_n are related to the q_n by

$$q_n=\frac{\partial}{\partial t}\log\frac{f_{n-1}}{f_n}. \tag{5.11}$$

Airhault[17] noted that q_n is a solution of P-II when $\alpha = n$, although it was subsequently Kametaka[14] who pointed out that (5.10) is related to the Toda lattice. Explicitly, this is done by using (5.10) to define the variable p_n as

$$p_n = \frac{f_{n+1} f_{n-1}}{f_n^2} = \frac{\partial^2}{\partial t^2} \log f_n - \frac{t}{4}. \tag{5.12}$$

The first and second of these equalities, in conjunction with (5.11), are then seen to satisfy Eqs. (2.1a) and (2.1b), respectively. Kametaka also pointed out that the polynomials f_n are those of Yablonskii[15] and Vorobiev.[16] These seem to have been the first to have shown that if q_n is related to the polynomials f_n, then q_n satisfies P-II for $\alpha = n$. Strangely, Yablonskii's and Vorobiev's polynomials are for *each particle*, i.e., there is a polynomial associated with each n, which is the particle label!

ACKNOWLEDGMENTS

This collaboration was made possible through a NATO grant. Michael Tabor is supported in part by U.S. Department of Energy Grant No. DEFG02-84ER13190 and an Alfred P. Sloan Research Fellowship.

APPENDIX: THE CONTINUUM LIMIT

Using (1.4) with $\epsilon = 1$ and going into the continuum limit $x = \delta_n$, where δ is the lattice spacing, we find that

$$\cdots(\delta^4/12)(S_{xxxx} + 6S_{xx}^2) + \delta^2 S_{xx} - S_{tt} = 0. \tag{A1}$$

Writing $u = S_{xx}$ and differentiating (A1) twice with respect to x, we find that (A1) becomes the Boussinesq equation

$$(\delta^4/12)(u_{xx} + 6u^2)_{xx} + \delta^2 u_{xx} - u_{tt} = 0. \tag{A2}$$

In (A2) we have neglected terms δ^6 and higher. Formally taking $\delta = 1$ and rescaling to remove the factor of $\frac{1}{12}$ in (A2), we can perform a Painlevé analysis on (A2) with

$$u = \phi^{-2} \sum_{j=0}^{\infty} u_j \phi^j. \tag{A3}$$

Reworking the calculation on (A2) already performed previously[2] in which there was no u_{xx} term we find that the Bäcklund transformation between two solutions u and $\bar{u}$ [where $\bar{u}$ is u_2 in (A3)] is

$$u = \bar{u} + \frac{\partial^2}{\partial x^2} \log \phi, \tag{A4a}$$

$$4\phi_x \phi_{xxx} - 3\phi_{xx}^2 + 12\bar{u}\phi_x^2 + \phi_x^2 - \phi_t^2 = 0, \tag{A4b}$$

$$\phi_{xxxx} + 12\bar{u}\phi_{xx} + \phi_{xx} - \phi_{tt} = 0. \tag{A4c}$$

Weiss had obtained the isospectral operator for the Boussinesq equation out of (A4).[3] It is Eqs. (A4b) and (A4c), which are the two separate equations, which tell us how to break down Eq. (3.1), with the linear equation in ϕ, (A4c) becoming (4.4a).

[1] M. J. Ablowitz, A. Ramani, and H. Segur, J. Math. Phys. **21**, 715 (1980).
[2] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983).
[3] J. Weiss, J. Math. Phys. **24**, 1405 (1983).
[4] Y. F. Chang, M. Tabor, and J. Weiss, J. Math. Phys. **23**, 531 (1982).
[5] M. Toda, Prog. Phys. Suppl. **45**, 174 (1970); Phys. Rep. **18**, 1 (1975).
[6] H. Flaschka, Phys. Rev. **139**, 1923 (1974).
[7] S. Manakov, Sov. Phys. JETP **40**, 269 (1975).
[8] M. Henon, Phys. Rev. **139**, 1921 (1974).
[9] T. Bountis, F. Vivaldi, and H. Segur, Phys. Rev. A **25**, 1257 (1982).
[10] G. Darboux, *Lecons sur la theore generale des surface et les applications geometriques du calcul infinitesimall II* (Chelsea, New York, 1972).
[11] A. Mikhailov, JETP Lett. **30**, 414 (1979).
[12] A. P. Fordy and J. Gibbons, Commun. Math. Phys. **77**, 21 (1980).
[13] R. Hirota, in *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer, Heidelberg, 1980).
[14] Y. Kametaka, Proc. Jpn. Acad. **59**, 358, 407 (1983).
[15] A. I. Yablonskii, Vestsi Akad. Navuk. B. SSR. Ser. Fiz.-Tekh. Navuk **3** (1959).
[16] A. P. Vorobiev, Differencial'nye Uravnenija **1**, 79 (1965).
[17] H. Airhault, Stud. Appl. Math. **61**, 31 (1979).
[18] J. D. Gibbon, P. Radmore, M. Tabor, and D. Wood, Stud. Appl. Math. **72**, 39 (1985).
[19] E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1944).
[20] E. Gambier, Acta Math. **33**, 1 (1910).
[21] J. Weiss, Phys. Lett. A **102**, 329 (1984).
[22] J. Weiss, Phys. Lett. A **105**, 387 (1984).

Canonical transformations theory for presymplectic systems

J. F. Cariñena
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, 50.009 Zaragoza, Spain
J. Gomis
Departamento de Física Teórica, Universitat de Barcelona, 08028 Barcelona, Spain
L. A. Ibort
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, 50.009 Zaragoza, Spain
N. Román
Departament de Física Teórica, Universitat Autonoma de Barcelona, Bellaterra (Barcelona), Spain

(Received 5 November 1984; accepted for publication 29 March 1985)

We develop a theory of canonical transformations for presymplectic systems, reducing this concept to that of canonical transformations for regular coisotropic canonical systems. In this way we can also link these with the usual canonical transformations for the symplectic reduced phase space. Furthermore, the concept of a generating function arises in a natural way as well as that of gauge group.

I. INTRODUCTION

Since the well-known Dirac's pioneer work[1] on constrained Hamiltonian systems, the interest in such theory has been growing because it provides an appropriate framework to deal with many physical theories either for finite-dimensional systems as time-dependent [or more generally (n-parameter)-dependent] systems, mechanical systems defined by singular Lagrangians, etc., or infinite-dimensional systems exhibiting gauge invariance. A good test of the relevance of these systems is the amount of papers trying to develop the mathematical framework for these systems, which has been shown to be that of presymplectic geometry, which has been possible thanks to the papers by Gotay,[2] Lichnerowicz,[3] Sniatycki,[4] and others, to whom we apologize for omitting their names. For a recent review see, e.g., Ref. 5.

The essential characteristic of these systems is the existence of contraint functions, limiting the possible values of the dynamical variables that Dirac classified in first and second class according to the vanishing or not of their mutual Poisson brackets. This classification was motivated because the second-class constraints may be eliminated from the theory up to a redefinition of Poisson brackets becoming now the so-called Dirac's brackets, and they may be considered as corresponding to spurious degrees of freedom. This aspect is really clarified when using appropriate coordinates as indicated by Shanmugadhasan.[6]

On the other hand, the invariance of the Poincaré–Cartan integral has also been proved to be a sound principle for the study of nondegenerate systems and it has motivated a recent paper[7,8] devoted to the study of the Hamilton–Jacobi method for degenerate systems. Our experience with regular systems suggests for us to look for a concept generalizing that of canonical transformation, and it has been carried out[9] for regular canonical systems by making use of a generalization of the Hwa-Chung theorem.[10] We aim in this paper to give a general concept of canonical transformation for any presymplectic system, as well as attempt to go deep in the analysis of this concept in order to characterize such transformations, studying the group structure of such a set of canonical transformations, some remarkable subgroups (in particular, the subgroup of gauge transformations), and the theory of the corresponding generating functions, which follows the track of Weinstein's theory for symplectic systems.[11]

The paper is organized as follows: Section II is devoted to analyzing the structure of locally Hamiltonian presymplectic systems, and the main result of this section, given in Theorem 3, is that the study of the locally Hamiltonian presymplectic systems can be done by means of its local structure coisotropic germ. The concept of canonical transformation for presymplectic systems is given in Sec. III and after a deep analysis it is shown that it is enough to consider the case of canonical regular systems because any other can be reduced to it. Section IV contains a study of the group structure of the set of canonical transformations. When the process of reduction of the presymplectic system is carried out, the canonical transformations pass to the quotient and it singularizes the subgroup of canonical transformations, inducing the identity in the quotient, called the gauge group. The concept of a generating function is introduced in Sec. V and Sec. VI is devoted to showing some interesting properties of the generating functions, which will be of interest to manage with in local coordinates.

II. THE STRUCTURE OF PRESYMPLECTIC SYSTEMS

The mathematical framework for a geometrical description of the Dirac's theory of constrained systems[1,12,13] has been shown to be that of presymplectic dynamical systems.[2–4,14,15] In this section we will analyze the local structure of such systems and it will be shown how it is possible to imbed a presymplectic manifold as a coisotropic submanifold of a symplectic manifold in which a family of locally Hamiltonian vector fields extending the dynamics of the original system can be constructed. This result is based on some theorems by Sniatycki[4] and Gotay,[16] which will be restated in order to make this paper more self-contained.

Definition 1: A presymplectic manifold is a pair (M,ω) where ω is a closed two-form of constant rank on the differentiable manifold M. If α is a closed one-form on M, the

 0022-2488/85/081961-09$02.50

triplet (M,ω,α) is said to be a locally Hamiltonian presymplectic dynamical system.

The Dirac–Bergmann theory of constrained systems corresponds to taking $M = D_{\mathscr{L}}(T^*Q)$ and ω, the pullback to M, of the canonical two-form ω_0 on T^*Q. Here $D_{\mathscr{L}}$ denotes the Legendre map $D_{\mathscr{L}}: TQ \to T^*Q$, with Q the configuration space and $\mathscr{L}$ the Lagrangian function which is assumed to be singular; that is, $D_{\mathscr{L}}$ is not a local diffeomorphism. Alternatively, we can consider in this case another presymplectic manifold $(M = TQ, \omega_{\mathscr{L}} = D^*_{\mathscr{L}}\omega_0)$.

There are a lot of other relevant presymplectic manifolds arising in physics. For instance, we can mention parameter-dependent systems where the mainfold M is $P \times \Lambda$ with (P,Ω) a symplectic manifold and Λ the parameter space. The closed two-form $\tilde{\omega}$ is given by $\tilde{\omega} = \pi^*\omega$, where π denotes the natural projection on the first factor $\pi: P \times \Lambda \to P$. This is the case of the usual way of dealing with time-dependent systems.[17]

Given a locally Hamiltonian presymplectic dynamical system, the constraint algorithm, developed by Gotay *et al.*,[2,14–16,18] provides a method for obtaining a maximal submanifold C, called the final constraint submanifold, for which the equation

$$\iota(\Gamma)\omega_{|C} = \alpha_{|C} \tag{2.1}$$

is meaningful, and it is possible to endow α with a dynamical sense. The vector field Γ is not uniquely defined and this ambiguity corresponds to what is usually called gauge freedom.[2,15]

When dealing with Dirac's constrained Hamiltonian systems, the functions locally defining this final constraint submanifold are both the primary constraints defining $M \subset T^*Q$ and the secondary constraints. But Dirac gave a new classification of constraints in first and second class depending on the possibility of eliminating the ambiguity in the corresponding multiplier in the expression of the total Hamiltonian. In the general case constraints of both classes can appear, but Sniatycki proved[4] that it is possible to imbed coisotropically the final constraint submanifold C in a symplectic manifold, in the very general case of C defining a regular canonical system, and the second-class constraints are eliminated.

Definition 2: Let (P,Ω) be a symplectic manifold and j: $C \hookrightarrow P$ a submanifold of P. Then (P,C,Ω) is said to be a regular canonical system if $\ker j^*\Omega \cap TC$ is a subbundle of the tangent bundle TC.

Theorem 1[4]: Let (P,C,Ω) be a regular canonical system. If C is a closed submanifold of P, there exist a symplectic submanifold of (P,Ω), $k: (\tilde{P},\tilde{\Omega}) \hookrightarrow (P,\Omega)$, and a coisotropic imbedding of C into $(\tilde{P},\tilde{\Omega})$, $l: (C, j^*\Omega) \hookrightarrow (\tilde{P},\tilde{\Omega})$, such that $k \circ l = j$.

The existence of a symplectic manifold (P,Ω) containing C may be forgotten for the presymplectic case if we make use of the coisotropic imbedding theorem recently given by Gotay.[16]

Theorem 2: Let (M,ω) be a presymplectic manifold. Then, we have the following.

(i) There exists a symplectic form Ω on a tubular neighborhood of the zero section of the dual bundle E^* of the characteristic bundle E of (M,ω), where M can be coisotropically imbedded.

(ii) Any two coisotropic imbeddings of (M,ω) are locally equivalent: if j_i $(M,\omega) \to (P_i,\Omega_i)$, $i = 1,2$, are two of such coisotropic imbeddings, there exist two neighborhoods $U_i = 1,2$, of $j_i(M)$ in P_i and a symplectomorphism ϕ: $U_1 \to U_2$ such that $\phi^*\Omega_2 = \Omega_1$ and $\phi \circ j_1 = j_2$.

We introduce next some definitions and notations we are going to use concerning functions and one-forms defined on a symplectic manifold (P,Ω).

Definition 3: Let C be a submanifold of (P,Ω). A function $f \in \mathscr{C}^\infty(P)$ is said to be a constraint function for C if $f_{|C}$ is constant, and the set of such functions will be denoted $C(P,C)$. A function $g \in \mathscr{C}(P,C)$ is called a first-class function if $\{f,g\}_{|C=0}$ $\forall f \in \mathscr{C}(P,C)$, and we will write $\mathscr{B}(P,C)$ for the set of all first-class functions. Finally, the first-class constraint functions are those of $\mathscr{B}(P,C) \cap \mathscr{C}(P,C)$, and the corresponding set will be denoted $\mathscr{A}(P,C)$.

Here $\{\ ,\ \}$ will denote the Poisson bracket defined on the set $\Lambda^1(P)$ of one-forms by the form Ω as follows: $\{\alpha,\beta\} = \hat{\Omega}[\hat{\Omega}^{-1}(\alpha),\hat{\Omega}^{-1}(\beta)]$ for any pair of one-forms $\alpha,\beta \in \Lambda^1(P)$. The map $\hat{\Omega}: \mathscr{X}(P) \to \Lambda^1(P)$ is defined by contraction with $\Omega, \hat{\Omega}(X) = \iota(X)\Omega$. When f and g are functions, the Poisson bracket is defined by $\{f,g\} = \Omega(\hat{\Omega}^{-1}(df),\hat{\Omega}^{-1}(dg))$. The concepts of constraint and first-class functions can be generalized for one-forms on P as follows.

Definition 4: A one-form $\alpha \in \Lambda^1(P)$ is a constraint one-form for C if $j^*\alpha = 0$, j being the immersion j: $C \hookrightarrow P$. The set of these one-forms will be denoted $C^1(P,C)$. The one-form β is a first-class one-form if $j^*\{\alpha,\beta\} = 0$, $\forall \alpha \in C^1(P,C)$, and the set of all such one-forms will be written $B^1(P,C)$. Finally, by $A^1(P,C)$ we will denote the set $A^1(P,C) = B^1(P,C) \cap C^1(P,C)$ of the first-class constraint one-forms.

Proposition 1: With the above notations, we have the following.

(i) $d\mathscr{A}(P,C) \subset A^1(P,C) \cap Z^1(P)$,

$d\mathscr{B}(P,C) \subset B^1(P,C) \cap Z^1(P)$,

$d\mathscr{C}(P,C) \subset C^1(P,C) \cap Z^1(P)$.

(ii) If $(C, j^*\Omega) \hookrightarrow (P,\Omega)$ is a coisotropic imbedding, then $C^1(P,C) \cap Z^1(P) \subset B^1(P,C) \cap Z^1(P)$ and therefore $A^1(P,C) \cap Z^1(P) = C^1(P,C) \cap Z^1(P)$.

Proof: (i) If $f \in \mathscr{C}(P,C)$ then $j^*df = d(j^*f) = 0$ and therefore $df \in C^1(P,C) \cap Z^1(P)$. Moreover, if $g \in \mathscr{B}(P,C)$, then $j^* d\{f,g\} = 0$ and therefore $j^*\{df,dg\} = 0$. But it implies that $j^*\{\alpha,dg\} = 0$ for any $\alpha \in C^1(P,C) \cap Z^1(P)$ because of the local existence of a neighborhood and a function $f \in \mathscr{C}(P,C)$ such that $\alpha = df$ according to the relative Poincaré lemma.[9]

(ii) If $\alpha \in C^1(P,C) \cap Z^1(P)$, the lemma of Poincaré shows that there is a function f (only locally defined) such that $df = \alpha$ and then $j^*\alpha = 0$ implies that j^*f is constant on the neighborhood $\mathscr{V}$ where f was defined. Now, if $\beta \in C^1(P,C) \cap Z^1(P)$ and g is a function (locally defined) such that $\beta = dg$, and $g \in \mathscr{C}(P,C)$, we see that $j^*\{\alpha,\beta\} = j^*\{df,dg\} = j^* d\{f,g\} = dj^*\{f,g\}$. If C is coisotropically imbedded in $P, \mathscr{C}(P,C) \subset \mathscr{B}(P,C)$ and therefore $j^*\{f,g\} = 0$, which implies $j^*\{\alpha,\beta\} = 0$. In order to prove that $j^*\{\alpha,\beta\} = 0$, $\forall \beta \in C^1(P,C)$, we remark that $C^1(P,C) \cap Z^1(P)$ generates locally $C^1(P,C)$ as a $C^\infty(P)$ module and for every $\beta \in C^1(P,C)$ there exist $b_i \in C^\infty(P)$ and $f^i \in \mathscr{C}(P,C)$ such that β can be written as $\beta = \Sigma b_i\, df^i$. Then, using the identity $\{\alpha,h\gamma\} = X_\alpha(h)\gamma + h\{\alpha,\gamma\}$,

$\forall h \in C^\infty(P)$, $\alpha,\gamma \in A^1(P)$, with $X_\alpha = \hat{\Omega}^{-1}(\alpha)$, we find that for every $\alpha \in C^1(P,C) \cap Z^1(P)$ and $\beta \in C^1(P,C)$,

$$\{\alpha,\beta\} = \left\{\alpha, \sum b_i\, df^i\right\} = \sum X_\alpha(b_i) df^i + \sum b_i \{\alpha, df^i\},$$

and therefore

$$j^*\{\alpha,\beta\} = \sum (X_\alpha(b_i) \circ j)\, d(j^* f^i) + \sum (b_i \circ j) j^* \{\alpha, df^i\} = 0.$$

The main goal of this section is the following theorem.

Theorem 3: Let (M,ω,α) be a locally Hamiltonian presymplectic system and $i: C \hookrightarrow M$ the final constraint submanifold. There exist a symplectic manifold (P,Ω) and a coisotropic imbedding $j: C \hookrightarrow P$ such that $j^*\Omega = i^*\omega$ and we have the following.

(i) For each vector field Γ on M, tangent to C, satisfying $i(\Gamma)\omega_{|C} = \alpha_{|C}$, there is a locally Hamiltonian vector field Γ_ξ on P, tangent to C, such that $\Gamma_{|C} = \Gamma_{\xi|C}$.

(ii) The vector fields Γ_ξ associated to the dynamical system Γ satisfying the above conditions are given by

$$\Gamma_\xi = \hat{\Omega}^{-1}(\alpha_P + \xi), \tag{2.2}$$

where α_P is a closed one-form on P such that $j^*\alpha_P = i^*\alpha$, and ξ any closed first-class constraint one-form on P for C, $\xi \in A^1(P,C) \cap Z^1(P)$.

(iii) (local uniqueness) The coisotropic imbedding and the family

$$D(P,C) = \{\hat{\Omega}^{-1}(\alpha_P + \xi) \mid \xi \in A^1(P,C) \cap Z^1(P)\}$$

are locally unique.

Here local uniqueness means that if $j': C \hookrightarrow P'$ is another coisotropic imbedding, there will exist a family of locally Hamiltonian vector fields

$$D(P',C) = \{\hat{\Omega}'^{-1}(\alpha_{P'} + \xi) \mid \xi \in A^1(P',C) \cap Z^1(P'),\ \alpha_{P'} \in Z^1(P')\},$$

and a local symplectomorphism ϕ from a neighborhood of $j(C)$ in P in a neighborhood of $j'(C)$ in P' such that $j' \circ \phi = \phi \circ j$ and maps $D(P,C)$ on $D(P',C)$.

Proof: According to Theorem 2, there is a symplectic manifold $(\tilde{P},\tilde{\Omega})$ and a coisotropic imbedding $l: (M,\omega) \hookrightarrow (\tilde{P},\tilde{\Omega})$. On the other hand, if $(C, i^*\omega)$ is the presymplectic manifold which is obtained from application of the constraint algorithm, Theorem 2 furnishes a new symplectic manifold where $(C, i^*\omega)$ is coisotropically imbedded. Let j_2 denote such an imbedding $j_2: C \to P_2$. The relation between both symplectic structures is given by Theorem 1. We can also see C as a submanifold $J: C \to P_1$ with $J = l \circ i$ and then Theorem 1 asserts the existence of a symplectic submanifold $k: (P_3,\Omega_3) \to (\tilde{P},\tilde{\Omega})$ and a coisotropic imbedding $j_3: (C, i^*\omega) \to (P_3, \Omega_3)$ such that $k \circ j_3 = J$. The local uniqueness part of Theorem 2 leads to the existence of a symplectomorphism ϕ of a neighborhood of $j_3(C)$ in P_3 on a neighborhood of $j_2(C)$ in P_2. If (P,Ω) is any of such neighborhoods and j the corresponding immersion of C in P, we have a coisotropic imbedding of C in (P,Ω). (See Fig. 1.)

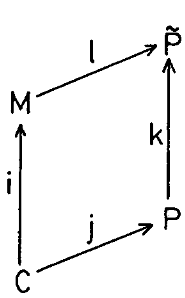


FIG. 1. Diagram of the coisotropic imbedding.

In order to prove the points concerning the dynamics, we remark that both $(\tilde{P},\tilde{\Omega})$ and (P,Ω) are neighborhoods of the zero sections of vector bundles over M and C, respectively. Let π_k, π_j, and π_l be the corresponding projections $\pi_k: \tilde{P} \to P$, $\pi_j: P \to C$, $\pi_l: \tilde{P} \to M$, verifying $\pi_k \circ k = \mathrm{id}_P$, $\pi_l \circ l = \mathrm{id}_M$, and $\pi_j \circ j = \mathrm{id}_C$. Let $\Gamma_\circ$ be a vector field in M tangent to C verifying the dynamical condition (2.1), i.e., $i(\Gamma_\circ)\omega_{|C} = \alpha_{|C}$. From the relation $k \circ j = l \circ i$ we see that the images of the manifold C under $l \circ i$ and $k \circ j$ are contained in $l(M)$ and $k(P)$, respectively, and then $C = l \circ i(C) \subset l(M) \cap k(P) \equiv W$. We define a vector field in $l(M)$ by $l_* \Gamma_\circ$ and take its restriction to W, that it is not necessarily tangent to W but it will be tangent to C because the tangency of $\Gamma_\circ$ to C implies that there exists a vector field $\Gamma'_\circ$ in C such that $i_* \Gamma'_\circ = \Gamma_\circ$ and therefore $l_* \Gamma_\circ = (l \circ i)_* \Gamma'_\circ$ is tangent to C. The map $\pi_k: k(P) \to P$ is a diffeomorphism, so that it is meaningful to take the restriction of π_k to $W = l(M) \cap k(P)$ and define the vector field $\pi_{k*}(l_* \Gamma_\circ)$ (the respective restrictions of l_* and π_k to W are understood). We remark that $\pi_{k*} l_* \Gamma_\circ$ is tangent to C in P because if we take the vector field $\Gamma'_\circ$ in C as above and compute $\pi_{k*} l_* \Gamma_\circ$ we find that $\pi_{k*} l_* \Gamma_\circ = j_* \Gamma'_\circ$. It is now easy to see that the vector field $\pi_{k*} l_* \Gamma_\circ$ defined in a submanifold of P satisfies on it the equation $i(\pi_{k*} l_* \Gamma_\circ)\Omega = \alpha_P$ with $\alpha_P = k^* \pi_l^* \alpha$. In fact, the following computation shows that we can associate the vector field $X_{\alpha_P} = \hat{\Omega}^{-1}(\alpha_P)$ with $\Gamma_\circ$ because

$$\begin{aligned} i(\pi_{k*} l_* \Gamma_\circ)\Omega(Y) &= k^* \Omega_1(\pi_{k*} l_* \Gamma_\circ, Y) \\ &= (\pi_l \circ k)^* \omega((\pi_k \circ l)_* \Gamma_\circ, Y) \\ &= \omega(\Gamma_\circ, (\pi_l \circ k)_* Y) \\ &= i(\Gamma_\circ)\omega(\pi_l \circ k)_* Y = \alpha_P(Y). \end{aligned}$$

The vector field $X_\xi = \hat{\Omega}^{-1}(\xi)$ corresponding to an element of $A^1(P,C) \cap Z^1(P)$ is tangent to C and is such that $X_{\xi|C} \in \ker i^*\omega$, and consequently the vector field $X_{\alpha_P} + X_\xi$ is a solution of the dynamical equation, too. Therefore, by addition of vector fields X_ξ with $\xi \in A^1(P,C) \cap Z^1(P)$ to the vector field X_{α_P} we obtain vector fields in P tangent to C. Noteworthy is that if Γ_1 is another vector field in M satisfying the dynamical equation, then the difference $(\Gamma_1 - \Gamma_\circ)_{|C}$ lies in $\ker i^*\omega$ and therefore $\pi_{k*} j_{1*} \Gamma_1 = \pi_{k*} j_{1*} \Gamma_\circ + X_\xi$, with $\xi \in A^1(P,C) \cap Z^1(P)$. Actually $\ker i^*\omega = \pi_{j*} \hat{\Omega}^{-1}(A^1(P,C) Z^1(P))$, because the closed first-class constraint one-forms generate via $\hat{\Omega}^{-1}$ the submodule $\Gamma(TC^\perp)$ of $\mathfrak{X}(P,C) = \{X \in \mathfrak{X}(P) \mid X_{|C} \in \Gamma(TC)\}$.

As far as the local uniqueness is concerned we must prove that given two coisotropic imbeddings j_1, j_2 into two symplectic manifolds (P_i, Ω_i), $i = 1,2$, there will be a local symplectomorphism $\phi: P_1 \to P_2$ mapping locally Hamiltonian vector fields on P_1 in locally Hamiltonian vector fields on P_2 and $\phi \circ j_1 = j_2$. Now, let (P,Ω) be a symplectic manifold where the final constraint manifold is coisotropically imbedded, obtained from any symplectic manifold (S,σ) in which (M,ω) is imbedded.

The local uniqueness part of the statement of Theorem 2 says that any two coisotropic imbeddings are locally equivalent and consequently the two symplectic manifolds we will obtain, either from (M,ω) using (S,σ) or from $(C,i^*\omega)$ using the coisotropic imbedding theorem, have to be locally symplectomorphic. The second assertion of the statement follows from the fact that any symplectomorphism preserves the locally Hamiltonian character of the vector fields, and from the condition $\phi \circ j_1 = j_2$, which says that ϕ transforms constraint one-forms into constraint one-forms; and, as we have shown that $C^1(P,C)\cap Z^1(P,C) = A^1(P,C)\cap Z^1(P,C)$, the proof ends.

Definition 5: If (P,C,Ω) is a regular canonical system such that the immersion j is a coisotropic imbedding we will say that (P,C,Ω) is a regular canonical coisotropic system.

The preceding results can also be presented in a different language using the concept of local manifold pair, as Weinstein does,[19] or that of a germ of a manifold as a submanifold of another one; that is, if C is a submanifold of M and (M,C) a pair of manifolds, we will say that (M',C) is equivalent to (M,C) if there is another pair (M'',C) such that M'' is an open submanifold of both M and M'. An equivalence class of pairs of manifolds is called a local manifold pair or germ of C in M and will be denoted $[M,C]$. A map between two germs is defined by an equivalence class of maps. This equivalence is defined as follows: two maps $f_i:(M_i,C)\rightarrowtail(M_i',C')$, $i = 1,2$ are said to be equivalent if there exists a map $g:(M_3,C)\rightarrowtail(M_3',C')$ such that M_3 is an open submanifold of both M_1 and M_2, and M_3' is an open submanifold of M_1' and M_2' with $f_{1|M_3} = f_{2|M_3} = g$.

A germ $[P,C]$ is said to be coisotropic if (P,C) is a pair where C is a coisotropic submanifold of the symplectic manifold (P,Ω). We can consider the category with objects the germs $[P,C]$ and morphisms the symplectic maps between germs $[\phi]:[P,C]\rightarrowtail[P',C']$. We will say that a germ $[P,C]$ is the local structure germ for a presymplectic germ if it verifies the universal property of being an initial object in this category, i.e., for every $[P',C]$ there is a morphism $[\phi]:[P,C]\rightarrowtail[P',C']$ such that $\phi_{|C} = id_C$. With this language Theorem 3 can be restated as follows: For every locally Hamiltonian presymplectic system (M,ω,α), there exists a local structure germ $[P,C]$, with C the final constraint manifold for (M,ω,α). It is uniquely defined and there is on it a family of locally Hamiltonian vector fields furnishing a dynamical description of the system.

III. CANONICAL TRANSFORMATIONS FOR PRESYMPLECTIC SYSTEMS

The traditional concept of canonical transformations for Hamiltonian dynamical systems as symplectomorphisms has recently been generalized[9] for application to regular canonical systems (P,S,Ω). The definition of canonical transformation depends on the choice of a particular kind of vector field, called locally weakly Hamiltonian fields relative to (P,S,Ω), and therefore depends on the immersion of S in the ambient manifold P. We aim to find a generalization of the concept of canonical transformation for a presymplectic system with no reference to an ambient symplectic manifold containing it, that it will reduce to that proposed in Ref. 9 in the case of a regular canonical system. Moreover, we will prove, by making use of the results of the preceding section, that the general problem of studying the canonical transformations of a presymplectic system can be reduced to that of the canonical transformations of a regular canonical system (P,C,Ω).

Definition 6: Let (M,ω,α) be a locally Hamiltonian presymplectic system and let $i_C:C\hookrightarrow M$ be the final constraint submanifold. A vector field $X\in\mathfrak{X}(M)$ is said to be a locally Hamiltonian vector field relative to C if (i) X is tangent to C, $X_{|C}\in\Gamma(TC)$ and (ii) there exists a closed one-form $\beta\in Z^1(M)$ such that

$$i_C^*(\iota(X)\omega - \beta) = 0. \tag{3.1}$$

The set of such vector fields will be denoted $\mathfrak{X}_{\rm LH}(M,C)$. It is to be remarked that the condition (ii) is weaker than $\iota(X)\omega_{|C} = \beta_{|C}$ and any vector field X satisfying this equation will satisfy (3.1), too. As an example, the dynamical vector fields provided by the Constraint algorithm are locally Hamiltonian vector fields relative to C. On the other hand condition (ii) is equivalent to $i_C^* L_X\omega = 0$.

As a corollary of the theorems of Ref. 9 we can write down the generalization of the Hwa-Chung theorem for presymplectic systems.

Theorem 4: Let (M,ω,α) be a locally Hamiltonian presymplectic system with final constraint submanifold $i_C:C\hookrightarrow M$ and rank $(i_C^*\omega) = 2r$. If $\beta\in\Lambda^p(M)$ is such that $i_C^*(L_X\beta) = 0$, $\forall X\in\mathfrak{X}_{\rm LH}(M,C)$, then we have the following.

(i) $i_C^*\beta = 0$ if $p > 2r$ or $p = 2l + 1$ with $l < r$.

(ii) If $p = 2l$, $l<r$, there exists a function $f\in C^\infty(M)$ such that $i_C^*(\beta - f\omega^{\wedge l}) = 0$ and $i_C^* f$ is constant on each connected component of C.

In this context the concept of canonical transformation generalizing that of Ref. 9 is the following one.

Definition 7: Let (M_k,ω_k,α_k), $k = 1,2$, be two locally Hamiltonian presymplectic systems and $i_k:C_k\rightarrow M_k$ the corresponding final constraint submanifolds. A pair (Φ,ϕ) of diffeomorphisms $\Phi:M_1\rightarrow M_2$ and $\phi:C_1\rightarrow C_2$ is said to be a canonical transformation between (M_1,ω_1,α_1) and (M_2,ω_2,α_2) if (i) $\Phi\circ i_1 = i_2\circ\phi$ and (ii) $\Phi_*(\mathfrak{X}_{\rm LH}(M_1,C_1))\subset\mathfrak{X}_{\rm LH}(M_2,C_2)$.

A characterization of a canonical transformation for such systems, which is a straightforward consequence of the former theorem, is given by the following.

Theorem 5: A pair (Φ,ϕ) of diffeomorphisms $\Phi:M_1\rightarrow M_2$ and $\phi:C_1\rightarrow C_2$, such that $\Phi\circ i_1 = i_2\circ\phi$, is a canonical transformation if and only if there is a real number c such that $i_1^*(\Phi^*\omega_2 - c\omega_1) = 0$.

Only the particular case $c = 1$ will be considered in the following. It corresponds to the restricted canonical transformations for Hamiltonian systems in the terminology of the book by Saletan and Cromer,[20] but we will omit the word restricted.

A convenient characterization of the locally Hamiltonian vector fields which is also an immediate consequence of Theorem 4 is given next.

Theorem 6: Let (M,ω,α) be a locally Hamiltonian presymplectic system and $i_C:C\rightarrow M$ the final constraint submanifold. A vector field X in M tangent to C is locally Hamiltonian relative to C if and only if the flow of X is a family of canonical transformations of (M,ω,α).

The fundamental result of this section concerns the reduction for a general presymplectic system to the case of a canonical system which is given by the structure theorem of the precedent section. In fact, the next theorem asserts that the set of canonical transformations of a presymplectic system can be seen as the set of canonical transformations of a regular canonical system coisotropically imbedded.

Theorem 7: With the same notations as in Definition 7, for each canonical transformation (Φ,ϕ) between (M_k,ω_k,α_k), $k=1,2$, if (P_k,C_k,Ω_k) are their corresponding regular canonical coisotropic systems given by Theorem 2, there exists a symplectomorphism Ψ between them such that $\Psi\circ j_1=j_2\circ\phi$, with j_k being the injections $j_k:C_k\to P_k$.

Proof: Let $l_2:M_2\hookrightarrow(\tilde P_2,\tilde\Omega_2)$ be the coisotropic imbedding to M_2 in $(\tilde P_2,\tilde\Omega_2)$, $k_2:(P_2,\Omega_2)\to(\tilde P_2,\tilde\Omega_2)$ the symplectic submanifold, and $j_2:C_2\hookrightarrow(P_2,\Omega_2)$ the coisotropic imbedding given by Gotay's and Sniatycki's theorems verifying $k_2\circ j_2=l_2\circ i_2$ as in Theorem 3. (see Fig. 2.)

Now, the point is that the composite map $j_2\circ\phi:C_1\hookrightarrow(P_2,\Omega_2)$ is a coisotropic imbedding satisfying $(j_2\circ\phi)^*\Omega_2=i_1^*\omega_1$. In fact, a little computation gives $(j_2\circ\phi)^*\Omega_2=\phi^*i_2^*\omega_2=(\Phi\circ i_1)^*\omega_2=i_1^*\omega_1$. In order to prove that $j_2\circ\phi$ is coisotropic, we must show that $TC_1^{\perp\Omega_2}\subset\mathbf{TC}_1$, where $\mathbf{TC}_1$ denotes the set of tangent vectors to C_1 through $j_2\circ\phi$, that is, $\mathbf{TC}_1=(j_2\circ\phi)_*(TC_1)$. Let $u\in TC_1^{\perp\Omega_2}|_p$, where $p=j_2\circ\phi(m_1)$, i.e., $\Omega_2(p)(u,v)=0$, $\forall v\in\mathbf{TC}_1|_p$. If $v\in\mathbf{TC}_1|_p$, there exists a tangent vector $v'\in TC_1|_{m_1}$ such that $(j_2\circ\phi)_*(m_1)v'=v$, so that $\Omega_2(m_2)(u,j_{2*}\phi_*(m_1)v')=0$, where $m_2=\phi(m_1)$, $\forall v'\in T_{m_1}C_1$, or in the same way $\Omega_2(p)(u,j_{2*}(m_2)v'')=0$, $\forall v''\in T_{m_2}C_2$, because ϕ is a diffeomorphism. Then, $u\in TC_2^{\perp\Omega_2}$, and from the coisotropy of C_2 we have that $u\in\mathbf{TC}_2$; but $\mathbf{TC}_2=\mathbf{TC}_1$ and $j_2\circ\phi:C_1\hookrightarrow(P_2,\Omega_2)$ is a coisotropic imbedding.

In this point the local uniqueness of Theorem 2 shows that there exists a symplectomorphism Ψ from (P_1,Ω_1), the initial symplectic manifold where C_1 is coisotropically imbedded, into (P_2,Ω_2), such that $\Psi\circ j_1=j_2\circ\phi$, and the proof ends.

Remarks: (i) The function Ψ is defined only locally on a neighborhood of $j_1(C)$ in P, but taking this neighborhood as the whole manifold the result still holds.

(ii) This theorem shows the possibility of studying canonical transformations for presymplectic systems using only their local structures as in Theorem 3. This simplification permits development of the study of the group of canonical transformations and its subgroups, so in the following sections we will use both points of view to deal with canonical transformations for a presymplectic system. That is, given a canonical transformation (Φ,ϕ) between (M_1,ω_1,α_1) and (M_2,ω_2,α_2), we use without mention of it the associated canonical transformation (Ψ,ϕ) between the associated coisotropic regular canonical systems (P_1,C_1,Ω_1) and (P_2,C_2,Ω_2).

(iii) It is also to be remarked that there are canonical transformations between canonical regular systems that are not symplectomorphisms. In fact, it is possible to consider canonical transformations between two canonical regular systems associated to presymplectic systems (M_k,ω_k,α_k) that are not symplectic transformations.

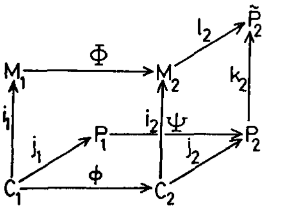


FIG. 2. Diagram displaying the maps of Theorem 7.

IV. THE GROUP OF CANONICAL TRANSFORMATIONS FOR PRESYMPLECTIC SYSTEMS

Instead of dealing with presymplectic systems as indicated in the preceding sections, there is an alternative way which is called the reduction of the phase space.[17,19] The kernel of the presymplectic form $\omega_C=i_C^*\omega$ defines an involutive distribution and therefore it is integrable because of the well-known Frobenius theorem. The maximal connected integral submanifolds are the leaves of a foliation that gives rise to an equivalence relation in C. Suppose we discard the points of C, where ω_C fails to be of constant rank, and denote $\pi_C:C\to\hat C$ the natural projection of C onto the quotient space. Then, if π_C is a submersion, there exists a symplectic form $\hat\omega$ defined on $\hat C$ such that $\pi_C^*\hat\omega=\omega_C=i_C^*\omega$. It is defined by means of $\hat\omega_{\hat m}(\hat X,\hat Y)=\omega_m(X,Y)$, where $m\in C,\pi_C(m)=\hat m$ and $X,Y\in T_mC$, $\hat X,\hat Y\in T_{\hat m}\hat C$ are related by $\pi_{C_*m}(X)=\hat X,\pi_{C_*m}(Y)=\hat Y$. The pair $(\hat C,\hat\omega)$ is called the reduced phase space.

This is the usual approach to the study of dynamical systems with gauge degrees of freedom, as Yang–Mills fields[21,23] and gravitational fields.[23] In this scheme the canonical transformations are but symplectomorphisms of the reduced structure. In this section both alternative definitions will be related; we will prove that there is a canonical epimorphism of the group of generalized canonical transformations we have defined onto the group $\mathrm{Sp}(\hat C,\hat\omega)$ of symplectomorphisms of $(\hat C,\hat\omega)$.

In order to explain this deep relation we need some notations referring to the group of (generalized) canonical transformations and its more relevant subgroups, which we present next.

We will denote $GC(P,C)$ the set of canonical transformations for the coisotropic canonical system (P,C,Ω) which can be endowed with a group structure in the natural way.

There are a lot of important subgroups of this with physical and mathematical meaning. For instance, $GS(P,C)=GC(P,C)\cap\mathrm{Sp}(P)$, which is not a normal subgroup in the general case. Now, if $\pi:P\to C$ denotes the above-mentioned projection, a very important subgroup of $GS(P,C)$ is made up by the elements that commute with π and leave invariant the symplectic form Ω. The set of such fibered symplectomorphisms is a subgroup to be denoted $FS(P,C)$, and it has been studied for time-dependent systems in Ref. 24. In a similar way we can define $FG(P,C)$ as made up from all fibered canonical transformations.

We will denote $TC(P,C)$ the set of canonical transformations that are trivial on C. This set is a normal subgroup of $GC(P,C)$ and has a subgroup to $TS(P,C)=TC(P,C)\cap GS(P,C)$.

The lattice of these subgroups as well as the relationship between them are shown in the diagram below. The symbol $\vdash\!\!\dashv$ means that the lower is normal in the upper one, and a subgroup in the link of two means that it is the intersection of both groups on the opposite edges

$$\begin{tikzcd}[column sep=small]
 & GC(PC) \arrow[dr, dash] & & SP(P) & \\
TC(P,C) \arrow[ur, dash] \arrow[dr, dash] & & GS(P,C) \arrow[ur, dash] \arrow[dl, dash] \arrow[dr, dash] & & FG(P,C) \\
 & TS(P,C) \arrow[dr, dash] & & FS(P,C) \arrow[ur, dash] & \\
 & & FTS(P,C) \arrow[ur, dash] & &
\end{tikzcd}$$

The group of the equivalence classes, $GC(P,C)/TC(P,C)$, will be denoted Can C and it is obvious that each class $[(\Phi,\phi)] \in \text{Can } C$ is uniquely defined by $\phi \in \text{Diff } C$, hence Can C is isomorphic to the group of those diffeomorphisms of C preserving the presymplectic structure $\Omega_C = j^*\Omega$. Another related matter is to know whether it is possible to choose a symplectic transformation of (P,Ω) in any class or not. All this and related questions will be dealt with in next section.

The main theorem in this section is based on the following proposition.

Proposition 2: Any canonical transformation (Φ,ϕ) for (P,C,Ω) leaves invariant the distribution defined by Ω_C.

Proof: If $v \in T_m C$ is in $\ker \Omega_C(m)$, X is a vector field defined in a neighborhood of m in P such that $X_m = v$ and $X_{|C} \in \Gamma(\ker \Omega_C)$, and we take into account that C is coisotropic in P, we can conclude that $\Gamma(\ker \Omega_C) = \Gamma(TC^\perp)$ and consequently $\Gamma(\ker \Omega_C)$ is generated by constraint first-class functions; namely, $f \in \mathscr{A}(P,C)$ will exist such that $X_f = \hat{\Omega}^{-1}(df) = X$. A canonical transformation maps the set of locally Hamiltonian vector fields tangent to C onto itself and the subset of those corresponding to constraint first-class functions on itself and therefore $\phi_* \ker \Omega_C = \ker \Omega_C$.

Theorem 8: With the same notations as above, the map $\hat{\phi}:\hat{C} \to \hat{C}$, defined by $\hat{\phi} \circ \pi_C = \pi_C \circ \phi$, is a symplectic map in $(\hat{C},\hat{\Omega})$.

Proof: The map is well defined because the foliation defined by $\ker \Omega_C$ is invariant under ϕ. Moreover, if we compute $\pi_C^* \hat{\phi}^* \hat{\Omega}$ we find the chain of identities $\pi_C^* \hat{\phi}^* \hat{\Omega} = (\phi \circ \pi_C)^* \hat{\Omega} = \phi^* \Omega_C = \Omega_C = \pi^* \hat{\Omega}$. Now, π being a submersion, we can conclude that $\phi^* \hat{\Omega} = \hat{\Omega}$.

Corollary 1: There is a canonical homomorphism p between Can C and $\text{Sp}(\hat{C},\hat{\Omega})$ given by $p(\phi) = \hat{\phi}$.

Definition 8: The kernel of the homomorphism p will be called the group $\mathscr{G}$ of (P,C,Ω) and is made up by the canonical transformations preserving every leaf of the foliation defined by $\ker \Omega_C$.

If $A(P,C)$ is the set of Hamiltonian constraint first-class vector fields in P over C, according to Gotay's notation,[2] namely, $A(P,C) = \{X_f = \hat{\Omega}^{-1}(df) | f \in \mathscr{A}(P,C)\}$, we can write an exact sequence of Lie algebras as indicated by the following theorem.

Theorem 9: With the above notations, the sequence

$$0 \to A(P,C) \xrightarrow{i_*} \mathfrak{X}_{\text{LH}}(P,C) \xrightarrow{\pi_{C*}} \mathfrak{X}_{\text{LH}}(\hat{C}) \to 0$$

is exact. Here i_* is the natural injection of $A(P,C)$ in $\mathfrak{X}(C)$.

Proof: Notice that the vector fields in $\mathfrak{X}_{\text{LH}}(C)$ are π_C projectable and therefore π_{C*} is well defined. The Hamiltonian constraint first-class vector fields generate $\ker \Omega_C = \Gamma(TC^\perp)$ and they are mapped by π_{C*} on the zero vector field. Conversely, if a vector field $X \in \mathfrak{X}_{\text{LH}}(C)$ is mapped by π_{C*} on the zero vector field, each integral curve is contained in a leaf of the foliation defined by $\ker \Omega_C$, so X is in $\Gamma(TC^\perp)$ and it belongs to $A(P,C)$.

Corresponding to this exact Lie algebra sequence we have another sequence of Lie groups

$$1 \to \mathscr{G} \to \text{Can } C \to \text{Sp}(\hat{C},\hat{\Omega}) \to 1.$$

It is noteworthy that in the case of Yang–Mills fields, the gauge group $\mathscr{G}$ is a Lie Hilbert group and $A(P,C)$ is actually the Lie algebra of this group.[21,25]

V. GENERATING FUNCTIONS FOR GENERALIZED CANONICAL TRANSFORMATIONS

The generating functions for canonical transformations of Hamiltonian systems arise as associated to the Lagrangian manifolds corresponding to the graph of the transformation in a symplectic product space.[11,17] If (P_1,Ω_1) and (P_2,Ω_2) are symplectic manifolds, a symplectic structure Ω_{12} is defined on the product manifold $P_1 \times P_2$ by $\Omega_{12} = \pi_1^* \Omega_1 - \pi_2^* \Omega_2$, where $\pi_i : P_1 \times P_2 \to P_i$ $(i = 1,2)$ are the canonical projections. Then $\phi:(P_1,\Omega_1) \to (P_2 \circ \Omega_2)$ is a symplectomorphism if and only if its graph is a Lagrangian submanifold of $(P_1 \times P_2, \Omega_{12})$.[11,17,26] Before trying to generalize the concept of generating function we establish a similar property characterizing the canonical transformations for presymplectic systems.

Theorem 10: Let (P_i,S_i,Ω_i), $i = 1,2$, be two canonical regular systems. A pair of diffeomorphisms (Φ,ϕ), $\Phi:P_1 \to P_2$, $\phi:S_1 \to S_2$ is a canonical transformation if and only if (i) $\Phi \circ j_1 = j_2 \circ \phi$, where $j_i:S_i \to P_i$ are the imbeddings of the submanifolds into the symplectic manifolds (P_i,Ω_i); and (ii) graph Φ is an isotropic submanifold of $(P_1 \times P_2, \Omega_{12})$.

Proof: Let $\bar{i}$ denote the canonical injection $\bar{i}$: graph $\phi \to S_1 \times S_2$ and i the canonical injection i:graph $\Phi \to P_1 \times P_2$. The map j:graph $\phi \to$ graph Φ defined by $j(x,\phi(x)) = (j_1(x), \Phi(j_1(x)))$, $\forall x \in S_1$ is such that $(j_1 \times j_2) \circ \bar{i} = i \circ j$. The map $i \circ j$:graph $\phi \to P_1 \times P_2$ is an imbedding and

$$(i \circ j)^* \Omega_{12} = j_1^*(\Omega_1 - \Phi^* \Omega_2).$$

Consequently $(i \circ j)^* \Omega_{12} = 0$ if and only if $j_1^* \times (\Omega_1 - \Phi^* \Omega_2) = 0$.

We recall that if $k:I \to P$ is an isotropic submanifold of the symplectic manifold (P,Ω), then $k^*\Omega = 0$, and if θ is a locally defined one-form such that $\Omega = d\theta$, the one-form $k^*\theta$ is closed and there will be a locally defined function S on I with $dS = k^*\theta$. Such a function S is called a generalized generating function for the isotropic submanifold I. The important point to be remarked is that the generating function for Lagrangian submanifolds describes the local structure of these,[11] whereas the generalized generating functions for isotropic submanifolds only partially describe such submanifolds. We can, however, define generalized generating functions for canonical transformations of presymplectic systems in a similar way as in the classical case of canonical transformations for Hamiltonian systems.

With the same notations as in Theorem 10, if $\mathscr{U}_1$ and $\mathscr{U}_2$ are two neighborhoods in P_1 and P_2, respectively, in which one-forms θ_i, $i=1,2$ are defined such that $d\theta_i = \Omega_i$, the one-form $\Theta_{12} = \pi_1^* \theta_1 - \pi_2^* \theta_2$ defined in $\mathscr{U}_{12} = \mathscr{U}_1 \times \mathscr{U}_2$ satisfies $\Omega_{12} = d\Theta_{12}$. The relation between θ_1 and $\Phi^* \theta_2$ for a canonical transformation of (P,S,Ω) is given by the following theorem.

Theorem 11: Let $\Phi: P_1 \to P_2$ be a map such that there exists $\phi: S_1 \to S_2$ with $\Phi \circ j_1 = j_2 \circ \phi$. Then, we have the following.

(i) (Φ,ϕ) is a canonical transformation if and only if there is a function G locally defined on graph ϕ such that $(i \circ j)^* \Theta_{12} = dG$.

(ii) (Φ,ϕ) is a canonical transformation if and only if there is a function F locally defined on S_1 such that $j_1^* \times (\theta_1 - \Phi^* \theta_2) = -dF$.

(iii) In the case of (Φ,ϕ) being a canonical transformation, there exist connected neighborhoods $\mathscr{V}$ in S_1 and $\mathscr{U}$ in graph ϕ such that $G \circ \rho - F$ is constant, where ρ is the inverse of the restriction of π_1 to $\mathscr{U}$.

Proof: (i) The submanifold graph Φ of $P_1 \times P_2$ is isotropic if and only if $0 = (i \circ j)^* \Omega_{12} = d(i \circ j)^* \Theta_{12}$ and therefore iff there exists a function G locally defined on a neighborhood $\mathscr{U}$ in graph ϕ for every point in graph ϕ with $(i \circ j)^* \Theta_{12|\mathscr{U}} = dG$.

(ii) The canonicity condition $j_1^*(\Phi^* \Omega_2 - \Omega_1) = 0$, when restricted to $\mathscr{U}_1 \times \mathscr{U}_2$, becomes the closedness of $j_1^* \times (\Phi^* \theta_2 - \theta_1)$ on a neighborhood $\mathscr{V}$ in S_1 such that $j_1(\mathscr{V}) \subset \mathscr{U}_1 \cap \Phi^{-1}(\mathscr{U}_2)$. It is equivalent to the local existence of a function F on $\mathscr{V}$ with

$$j_1^*(\Phi^* \theta_2 - \theta_1) = -dF.$$

(iii) Let ρ be the inverse map for the restriction of π_1 to graph ϕ.

Then, $\rho^* \, dG = j_1^*(\theta_1 - \Phi^* \theta_2) = dF$ and therefore $G \circ \rho - F$ is constant.

Definition 9: The functions F and G locally defined as above on S_1 and graph ϕ, respectively, will be called Poincaré and Weinstein generating functions for the canonical transformation (Φ,ϕ).

These functions are but generalizations of the corresponding concepts for Hamiltonian systems[17,24] and they admit extensions to open neighborhoods in P_1 and graph Φ, respectively, as shown in the following proposition.

Proposition 3: With the notations of Theorem 10, if $\{\xi^i\}_{i=1}^k$ is a set of independent functions defining $\mathscr{V}$ in P_1 and $\{\zeta^i\}_{i=1}^k$ is another defining $\phi(\mathscr{V})$ in P_2, then we have the following.

(i) The neighborhood $\mathscr{U} = \rho(\mathscr{V})$ in graph ϕ can be defined in graph Φ by the set $\{\eta^i\}_{i=1}^k$ of independent functions given by $\eta^i = \pi_1^* \xi^i + \pi_2^* \zeta^i$.

(ii) There exists a function $\tilde{G}$ defined on $j \circ \rho(\mathscr{V}) \subset \text{graph}\ \Phi$ such that $i^* \Theta_{12|_{\text{graph}\ \phi}} = d\tilde{G}_{|\text{graph}\ \phi}$ and $j^* \tilde{G} = G$.

(iii) There exists a function $\tilde{F}$ defined on a neighborhood $\tilde{\mathscr{V}}$ of P_1 such that $\tilde{\mathscr{V}} \cap C_1 = \mathscr{V}$, $j_1^* \tilde{F} = F$, $\tilde{G} \circ \rho - \tilde{F}$ is constant, and $(\Phi^* \theta_2 - \theta_1)|_{C_1} = dF_{|C_1}$.

Proof: (i) Let $(y, \Phi(y))$ be an element of graph Φ. Then, from the identity

$$\eta(y,\Phi(y)) = (\pi_1^* \xi^i + \pi_2^* \zeta^i)(y,\Phi(y)) = \xi^i(y) + \zeta^i(\Phi(y)),$$

it follows that $\eta^i(y,\Phi(y)) = 0$ $(i=1,\ldots,k)$ is equivalent to $y \in \mathscr{V}$ and $\Phi(y) \in \Phi(\mathscr{V})$.

(ii) Let G_e be an arbitrary but fixed extension of G to $j(\rho(\mathscr{V}))$. Every extension $\tilde{G}$ can be written as $\tilde{G} = G_e + \Sigma_{i=1}^k f_i \eta^i$ and therefore

$$j^* \tilde{G} = j^* G_e + \sum_{i=1}^k (f_i \circ j)(\eta_i \circ j) = j^* G_e = G.$$

Finally, since $j^*(i^* \Theta_{12}) = j^* d\tilde{G}$ and using Lagrange's multiplier theorem, we can conclude that $i^* \Theta_{12|\mathscr{U}} = d\tilde{G}_{|\mathscr{U}}$.

(iii) Let F_e be defined as $F_e = G_e \circ \rho$, and defining $\tilde{F}$ on a neighborhood $\tilde{\mathscr{V}}$ of S_1 in P_1 by means of $\tilde{F} = F_e + \Sigma_{i=1}^k (f_i \circ \rho) \xi^i$, such a function F is such that $j_1^* \tilde{F} = F$, and furthermore,

$$d\tilde{F} = dF_e + \sum_{i=1}^k d(f_i \circ \rho) \xi^i + \sum_{i=1}^k (f_i \circ \rho) d\xi^i = d(\tilde{G} \circ \rho),$$

where the functions ξ^i and η^i defining $\mathscr{V}$ and $\phi(\mathscr{V})$ have been assumed to be chosen as $\xi^i = \eta^i \circ \rho$. Finally, since $j_1^*(\theta_1 - \Phi^* \theta_2) = dF = j_1^* \, d\tilde{F}$ we obtain $(\theta_1 - \Phi^* \theta_2)_{|\mathscr{V}} = d\tilde{F}_{|\mathscr{V}}$.

Before ending this section we want to remark that even if $\tilde{F}$ and $\tilde{G}$ seem to play the same role as the classical Poincaré and Weinstein generating functions, they only define locally a symplectic transformation. The point is that in some cases they define a global symplectomorphism $\Phi: P_1 \to P_2$. This case was the one considered in Ref. 24 but it is not the general case in which we are only capable of relating the coordinates of the points in S_1 with those of S_2. The next section is devoted to explaining how to get the explicit form of $\phi: S_1 \to S_2$ from the generating function $\tilde{G}$ (or G) as well as to presenting some remarkable results concerning the generating functions F and G.

VI. LOCAL PROPERTIES OF GENERATING FUNCTIONS

In this section we will analyze the local reconstruction of a generalized canonical transformation $\phi \in \text{Can}\ C$ starting from its Weinstein generating function, as well as its relation with the corresponding generating function in the reduced phase space. Let (P_i, C_i, Ω_i), with $i=1,2$, be two coisotropic regular systems. Then, it is to be remarked that if (Φ,ϕ) is a generalized canonical transformation between (P_1, C_1, Ω_1) and (P_2, C_2, Ω_2), then graph ϕ is an isotropic submanifold of $(P_1 \times P_2, \Omega_{12})$ while $C_1 \times C_2$ is a coisotropic submanifold. The canonical projection of C_i on the corresponding reduced space will be denoted η_i instead of the more cumbersome notation π_{C_i}. The reduced phase space $\widehat{C_1 \times C_2}$ is but $\hat{C}_1 \times \hat{C}_2$ and the projection on the reduced phase space $\widehat{C_1 \times C_2}$ is denoted $\Pi: C_1 \times C_2 \to \widehat{C_1 \times C_2}$, which coincides with $\eta_1 \times \eta_2$. We also recall that if $\hat{\phi}$ is a symplectomorphism between $(\hat{C}_1, \hat{\Omega}_1)$ and $(\hat{C}_2, \hat{\Omega}_2)$, the set graph $\hat{\phi}$ is a Lagrangian submanifold of $(\hat{C}_1 \times \hat{C}_2, \hat{\Omega}_{12})$ with $\hat{\Omega}_{12}$ defined usually as $\hat{\Omega}_{12} = \hat{\pi}_1^* \hat{\Omega}_1 - \hat{\pi}_2^* \hat{\Omega}_2$, where $\hat{\pi}_i: \hat{C}_1 \times \hat{C}_2 \to \hat{C}_1$ $(i=1,2)$ is the canonical projection. With these notations, we can state the following proposition.

Proposition 4: Let G be a locally defined Weinstein function for the canonical transformation (Φ,ϕ). Then, there exists a Weinstein generating function $\hat{G}$ for the reduced symplectomorphism $\hat{\phi}: \hat{C}_1 \to \hat{C}_2$ such that $G = \Pi^* \hat{G}$.

Proof: If $\hat{\theta}_i$ and θ_i are locally defined one-forms such that $d\hat{\theta}_i = \hat{\Omega}_i$ and $d\theta_i = \Omega_i$, the identity $\eta_i^* \hat{\Omega}_i = j_i^* \Omega_i$ implies that there are locally defined functions f_i on neighborhoods of the C_i's with $\eta_i^* \hat{\theta}_i = j_i^* \theta_i + df_i$. If θ_i' is defined as $\theta_i' = \theta_i + d\pi_{ij}^* f_j$, where $\pi_{ij}: P_i \to C_j$ denotes the projection along the fiber structure of P_i over C_j, then $j_i^* \theta_i' = \eta_i^* \hat{\theta}_i$. Let G be the Weinstein generating function defined in Sec. V using the one-form $\Theta_{12}' = \pi_1^* \theta_1' - \pi_2^* \theta_2'$. The following relation holds locally: $\bar{i}^*(j_1 \times j_2)^* \Theta_{12}' = dG$. If $\hat{i}$ is the natural inclusion of graph $\hat{\phi}$ in $\hat{C}_1 \times \hat{C}_2$, we have $\hat{i} \circ \Pi = \Pi \circ \bar{i}$ and $\hat{\pi}_i \circ \Pi \circ \bar{i} = \eta_i \circ \pi_i$ $(i = 1,2)$. Consequently, the one-form $\hat{\Theta}_{12}$ defined by $\hat{\Theta}_{12} = \hat{\pi}_1^* \hat{\theta}_1 - \hat{\pi}_2^* \hat{\theta}_2$ defines a generating function $\hat{G}$ such that $\Pi^* d\hat{G} = dG$ because

$$\Pi^* d\hat{G} = \Pi^* \hat{i}^* \hat{\Theta}_{12} = (\eta_1 \circ \pi_1)^* \hat{\theta}_1 - (\eta_2 \circ \pi_2)\hat{\theta}_2$$
$$= \bar{i}^* \circ (j_1 \times j_2)^* \Theta_{12}' = dG.$$

As a straightforward consequence we can state the following corollaries.

Corollary 2: In the same conditions as the above proposition, we can find Poincaré generalized generating functions for ϕ and $\hat{\phi}$, respectively, that are related by $F = \eta_1^* \hat{F}$.

Corollary 3: Let $\mathscr{U}_1$ be a coordinate neighborhood of a point $x_1 \in C_1$ in P_1 and $(q^1,\dots,q^n, p_1,\dots,p_n)$ be Darboux coordinates such that the equations $p_1 = \cdots = p_k = 0$ locally define $C_1 \cap \mathscr{U}_1$. Then, there is a Poincaré generating function such that $\partial F/\partial q^i = 0$, $i = 1,\dots,k$, for each canonical transformation.

Proof: It is an obvious consequence of the form $F = \eta_1^* \hat{F}$ because of the tangency of the vector fields $\{\partial/\partial q^i\}_{i=1}^k$ to the kernel of $j_1^* \Omega_1$ in C_1.

This fact is worthy of note: the generating functions F do not depend on the gauge variables.

Before studying mixed generating functions for generalized canonical transformations we introduce some notations. The neighborhoods of P_i in which θ_i is locally defined will be denoted by $\mathscr{U}_i$ $(1 = 1,2)$. By $x_1 = (q^1,\dots,q^n, p_1,\dots,p_n)$ we mean a set of canonical coordinates for $\mathscr{U}_1$ such that the set $\mathscr{V}_1 = \mathscr{U}_1 \cap C_1$ is defined by the vanishing of the first k p's.

Lemma 1: Let $(\mathscr{U}_1, x_1)$ be a canonical neighborhood of $m_1 \in C_1$ as defined above. For every canonical transformation (Φ,ϕ) from (P_1,C_1,Ω_1) to (P_2,C_2,Ω_2), there exists a canonical neighborhood $(\mathscr{U}_2, x_2)$ of $\phi(m_1)$ such that $\phi(\mathscr{V}_1) \subset \mathscr{U}_2 \subset \Phi(\mathscr{U}_1)$ and if $x_2 = (Q^1,\dots,Q^n, P_1,\dots,P_n)$, we have $Q^i \circ \phi = q^i$, $i = 1,\dots,n$ and $P_{k+i} \circ \phi = p_i$, $i = 1,\dots,n-k$.

Proof: The point is that as Φ is not a symplectomorphism, $\Phi(\mathscr{U}_1)$ is not a canonical neighborhood. We remark that $\phi(\mathscr{V}_1) \subset C_2$ because $\phi(C_1) \subset C_2$. There exists a canonical neighborhood $\mathscr{U}'$ of $\phi(m_1)$ in P_2 such that $\mathscr{U}' = \Phi(\mathscr{U}_1)$, but what we need is that $\phi(\mathscr{V}_1) \subset \mathscr{U}' \subset \Phi(\mathscr{U}_1)$, and it can be found as follows: $\phi(\mathscr{V}_1)$ is a coisotropic submanifold of (P_2,Ω_2) and we know that there is a tubular neighborhood $\mathscr{W}$ of $\phi(\mathscr{V}_1)$ in (P_2,Ω_2) symplectomorphic to a tubular neighborhood of the canonical coisotropic imbedding of $\phi(\mathscr{V}_1)$. Then, we can choose $\mathscr{U}_2 = \Phi(\mathscr{U}_1) \cap \mathscr{W}$ and the coordinate functions given by those of the coisotropic imbedding using the identification by the local symplectomorphism, and on $\phi(\mathscr{V}_1)$ the set of coordinates given by $Q^i = q^i \circ \phi^{-1}$, $P_i = p_i \circ \phi^{-1}$. This is a canonical set satisfying the required conditions.

Instead of using the projection of graph $\phi \cap (\mathscr{U}_1 \times \mathscr{U}_2)$ on $\mathscr{V}_1$ we can also project on other sets and in this way we can define generating functions that are not of type I. The neighborhood $\mathscr{U}_{12} = \mathscr{U}_1 \times \mathscr{U}_2$ is identified with an open set of $\mathbb{R}^{2n} \times \mathbb{R}^{2n}$ via the map $x_1 \times x_2$. If we think of $\mathbb{R}^{2n} \times \mathbb{R}^{2n}$ as the product $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$, each factor being labeled by a number $\alpha = 1,\dots,4$, the map that projects canonically in the $i \times j$ factor will be called π^{ij}. So we can construct the following six functions from $\mathscr{U}_{12}$ to $\mathbb{R}^n \times \mathbb{R}^n$, $\{\tau^{ij} = \pi^{ij} \circ (x_1 \times x_2)\}$. This family of functions defines a family of functions parametrized with different sets of variables associated by the Weinstein function defined on graph ϕ that we will denote by $F_{ij} = G(\tau^{ij})^{-1}$ when $(\tau^{ij})^{-1}$ exists. The first function F_{12} is the coordinate representation of the usual Poincaré generating function for (Φ,ϕ), and the last one, F_{34}, the Poincaré generating function for (Φ^{-1},ϕ^{-1}). The other ones are the generalized mixed generating functions of type (ij) for (Φ,ϕ) and their main properties will be described in the theorem below. There is an important point to be remarked here, on the definition of the mixed generating functions F_{ij}. We have pointed out that it is necessary that there exists $(\tau^{ij})^{-1}$ and this is equivalent to the fact that the submanifold graph ϕ in $P_1 \times P_2$ is transverse to the function τ^{ij}; that is, denoting by $T^{ij}(p)$ the set of points which are mapped in $p \in \mathbb{R}^n \times \mathbb{R}^n$ by τ^{ij}, $T^{ij}(p) = \tau^{ij-1}(p)$, $p \in \tau^{ij}(\mathscr{U}_{12})$, graph ϕ will be transverse in the point $(x,\phi(x)) \in$ graph ϕ to τ^{ij} if

$$T_{(x,\phi(x))}\,\text{graph}\,\Phi \oplus T_{(x,\phi(x))} T^{ij}(\tau^{ij}(x,\phi(x))) = T_{(x,\phi(x))}(P_1 \times P_2).$$

If it occurs we will be able to parametrize locally the submanifold graph ϕ (or graph Φ) by means of the function τ^{ij} and then there will exist τ^{ij-1}. Using these conditions in the following we can state Theorem 12.

Theorem 12: With the notation defined above, if (Φ,ϕ) is a canonical transformation from (P_1,C_1,Ω_1) to (P_2,C_2,Ω_2), locally we have

$$p_i = \frac{\partial F_{13}}{\partial q^i}, \quad i = k+1,\dots,n,$$
$$P_i = -\frac{\partial F_{13}}{\partial Q^i}, \quad i = k+1,\dots,n,$$

F_{13} being the mixed generating function of type (1.3) for the transformation.

Proof: The proof is a simple matter of computing the coordinate expression of the Weinstein generating function. That is, since

$$\Theta_{12} = \pi_1^* \theta_1 - \pi_2^* \theta_2 = \sum_{i=1}^n p_i\, dq^i - \sum_{i=1}^n P_i\, dQ^i,$$

then

$$dF_{12} = (\tau^{13})^{-1*}\, dG$$
$$= (\tau^{13})^{-1*}\Big(\sum (p_i\, dq^i - P_i\, dQ^i)\Big)$$

and then

$$\frac{\partial F_{13}}{\partial q^i} = p_i, \quad i = k+1,\dots n$$

and

$$-\frac{\partial F_{13}}{\partial Q^i} = P_i, \quad i = k+1,\dots n.$$

Finally, by the construction of the canonical neighborhood $\mathcal{U}_2$ in P_2 we have that $P_i \circ \phi = p_i$, $i = k+1,...n, Q^i \circ \phi = q^i, i = 1,...,n$.

There exists a similar theorem for each mixed function F_{14}, F_{23}, and F_{24}, and it is very interesting to notice that they define locally the canonical transformation only for $(n-k)\times(n-k)$ variables on the submanifold C_1. In the particular case of the mixed function of type (1,4), the equations before become

$$\frac{\partial F_{14}}{\partial q^i} = p_i, \quad i = k+1,...,n,$$
$$-\frac{\partial F_{14}}{\partial P_i} = Q_i, \quad i = k+1,...,n,$$

showing that the gauge ambiguity does not permit the complete reconstruction of the transformation on the manifold C_1 from the generating functions.

The extended generating functions $\tilde{G},\tilde{F}$, defined in Sec. V, give locally a symplectomorphism $(\tilde{\Phi},\phi)$ such that it coincides with ϕ in $\mathscr{V}_1$, but in general, as pointed out in Sec. IV, it will not be possible to extend such a symplectomorphism to a global one, and it will not be possible to construct smoothly a set of such functions such that their graphs overlap correctly.

VII. CONCLUSIONS

We have introduced the concept of canonical transformation that generalizes the concept introduced for regular systems (see, e.g., Ref. 11), time-dependent systems,[24] and canonical systems.[9] The generalization is based on Theorem 3 where the possibility of finding a symplectic manifold P in an essentially unique way is shown, such that the final constraint submanifold C is coisotropically imbedded in P and for any dynamical vector field Γ compatible with C there is a (no uniquely defined) vector field on P with the same restriction on C (up to identification of C with its image). Furthermore, the result of Theorem 7 shows the possibility of studying the canonical transformations using only their local structure and the crucial point is that every canonical transformation defines a symplectic transformation in the (symplectic) reduced space and it is possible to define canonical transformations of the presymplectic space that are trivial on the quotient space; they will be called gauge transformations. In fact, if we start with a gauge theory as is usually meant it will be a presymplectic system and the group of gauge transformations as defined above coincides with the gauge group of the theory.

It is remarkable that the equations of motion can now be considered as a one-parameter family of canonical transformations. Moreover, the equation for the determination of the generating function is but the generalized Hamilton–Jacobi equation. These and other applications will be given in a subsequent paper.

[1] P. A. M. Dirac, "Generalized Hamiltonian dynamics," Can. J. Math. **2**, 129 (1950).
[2] M. J. Gotay, Ph.D thesis, University of Maryland, 1979.
[3] A. Lichnerowicz, "Varieté symplectique et dynamique associée a une sousvarieté," C. R. Acad. Sci. Paris Ser. A **280**, 523 (1975).
[4] J. Sniatycki, "Dirac brackets in geometric dynamics," Ann. Inst. H. Poincaré **20**, 365 (1974).
[5] G. Marmo, N. Mukunda, and J. Samuel, "Dynamics and symmetry for constrained systems: A geometrical analysis," Riv. Nuovo Cimento **6**, (1983).
[6] S. Shanmugadhasan, "Canonical formalism for degenerate Lagrangians," J. Math. Phys. **14**, 677 (1973).
[7] D. Dominici and J. Gomis, "Poincaré–Cartan integral invariant canonical transformations for singular Lagrangians," J. Math. Phys. **21**, 2124 (1980); Addendum **23**, 256 (1982).
[8] D. Dominici, G. Longhi, J. Gomis, and J. M. Pons, "Hamilton–Jacobi theory for constrained systems," J. Math. Phys. **25**, 2439 (1984).
[9] J. Gomis, J. Llosa, and N. Román, "Lee-Hwa-Chung theorem for presymplectic manifolds. Canonical transformations for constrained systems," J. Math. Phys. **25**, 1348 (1984).
[10] L. Hwa-Chung, "The universal integral invariants of Hamiltonian systems and application to the theory of canonical transformations," Proc. R. Soc. London Ser. A **12**, 237 (1947).
[11] A. Weinstein, "Lectures on symplectic manifolds," C.B.M.S. Regional Conf. Ser. Math. **29** (1979).
[12] P. A. M. Dirac, "Lectures on Quantum Mechanics," Belfer Graduate School of Science, Yeshiva University, 1964.
[13] P. G. Bergmann and I. Goldberg, "Dirac bracket transformations in phasespace," Phys. Rev. **98**, 531 (1955).
[14] M. J. Gotay, J. M. Nester, and G. Hinds, "Presymplectic manifolds and the Dirac–Bergmann theory of constraints," J. Math. Phys. **19**, 2388 (1978).
[15] M. J. Bergvelt, "Yang–Mills theories as constrained Hamiltonian systems," Preprint ITFA; 83-5, University of Amsterdam, 1983.
[16] M. J. Gotay, "On coisotropic imbeddings of presymplectic manifolds," Proc. Am. Math. Soc. **84**, 111 (1982).
[17] R. Abraham and J. Marsden, *Foundations of Mechanics* (Benjamin–Cummings, Reading, MA, 1978), 2nd ed.
[18] M. J. Gotay and J. M. Nester, "Presymplectic Lagrangian systems I. The constraint algorithm and the equivalence theorem," Ann. Inst. H. Poincaré **30**, 129 (1979).
[19] A. Weinstein, "Symplectic manifolds and their Lagrangians submanifolds," Adv. Math. **6**, 329 (1971).
[20] E. J. Saletan and A. H. Cromer, *Theoretical Mechanics* (Wiley, New York, 1971).
[21] P. K. Mitter, "Geometry of the space of gauge orbits and the Yang–Mills dynamical system," in *Recent Developments in Gauge Theories*, edited by G.'t Hooft and P. K. Mitter, (Plenum, New York, 1979), pp. 265–292.
[22] M. S. Narasimhan and T. R. Ramadas, "Geometry of SU(2) gauge fields," Commun. Math. Phys. **67**, 121 (1979).
[23] J. Marsden, *Applications of Global Analysis in Mathematical Physics* (Publish or Perish, Boston, 1974).
[24] M. Asorey, J. F. Cariñena, and L. A. Ibort, "Generalized canonical transformations for time-dependent systems," J. Math. Phys. **24**, 2745 (1983).
[25] P. K. Mitter, and C. M. Viallet, "On the bundle of connections and the gauge orbit manifold in Yang–Mills theory," Commun. Math. Phys. **79**, 457 (1981).
[26] J. Sniatycki and N. M. Tulczyjew, "Generating forms of Lagrangian submanifolds," Indiana Univ. Math. J. **22**, 267 (1972).

How to construct integrable Fokker–Planck and electromagnetic Hamiltonians from ordinary integrable Hamiltonians

J. Hietarinta
Wihuri Physical Laboratory and Department of Physical Sciences, University of Turku, 20500 Turku, Finland

(Received 28 November 1984; accepted for publication 12 April 1985)

Integrable Hamiltonians with velocity-dependent potentials, including those of Fokker–Planck-type $H = \frac{1}{2}(p_x^2 + p_y^2) + K_x p_x + K_y p_y$, are constructed from integrable Hamiltonians of type $H = \frac{1}{2}(p_x^2 + p_y^2) + V(x,y)$ using certain canonical and noncanonical transformations. Some of the Hamiltonians obtained this way are integrable only for zero energy. Candidates for the Φ potential, which is of interest for Fokker–Planck models, are constructed in several cases.

I. INTRODUCTION

The concepts of integrability have been applied to an increasing number of physical systems. Recently the integrability of Fokker–Planck-type Hamiltonians

$$H = \tfrac{1}{2} Q^{ij} p_i p_j + K^i p_i \tag{1.1}$$

has received attention.[1,2] Many statistical systems can be described by a Markov process whose probability density P solves the Fokker–Planck equation. In the weak noise limit one can make a semiclassical-type approximation for P and in that limit the equation for the stationary probability density reduces to the form

$$\tfrac{1}{2} Q^{ij}(\partial_i \Phi)(\partial_j \Phi) + K^i(\partial_i \Phi) = 0, \tag{1.2}$$

where Φ is the nonequilibrium potential (action).[1] Equation (1.2) can now be interpreted as the zero-energy Hamilton–Jacobi equation for the Hamiltonian (1.1).

Of particular interest in this context is the relationship between the integrability of the Hamiltonian (1.1) and the existence of a smooth nonequilibrium potential Φ. If the Hamiltonian H is integrable we can construct solutions of (1.2) directly: From H and the $N-1$ other invariants we can solve for p_i in terms of the q_i's and then the Hamilton–Jacobi theory states that $p_i = \partial_i \Phi$, from which Φ can be integrated. The solution can be interpreted as a nonequilibrium potential,[1,2] if first Q_{ij} is positive semidefinite, and second Φ is real, bounded from below, approaches infinity as $|q_i| \to \infty$, and is stationary on the limit sets of $\dot{q}_i = K_i$. This last condition implies that the invariants are given the constant values corresponding to $p_i = 0$ (see Ref. 2). In the following we solve for Φ in several cases using this necessary condition. All solutions will be called Φ potentials regardless of whether or not they satisfy all of the additional conditions mentioned above. In fact some of the models below cannot be interpreted as weak-noise limits of Fokker–Planck models, nevertheless the results still illustrate the connection between integrability of (1.1) and the solutions of (1.2), which is one of the main concerns of this paper.

Another physically interesting case similar to (1.1) is the Hamiltonian

$$H = \tfrac{1}{2}(p_x - A_x)^2 + \tfrac{1}{2}(p_y - A_y)^2 + V. \tag{1.3}$$

In suitable units it describes the motion of a particle in an electromagnetic field, whose vector potential is **A**, and scalar potential V. Very few integrable models of this type are known; some results in this direction have been obtained in Ref. 3. In the following we will discuss the results from this point of view as well.

In this paper we will construct several integrable Fokker–Planck or electromagnetic-type Hamiltonians. We will only consider two-dimensional systems and often transform the results in the form where $Q = I$, i.e., so that the Hamiltonian has either the form (1.3) or

$$H = \tfrac{1}{2}(p_x^2 + p_y^2) + p_x K_x + p_y K_y. \tag{1.4}$$

The models that we present have not been found by brute force searches but rather by using certain transformations on already-found integrable Hamiltonians of type

$$H = \tfrac{1}{2}(p_x^2 + p_y^2) + V(x,y). \tag{1.5}$$

The power of this approach has been shown in previous publications,[4] where, for example, the integrability of the Holt and Fokas–Lagerstrom Hamiltonians were shown to follow from that of the Henon–Heiles and r^4 potentials. An interesting feature of these transformations is that although the Hamiltonian changes in a very controlled fashion, the change in the new invariant can be more impressive. The point is that often these transformations mix coordinates and momenta and therefore they can change the order of the invariant (see, e.g., Ref. 4), and even generate rational invariants from polynomial ones. Thus we get integrable models with second invariants whose p dependence we might not have guessed.

The results given in this paper can be divided into four groups: (1) models that follow from standard-type Hamiltonians by a gauge transformation, for them $K_i = -\frac{1}{2}\partial_i \Phi$; (2) models obtained from Toda-type potentials; (3) models that follow from any integrable homogeneous potential $V(x,y)$, yielding $K_x = xV(x,y)$ and $K_y = yV(x,y)$; and (4) models that follow from quartic standard-type Hamiltonians via a more complicated set of transformations, they will have quadratic K_x and K_y. It will turn out that the same ordinary-type integrable Hamiltonian can give rise to several different Fokker–Planck or electromagnetic models. The transformations used in groups (1), (2), and (4) preserve integrability at any energy, while for (3) we obtain integrability only at zero energy. For groups (2) and (3) the results can be interpreted also as electromagnetic Hamiltonians.

 0022-2488/85/081970-06$02.50

II. THE EFFECT OF GAUGE TRANSFORMATIONS

Let us start with the Hamiltonian (1.5), which we assume to be integrable. The simplest method for obtaining Fokker–Planck-type integrable models is to solve for the Φ potential from the equation

$$(\partial_x \Phi)^2 + (\partial_y \Phi)^2 + 8V(x,y) = 0. \tag{2.1}$$

If a well-behaved solution exists and we then make the canonical gauge transformation

$$p_x = P_x - \tfrac{1}{2}\partial_x \Phi, \quad p_y = P_y - \tfrac{1}{2}\partial_y \Phi, \tag{2.2}$$

the standard-type Hamiltonian (1.5) will change to the form (1.4), with $K_i = -\frac{1}{2}\partial_i \Phi$, $i = x, y$. In this case the order of the invariant does not change.

As examples of systems obtained this way let us consider

$$\Phi = 2\alpha y^3 + x^2 y, \tag{2.3}$$

hence

$$K_x = -xy, \quad K_y = -(\tfrac{1}{2}x^2 + 3\alpha y^2). \tag{2.4}$$

The corresponding V is

$$V(x,y) = -(x^4/8 + (1+3\alpha)/2x^2y^2 + 9\alpha^2/2y^4),$$

which is known to be integrable for $\alpha = -\frac{1}{6}, \frac{1}{6}$, and $\frac{2}{3}$ [Ref. 5, models (4)2, (4)3′, and (4)4′, respectively]. Using the transformation (2.2) we get the second invariants of the Fokker–Planck system (1.4) with (2.4),

$$I(\alpha = -\tfrac{1}{6}) = y p_x - x p_y, \tag{2.5a}$$

$$I(\alpha = \tfrac{1}{6}) = p_x p_y - \tfrac{1}{4}(x^2 + y^2) p_x - xy p_y, \tag{2.5b}$$

$$I(\alpha = \tfrac{2}{3}) = x p_x p_y - y p_x^2 - \tfrac{1}{2}x^3 p_x - x^2 y p_y. \tag{2.5c}$$

Next consider

$$\Phi = a(x^2 + y^2) + b/2(x^4 + 6x^2y^2 + y^4),$$

which leads to the corresponding V,

$$V(x,y) = -\tfrac{1}{2}a^2(x^2+y^2) - ab(x^4 + 6x^2y^2 + y^4)$$
$$- \tfrac{1}{2}b^2(x^6 + 15x^4y^2 + 15x^2y^4 + y^6).$$

This V is a sixth-degree generalization of the integrable models (3)2′ and (4)3 of Ref. 5, and is therefore integrable. The corresponding Fokker–Planck-type model has

$$K_x = -ax - b(x^3 + 3xy^2),$$
$$K_y = -ay - b(3x^2y + y^3), \tag{2.6}$$
$$I = p_x p_y + K_y p_x + K_x p_y.$$

Note the simple form of the invariant, which also holds for $I(\alpha = \frac{1}{6})$ above, and can in fact be extended to any system of the form

$$K_x = f'(x+y) + g'(x-y), \quad K_y = f'(x+y) - g'(x-y), \tag{2.7}$$

$$\Phi = -2\{f(x+y) + g(x-y)\}, \tag{2.8}$$

The Φ potentials that were given above were used to construct the integrable model by way of (2.2), but they are not the only potentials associated with these integrable Hamiltonians. Let us therefore take a closer look at the problem of constructing Φ potentials for $H = 0$, $I = 0$. Solving for p_x, say, from (2.6) and substituting it to the Hamiltonian (1.4) yields a fourth-order polynomial in p_y, whose roots are

$$p_y = 0, \quad p_y = -2K_y, \quad p_y = -K_y \pm K_x. \tag{2.9}$$

From these, Φ can be solved by virtue of the integrability of the system. The first gives the trivial solution, and the second (2.8), which was used to construct the system by the gauge transformation (2.2). The last pair gives two additional potentials, for which $K_i \neq -\frac{1}{2}\partial_i \Phi$, namely

$$\Phi_+ = -2g(x-y), \quad \Phi_- = -2f(x+y). \tag{2.10}$$

Similarly for the slightly less trivial $\alpha = \frac{2}{3}$ case we get from (2.5c)

$$p_y = p_x(y p_x - \tfrac{1}{2}x^3)/(x p_x - x^2 y)$$

and substituting this into the Hamiltonian we get an equation whose roots are

$$p_x = 0, \quad p_x = 2xy, \quad p_x = xy \pm x(\tfrac{1}{2}x^2 + y^2)(x^2 + y^2)^{-1/2}. \tag{2.11}$$

The second root corresponds again to the original potential, in this case (2.3), while the last pair yields

$$\Phi = \{3x^2y + 4y^3 \pm (x^2 + 4y^2)(x^2 + y^2)^{1/2}\}/6. \tag{2.12}$$

The above steps for obtaining the Φ potential can be applied to many of the following models. Sometimes the second invariant factors nicely and each factor yields a polynomial potential while in the typical case a polynomial (in p_x, p_y, x, and y) invariant gives potentials that are only integrals of algebraic functions in x and y. It is probable that only some of these potentials have the extremal property discussed in the Introduction. But we have not investigated that.

III. TODA-TYPE POTENTIALS

As the next type let us consider the Toda potentials. For the purposes of this paper we are only interested in the so-called free end lattice. After eliminating the center of mass motion the Hamiltonian can be put in the form

$$H = \tfrac{1}{2}(p_x^2 + p_y^2) - e^{ax+cy} - e^{bx+dy}. \tag{3.1}$$

This can further be scaled and rotated so that, e.g., $c = d = 1$, $a > 0$, and $a^2 > b^2$, a convention which we adopt in the following. Another form that is sometimes more useful is obtained by the canonical transformation

$$X = ax + y, \quad Y = bx + y, \quad p_x = aP_X + bP_Y,$$
$$p_y = P_x + P_Y, \tag{3.2}$$

which produces (after scaling) the Hamiltonian

$$H = \tfrac{1}{2}(\alpha P_X^2 - 2P_X P_Y + \beta P_Y^2) - e^X - e^Y, \tag{3.3}$$

where $\alpha = (a^2+1)/(ab+1)$, $\beta = -(b^2+1)/(ab+1)$. (This does not apply to the separable case $b = -1/a$.) The Hamiltonian (3.3) was studied in Ref. 6 and it was found integrable for $\alpha = 2, \beta = 2, 1$, and $\frac{2}{3}$, corresponding to

$$a = \sqrt{3}, \quad b = -\sqrt{3}; \quad a = 3, \quad b = -2;$$
$$a = 3\sqrt{3}, \quad b = -5/\sqrt{3}; \tag{3.4}$$

and having invariants of order 3, 4, and 6, respectively.

Equation (3.3) is also transparent for creating Fokker–Planck-type Hamiltonians. There are now two canonical transformations that can be used for that purpose, namely

$$X = u + \log(p_u), \quad P_X = p_u, \tag{3.5a}$$
$$Y = v + \log(p_v), \quad P_Y = p_v, \tag{3.5b}$$
leading to
$$H_s = \tfrac{1}{2}(\alpha p_u^2 - 2 p_u p_v + \beta p_v^2) - e^u p_u - e^v p_v, \tag{3.6}$$
and
$$\begin{aligned} X &= u + \log(p_u + p_v), \quad P_X = p_u, \\ Y &= v + \log(p_u + p_v), \quad P_Y = p_v, \end{aligned} \tag{3.7}$$
$$H_d = \tfrac{1}{2}(\alpha p_u^2 - 2 p_u p_v + \beta p_v^2) - (e^u + e^v)(p_u + p_v). \tag{3.8}$$

Before going into the specific integrable cases let us note that we can immediately construct two Φ potentials to the Hamiltonian H_s, namely
$$\Phi_{s1} = 2/\alpha e^u, \quad \Phi_{s2} = 2/\beta e^v. \tag{3.9}$$
Since these Φ's will stay potentials in a canonical point transformation, this will be an easy way to construct potentials for systems of type (3.6), where we can even have arbitrary functions of u and v, respectively, in place of e^u and e^v. This result illustrates also the fact that the existence of a solution to (1.2) is a much weaker condition than integrability, since two solutions can immediately be constructed for any system of this type, but only in rare cases is the system integrable. Both of the potentials (3.9) are also smooth and bounded from below, but we leave open the question of which parameter values provide systems where these potentials satisfy the extremal properties mentioned in the Introduction.

Let us now consider the special case $a = -b = \sqrt{3}$ or $\alpha = \beta = 2$. The Hamiltonian (3.3) has in that case the second invariant
$$I(\beta = 2) = P_X^2 P_Y - P_X P_Y^2 + P_X e^Y - P_Y e^X. \tag{3.10}$$
When the transformation (3.5) is applied to this invariant it becomes
$$I(\beta = 2)_s = p_u p_v (p_u - p_v - e^u + e^v). \tag{3.11}$$
The transformation (3.5) is valid for any energy so the system $\alpha = \beta = 2$ is integrable for any energy and for any value of the constant of motion I.

When we construct the Φ potential we need to consider only those values of the energy and the constant of motion corresponding to $p_u = p_v = 0$, i.e., $E = 0$, $I = 0$. Since I in (3.11) has three factors we can obtain at least three potentials for $I = 0$: The cases corresponding to $p_u = 0$ or $p_v = 0$ were obtained before directly, for the third case we solve for p_v from the third factor and substitute it to the Hamiltonian (3.6) to obtain
$$p_u^2 - 2e^u p_u + e^{2u} - e^{u+v} = 0.$$
The two solutions to this provide us with two additional potentials
$$\Phi = \{e^{(1/2)u} \pm e^{(1/2)v}\}^2. \tag{3.12}$$

Similar calculations can be done to the remaining five Fokker–Planck systems and the calculations get progressively more difficult. The potentials that were obtained directly appear as special cases, but for the other cases one needs to solve a higher-order equation for p_u and then integrate it. It is not clear whether any single-valued potentials would be obtained this way. We will therefore just quote the invariants for the remaining integrable cases of (3.3) from which the interested reader can start with the canonical transformation (3.5) or (3.7) (for each case $\alpha = 2$):
$$\begin{aligned} I(\beta = 1) = {}& 4P_X^2 P_Y^2 - 4P_X P_Y^3 + P_Y^4 \\ & + 8e^Y P_X P_Y - 4P_Y^2(e^X + e^Y) + 4e^{2Y}, \end{aligned} \tag{3.13}$$
$$\begin{aligned} I(\beta = \tfrac{2}{3}) = {}& 4P_X^6 - 12P_X^5 P_Y + 13P_X^4 P_Y^2 - 6P_X^3 P_Y^3 + P_X^2 P_Y^4 \\ & - 12P_X^4(e^X + e^Y) + 6P_X^3 P_Y(4e^X + 3e^Y) \\ & - 2P_X^2 P_Y^2(7e^X + 3e^Y) + 2e^X P_X P_Y^3 \\ & + 3P_X^2(4e^{2X} + 8e^{X+Y} + 3e^{2Y}) \\ & - 6e^X P_X P_Y(2e^X + 3e^Y) \\ & + e^{2X} P_Y^2 - 4e^{2X}(e^X + 3e^Y). \end{aligned} \tag{3.14}$$

The canonical transformations (3.5) and (3.7) preserve integrability at any energy and therefore the above results can also be interpreted (after a rotation) as electromagnetic Hamiltonians. In addition to (3.6) and (3.8) we can obtain still other kinds of models by making the $\log(p)$ translation in only one of the coordinates, i.e., using (3.5a) or (3.5b). Thus the same integrable Toda-type Hamiltonian produces the following integrable systems (where we have rotated back to $Q = 1$, and scaled):
$$H_{5a} = \tfrac{1}{2}(p_x^2 + p_y^2) - (p_x - bp_y)e^{ax+y} - e^{bx+y}, \tag{3.15a}$$
$$H_{5b} = \tfrac{1}{2}(p_x^2 + p_y^2) - (p_x - ap_y)e^{bx+y} - e^{ax+y}, \tag{3.15b}$$
$$H_5 = \tfrac{1}{2}(p_x^2 + p_y^2) - (p_x - bp_y)e^{ax+y} + (p_x - ap_y)e^{bx+y}, \tag{3.16}$$
$$H_7 = \tfrac{1}{2}(p_x^2 + p_y^2) - p_y(e^{ax+y} + e^{bx+y}). \tag{3.17}$$

The second invariants for these systems can be obtained by applying the transformations (3.5a), (3.5b), and (3.7), respectively, on the invariants (3.10), (3.13), or (3.14) and then transforming back by (3.2), where now X and Y should be replaced by u and v. The canonical transformations (3.5) and (3.7) do not change the total order of the invariant, rather they will cause factorization, as was seen in the step form (3.10) to (3.11).

IV. HOMOGENEOUS POTENTIALS

Let us next assume that the integrable Hamiltonian (1.5) has a homogeneous V potential, whose degree M differs from -2. Many such cases are known, see, e.g., Refs. 5 and 7. In the following we will only need integrability at zero energy. In polar coordinates the Hamiltonian will be
$$H = \tfrac{1}{2}p_r^2 + \tfrac{1}{2}r^{-2}p_\theta^2 + r^M v(\theta). \tag{4.1}$$
If we now make the canonical transformation
$$r = R(RP_R)^k, \quad p_r = P_R(RP_R)^{-k}, \tag{4.2}$$
θ and p_θ unchanged, the Hamiltonian (4.1) becomes
$$\begin{aligned} H' = {}& (RP_R)^{-2k}\{\tfrac{1}{2}P_R^2 + \tfrac{1}{2}R^{-2}p_\theta^2 \\ & + (RP_R)^{k(M+2)} R^M v(\theta)\}. \end{aligned} \tag{4.3}$$
Apart from the overall factor this will be of the form (1.4) (in

polar coordinates) if we choose $k = 1/(M+2)$. (Here we need the assumption $M \neq -2$.) As we mentioned in the Introduction one is often interested in integrability at zero energy and then the above overall factor can be ignored. Thus at zero energy the integrability of (4.1) implies the integrability of

$$H = \tfrac{1}{2}(p_x^2 + p_y^2) + p_x x V(x,y) + p_y y V(x,y). \qquad (4.4)$$

Here the transformation mixes p's and q's so its effects on the invariant are interesting.

To be specific let us consider $V = -ax^2 - by^2$, which is integrable:

$$I = p_x^2 - 2ax^2. \qquad (4.5)$$

We now apply the transformation (4.2) with $k = \frac{1}{4}$ and obtain

$$K_x = -ax^3 - bxy^2, \quad K_y = -ax^2y - by^3, \qquad (4.6)$$

$$I = (xp_x + yp_y)^{-1/2}(p_x^2 - 2ax^2(xp_x + yp_y)). \qquad (4.7)$$

Now the invariant involves a square root and therefore we should actually take I^2 as the new invariant. In this way the simple polynomial invariant (4.5) becomes a rotational invariant. For this example we have also found two Φ potentials for $I = 0$,

$$\Phi = \tfrac{1}{2}(\sqrt{a}\,x^2 \pm \sqrt{b}\,y^2)^2. \qquad (4.8)$$

Most of the known integrable homogeneous potentials have an invariant that is polynomial in p (see Refs. 5 and 7). In such a case one can show in general[5] that the invariant is of the form

$$I = \sum_{n=0}^{[N/2]} \sum_{m=0}^{N-2n} p_x^m p_y^{N-2n-m} d^{\,m,N-2n}(x,y), \qquad (4.9)$$

where $d^{\,m,s}$ is a homogeneous function of x and y with degree $n_c + \frac{1}{2}M(N-s)$, where n_c is an integer, $0 \leqslant n_c \leqslant N$. When the transformation (4.2) is applied to (4.9) it gives

$$I = (xp_x + yp_y)^{(n_c - N)/(M+2)} \times \sum_{n=0}^{[N/2]} \sum_{m=0}^{N-2n} p_x^m p_y^{N-2n-m} (xp_x + p_y)^n d^{\,m,N-2n}(x,y).$$

From this we see that depending on the numbers n_c, N, and M, the invariant can have a rather bad rational dependence on the momenta. Clearly the integrability for these K_i could not be found by using a polynomial ansatz and even for rational ones one would have needed rather high powers. Considering the simple form of the Hamiltonian (4.4) this suggests that there are still many more integrable models, with rather complicated second invariants, which are waiting to be found.

As a model whose invariant is not even rational in momenta let us consider

$$H = \tfrac{1}{2}(p_x^2 + p_y^2) + x/y, \qquad (4.10)$$

which was found integrable in Ref. 8. Let us now make the canonical transformation (4.2) using $k = \frac{1}{2}$. Since the potential is of degree zero the energy of (4.10) does not have to be zero, rather it will transform to a coupling constant, providing an example of partial coupling constant metamorphosis.[4] The resulting zero-energy Hamiltonian is singular at $y = 0$ and therefore, following Ref. 2, we will make a further canonical transformation $y = 1/Y$, $p_y = -Y^2 p_Y$. The new integrable (at zero energy) Fokker–Planck-type Hamiltonian is finally

$$H = \tfrac{1}{2}(p_x^2 + y^4 p_y^2) + (-ex + x^2y)p_x + (ey - xy^2)p_y. \qquad (4.11)$$

Even though the Hamiltonian (4.11) looks quite harmless, its second (and third) invariants can be expressed only in terms of the parabolic cylinder functions W (see Ref. 8), e.g.,

$$I = z/(2y)\{p_y/zW(\tfrac{1}{2}\sqrt{e}, p_x/z) + 2W'(\tfrac{1}{2}\sqrt{e}, p_x/z)\}^2, \qquad (4.12)$$

where we have denoted $z = (xp_x - yp_y)^{1/2}$. This invariant could still have branch-type singularities due to the square root in z. To avoid these we must choose for W an even or odd solution (y_1 or y_2 of Ref. 9), rather than a standard solution W_+ or W_- used in Ref. 8. This will take care of the term in curly brackets, but due to the overall z factor we should actually take I^2 as the invariant. Another invariant would be $I(y_+)/I(y_-)$. The question of constructing the Φ potential is left open.

If the potential also has a piece whose homogeneous degree is -2 the above transformation can still be carried out, and since r^{-2} transforms like p_r^2 we see that this additional part of the potential will not change at all. Thus at zero energy the integrability of

$$H = \tfrac{1}{2}p_r^2 + \tfrac{1}{2}r^{-2}p_\theta^2 + r^M v(\theta) + r^{-2}w(\theta) \qquad (4.13)$$

implies the integrability of

$$H = \tfrac{1}{2}p_r^2 + \tfrac{1}{2}r^{-2}p_\theta^2 + (rp_r)r^M v(\theta) + r^{-2}w(\theta). \qquad (4.14)$$

Even though these Hamiltonians are only integrable for zero energy the system can be relevant for the motion of real particles if the w potential is negative enough to bind the motion.

Models with r^{-2}-type additional pieces were discussed in Ref. 10, where they were used to construct higher-dimensional integrable theories. It was found that the additional term is usually x^{-2} or y^{-2} and in rare cases $\alpha x^{-2} + \beta y^{-2}$. Reference 10 contains a table where the known potentials and invariants are given.

V. QUARTIC POTENTIALS

In this section we assume that the potential is quartic, i.e., that

$$H = \tfrac{1}{2}(p_x^2 + p_y^2) + ax^4 + bx^2y^2 + cy^4 \qquad (5.1)$$

is integrable. Let us make the point transformation

$$x = \sqrt{X}, \quad p_x = 2\sqrt{X}\,P_X, \quad y = \sqrt{Y}, \quad p_y = 2\sqrt{Y}\,P_Y, \qquad (5.2)$$

and follow it by the $p \leftrightarrow q$ reflection

$$X = -p_u, \quad P_X = u, \quad Y = -p_v, \quad P_Y = v. \qquad (5.3)$$

The resulting Hamiltonian is of the Fokker–Planck form with nondiagonal Q:

$$H' = ap_u^2 + bp_u p_v + cp_v^2 - 2u^2p_u - 2v^2p_v. \qquad (5.4)$$

Due to the simple form of K_u and K_v we can immediately construct two different Φ's for (5.4), just as was done in Sec. III:

$$\Phi = 2/(3a)u^3, \quad \Phi = 2/(3c)v^3. \qquad (5.5)$$

Again two potentials can be constructed for any value of a, b, and c, but the system is known to be integrable only for six ratios $a{:}b{:}c$. In this case the Φ's are not bounded from below, and furthermore Q is not positive definite for any of the integrable systems.

Let us now apply this to the specific example of

$$H = \tfrac{1}{2}(p_x^2 + p_y^2) + x^4/6 + x^2y^2 + y^4/6, \tag{5.6}$$

whose second invariant is [case (4)3′ in Ref. 5]

$$I = p_x p_y + 2/3xy(x^2 + y^2). \tag{5.7}$$

After the transformations (5.2) and (5.3) the Hamiltonian and the invariant become

$$H' = \tfrac{1}{6}(p_x^2 + p_y^2) + p_x p_y - 2x^2p_x - 2y^2p_y, \tag{5.8}$$

$$I' = (p_x p_y)^{1/2}(p_x + p_y - 6xy)\, 2/3. \tag{5.9}$$

However, this invariant is not acceptable as it involves a square root. Thus we must take I'^2 as the new polynomial invariant. In principle there could also be the possibility that $p_x p_y$ is a square of something *on the affine manifold* $H = 0$, but in the present case one can show explicitly that this does not work.

In general a rotation of type $(\alpha\delta - \beta\tau \neq 0)$

$$u = \alpha x + \beta y, \quad p_u = (\delta p_x - \tau p_y)/(\alpha\delta - \beta\tau),$$

$$v = \tau x + \delta y, \quad p_v = (-\beta p_x + \alpha p_y)/(\alpha\delta - \beta\tau),$$

will put the Hamiltonian (5.4) into the form (1.5). We fix this final rotation by requiring that K_x has no y^2 part. This defines the parameter values up to an overall factor, which we choose for convenience only. For the Hamiltonian (5.8) we take

$$\alpha = -\tau = i(3\sqrt{2})^{-1/3}, \quad \beta = \delta = -(\tfrac{2}{3})^{1/3},$$

and these yield finally a Fokker–Planck-type Hamiltonian

$$H = \tfrac{1}{2}(p_x^2 + p_y^2) + 4xyp_x + (-x^2 + 2y^2)p_y \tag{5.10}$$

and the invariant

$$I = (2p_x^2 + p_y^2)(p_y + 2x^2 + 4y^2)^2. \tag{5.11}$$

Thus starting from a rather trivial integrable model (5.6) with an invariant (5.7) quadratic in p we have obtained through a sequence of canonical transformations the above integrable Fokker–Planck-type model with a fourth-order invariant. The order of the invariant had to doubled to eliminate the square roots.

The above computations can be carried out also for the two other integrable models of type (5.6) [cases (4)4 and (4)5 of Ref. 5]. The results are

$$K_x = x(i3\sqrt{5}x + 4y), \quad K_y = 5x^2 + 2y^2, \tag{5.12}$$

$$I = (2p_x - i\sqrt{5}p_y)(p_x - i\sqrt{5}p_y)^2(p_x + 5x(i\sqrt{5}x + 2y)^2), \tag{5.13}$$

and

$$K_x = x(i7x + 2y), \quad K_y = 10x^2 + y^2, \tag{5.14}$$

$$\begin{aligned} I = (p_x - i5p_y)^2[\, & p_x^2 - i2p_xp_y - p_y^2 \\ & + 4(i5x^2 + 2xy - i2y^2)p_x \\ & - 4(2x^2 - i8xy + y^2)p_y - 4(i2x + y)^4]. \end{aligned} \tag{5.15}$$

Note that the second invariants (5.11) and (5.13) factorize completely over the ring of polynomials of x and y. This gives us a straightforward method of obtaining Φ potentials for $I = 0$: For each linear (in p) factor of I in (5.11) and (5.13) we can solve for p_x and p_y at $H = 0$ without square roots and then a simple integration produces a polynomial Φ. Our results are

$$\begin{aligned} \Phi &= \tfrac{1}{3}(\pm i\sqrt{2x} - 2y)^3 \\ &\quad \text{or } -\tfrac{1}{3}(\pm i\sqrt{2x} - 2y)^2(\pm i\sqrt{2x} + y), \end{aligned} \tag{5.16}$$

$$\begin{aligned} \Phi &= \tfrac{2}{3}(i\sqrt{5x} + 2y)^3 \text{ or } \tfrac{1}{3}(i\sqrt{5x} + y)^3 \\ &\quad \text{or } -\tfrac{5}{3}{}^2(i\sqrt{5x} + 3y), \end{aligned} \tag{5.17}$$

for (5.11) and (5.13), respectively. From (5.15) only one linear factor can be extracted, although from (5.5) we know that there are two polynomial potentials

$$\Phi = \tfrac{1}{36}(i5x + y)^3 \text{ and } \tfrac{2}{9}(i2x + y)^3. \tag{5.18}$$

We have not searched for the remaining potentials corresponding to the second factor of (5.15).

All of the potentials above are complex and the imaginary unit is always associated with the x coordinate. It might therefore be appropriate to scale the systems (5.16), (5.17), and (5.18) by $x = ikX$, $p_x = -iP_X/k$, with $k = \sqrt{2}, 1/\sqrt{5}$, 1, respectively. In this way one obtains hyperbolic systems, whose invariants and potentials are all real and have integer coefficients.

VI. CONCLUSIONS

In the results above we have illustrated the fact that one can obtain many new integrable models using suitable transformations on old ones. This is quite economical and should always be tried when one applies the concepts of integrability in a new situation. Several transformations, canonical or noncanonical, have been applied previously in Ref. 4 to relate the Holt and Henon–Heiles Hamiltonians, and in Ref. 10 to generalize to any dimension some models that were previously found integrable in two dimensions. Some of the resulting Hamiltonian systems above were integrable only for zero energy, but that is sufficient for Fokker–Planck models, and sometimes for other systems as well. In other cases integrability at any energy was retained, which permits further applications, e.g., describing the motion of a particle moving subject to velocity-dependent forces.

ACKNOWLEDGMENT

I would like to thank D. Roekaerts for discussions about Fokker–Planck models, for sending me Ref. 2 before publication, and for comments about the manuscript.

[1]R. Graham and T. Tel, Phys. Rev. Lett. **52**, 9 (1984); J. Stat. Phys. **35**, 729 (1984); **37**, 709 (1984).

[2]R. Graham, D. Roekaerts, and T. Tel, "On the integrability of Hamiltonians associated with Fokker–Planck equations," Universitat Essen preprint, 1984.

[3]L. S. Hall, Physica (Utrecht) D **8**, 90 (1983); B. Dorizzi, B. Grammaticos, A. Ramani, and P. Winternitz, preprint CRMA-1228, 1984.

[4]J. Hietarinta, B. Grammaticos, B. Dorizzi, and A. Ramani, Phys. Rev. Lett. **53**, 1707 (1984).
[5]J. Hietarinta, Phys. Lett. A **96**, 276 (1983).
[6]B. Dorizzi, B. Grammaticos, R. Padjen, and V. Papageorgiou, J. Math. Phys. **25**, 2200 (1984).
[7]J. Hietarinta, Phys. Rev. A **28**, 3670 (1983); B. Grammaticos, B. Dorizzi, and A. Ramani, J. Math. Phys. **25**, 3470 (1984).
[8]J. Hietarinta, Phys. Rev. Lett. **52**, 1057 (1984).
[9]*Handbook of Mathematical Functions*, edited by M. Abramowitz and I. Stegun (Dover, New York, 1972), Chap. 19.
[10]B. Grammaticos, B. Dorizzi, A. Ramani, and J. Hietarinta, "Extending integrable Hamiltonian systems from 2 to N dimensions," preprint, 1984.

Lax-pair operators for squared-sum and squared-difference eigenfunctions

Yoshi-Hiko Ichikawa
Institute of Plasma Physics, Nagoya University, Nagoya 464, Japan
Kazu-Hiro Ino
Taiko Electronics Co. Ltd., Kanda, Tokyo 101, Japan

(Received 23 November 1984; accepted for publication 5 April 1985)

An interrelationship between various representations of the inverse scattering transformation is established by examining eigenfunctions of Lax-pair operators of the sine–Gordon equation and the modified Korteweg–de Vries equation. In particular, it is shown explicitly that there exist Lax-pair operators for the squared-sum and squared-difference eigenfunctions of the Ablowitz–Kaup–Newell–Segur inverse scattering transformation.

I. INTRODUCTION

In a series of analyses of the inverse scattering transformation for typical nonlinear evolution equations,[1,2] we have shown that the Lax-pair operators **L** and **A** constructed by the Chen–Lee–Liu algorithm[3] are nothing but the Lax-pair operators for the squared eigenfunctions of the Ablowitz–Kaup–Newell–Segur scheme (in short the AKNS scheme). Although the eigenvalue problem for the squared eigenfunctions has been studied in the literature to prove the closure and orthogonality properties of the squared eigenfunctions,[4,5] it was not shown before our studies that the squared eigenfunctions possess their own Lax-pair operators **L** and **A**.

Recently, we have derived alternative Lax-pair operators for the sine–Gordon equation,[6] which stands for the squared-sum eigenfunctions of the AKNS scheme. Here, first, we are going to present another set of Lax-pair operators for the sine–Gordon equation in the second section. We notice that this new set of Lax-pair operators stand for the squared-difference eigenfunctions of the celebrated AKNS scheme.

Turning to the modified Korteweg–de Vries equation, we have shown that the Chen–Lee–Liu algorithm gives rise to the Lax-pair operators for the squared sum eigenfunction of the modified Korteweg–de Vries equation. Thus, we are led to examine the Lax-pair operator for the squared-difference eigenfunction of the modified Korteweg–de Vries in the third section.

II. LAX-PAIR OPERATORS OF THE SINE–GORDON EQUATION

For the sine–Gordon equation in light-cone coordinates

$$\frac{\partial^2}{\partial t\,\partial x}\theta = \sin\theta, \tag{1}$$

following the Chen–Lee–Liu algorithm, we take an infinitesimal transformation

$$\theta \to \theta + \epsilon\phi, \tag{2}$$

and linearize the given nonlinear evolution equation, which yields

$$\frac{\partial^2}{\partial t\,\partial x}\phi = \cos\theta\phi. \tag{3}$$

Defining a function Φ by

$$\Phi = \frac{\partial}{\partial x}\phi, \tag{4}$$

we can express Eq. (3) in the form

$$\frac{\partial}{\partial t}\Phi = A_{(-)}\Phi, \tag{5}$$

where the temporal evolution operator is defined by

$$\mathbf{A}_{(-)} = \cos\theta\int_{-\infty}^{x} dy. \tag{6}$$

Since it will be shown later that Eq. (5) stands for the squared-difference eigenfunctions of the AKNS scheme, we attached the subscript $(-)$ for the operator defined by Eq. (6).

Corresponding to this operator $\mathbf{A}_{(-)}$, we can construct an eigenvalue operator $\mathbf{L}_{(-)}$ along the line suggested by the Chen–Lee–Liu algorithm, but skipping a step of deriving conserved densities. Assuming the operator $\mathbf{L}_{(-)}$ in the form of

$$\mathbf{L}_{(-)} = \frac{\partial^2}{\partial x^2} + a, \tag{7}$$

where a is at most of order θ^2, we proceed to determine $\mathbf{L}_{(-)}$ on the basis of perturbational considerations. We express

$$\theta^{(1)} = \sum \theta_m^{(1)}\exp\left(\frac{t}{m} + mx\right), \tag{8}$$

$$\Phi^{(0)} = \sum \Phi_k^{(0)}\exp\left(\frac{t}{k} + kx\right). \tag{9}$$

Then, up to the second order in θ, we can decompose the operators $\mathbf{L}_{(-)}$ and $\mathbf{A}_{(-)}$ according to

$$\mathbf{L}_{(-)}^{(0)} = \frac{\partial^2}{\partial x^2}, \quad \mathbf{L}_{(-)}^{(2)} = a^{(2)}, \tag{10a}$$

$$\mathbf{A}_{(-)}^{(0)} = \int_{-x}^{x} dy, \quad \mathbf{A}_{(-)}^{(2)} = -\frac{\theta^2}{2}\int_{-\infty}^{x} dy. \tag{10b}$$

A second-order expansion of the compatibility operator equation,

$$\frac{\partial}{\partial t}L_{(-)}^{(2)} = [A_{(-)}^{(0)}, L_{(-)}^{(2)}] + [A_{(-)}^{(2)}, L_{(-)}^{(0)}], \tag{11}$$

gives rise to

 0022-2488/85/081976-03$02.50

$$(a\Phi)^{(2)}_{k,l,m} = \sum_{l,m}\left\{lm + \frac{1}{2}lm\left(\frac{m}{l+k} + \frac{k}{k+m}\right)\right\}\theta^{(1)}_{l}\theta^{(1)}_{m}\Phi^{(0)}_{k}. \tag{12}$$

Hence, taking the inverse Fourier transformation, we get

$$L^{(2)}_{(-)} = \theta^2_x + \theta_{xx}\int_{-\infty}^{x} dy\,\theta_y, \tag{13}$$

where the suffix x denotes a partial differentiation with respect to x. It is straightforward to confirm that the eigenvalue equation

$$\mathbf{L}_{(-)}\Phi = \lambda_{(-)}\Phi, \tag{14a}$$

with

$$\mathbf{L}_{(-)} = \frac{\partial^2}{\partial x^2} + \theta^2_x + \theta_{xx}\int_{-\infty}^{x} dy\,\theta_y, \tag{14b}$$

and the temporal evolution equation (5) form the Lax-pair operator equations for the sine–Gordon equation (1).

Furthermore, we can explicitly show that the eigenfunction Φ and the eigenvalue $\lambda_{(-)}$ of Eq. (14a) are expressed as

$$\Phi = v_1^2 - v_2^2, \quad \lambda_{(-)} = -4\zeta^2 \tag{15}$$

in terms of the eigenfunctions v_1, v_2, and the eigenvalue ζ of the AKNS scheme,

$$\mathscr{L}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = -i\zeta\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \tag{16a}$$

$$\frac{\partial}{\partial t}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \mathscr{A}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \tag{16b}$$

with the definition of

$$\mathscr{L} = \begin{pmatrix} \dfrac{\partial}{\partial x} & \dfrac{1}{2}\theta_x \\ -\dfrac{1}{2}\theta_x & \dfrac{\partial}{\partial x} \end{pmatrix}, \tag{17a}$$

$$\mathscr{A} = \frac{i}{4\zeta}\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{17b}$$

That is to say, the Lax-pair operators $\mathbf{L}_{(-)}$ and $\mathbf{A}_{(-)}$ stand for the squared-difference eigenfunction Φ of the AKNS scheme.

We have shown[2] that the squared-sum eigenfunction Ψ defined as

$$\Psi = v_1^2 + v_2^2, \tag{18}$$

possesses the Lax-pair operators $\mathbf{L}_{(+)}$ and $\mathbf{A}_{(+)}$ defined by

$$\mathbf{L}_{(+)}\Psi = \lambda_{(+)}\Psi, \tag{19a}$$

$$\mathbf{L}_{(+)} = \frac{\partial^2}{\partial x^2} + \theta_x\int_{-\infty}^{x} dy\,\theta_y\frac{\partial}{\partial y}, \tag{19b}$$

and

$$\frac{\partial}{\partial t}\Psi = \mathbf{A}_{(+)}\Psi, \tag{20a}$$

$$\mathbf{A}_{(+)} = \int_{-\infty}^{x} dy\cos\theta. \tag{20b}$$

It is worthwhile to remark that Eqs. (20a) and (20b) are adjoint equations of Eqs. (5) and (6).

III. ALTERNATIVE LAX-PAIR OPERATORS OF THE MODIFIED KORTEWEG–DE VRIES EQUATION

Having found that the squared-difference eigenfunction has its own Lax-pair operators for the sine–Gordon equation, we turn to examine a case of the modified Korteweg–de Vries equation,

$$\frac{\partial}{\partial t}q + 6q^2\frac{\partial}{\partial x}q + \frac{\partial^3}{\partial x^3}q = 0, \tag{21}$$

for which the Lax-pair operators in the AKNS scheme are defined as

$$\mathscr{L}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = i\zeta\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \tag{22a}$$

$$\mathscr{L} = \begin{pmatrix} -\dfrac{\partial}{\partial x}, & q \\ q, & +\dfrac{\partial}{\partial x} \end{pmatrix}, \tag{22b}$$

and

$$\frac{\partial}{\partial t}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \mathscr{A}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} A & B \\ C & -A \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \tag{23a}$$

with

$$A = -i\zeta^3 + 2iq^2\zeta, \tag{23b}$$

$$B = 49\zeta^2 + 2iq_x\zeta - 2q^3 - q_{xx}, \tag{23c}$$

$$C = -49\zeta^2 + 2iq_x\zeta + 2q^3 + q_{xx}. \tag{23d}$$

Following the Chen–Lee–Liu algorithm, we have shown that Eq. (21) possesses the Lax-pair operators $\mathbf{L}_{(+)}$ and $\mathbf{A}_{(+)}$ defined as

$$\mathbf{L}_{(+)}\Psi = \lambda_{(+)}\Psi, \tag{24a}$$

$$\frac{\partial}{\partial t}\Psi = \mathbf{A}_{(+)}\Psi \tag{24b}$$

with

$$\mathbf{L}_{(+)} = \frac{\partial^2}{\partial x^2} + 4q\int_{-\infty}^{x} dy\,q\frac{\partial}{\partial y}, \tag{25a}$$

$$\mathbf{A}_{(+)} = -6q^2\frac{\partial}{\partial x} - \frac{\partial^3}{\partial x^3}. \tag{25b}$$

The eigenfunction Ψ and the eigenvalue $\lambda_{(+)}$ are expressed in terms of the eigenfunction and eigenvalue of the AKNS scheme by

$$\Psi = v_1^2 + v_2^2, \quad \lambda_{(+)} = -4\zeta^2, \tag{26}$$

namely, the eigenfunction Ψ is the squared-sum eigenfunction of the AKNS scheme.

Referring to the case of the sine–Gordon equation, we construct the eigenvalue problem for the squared-difference eigenfunction

$$\Phi = v_1^2 - v_2^2. \tag{27}$$

Briefly, we obtain

$$\mathbf{L}_{(-)}\Phi = \lambda_{(-)}\Phi, \tag{28a}$$

with

$$\mathbf{L}_{(-)} = \frac{\partial^2}{\partial x^2} + 4q^2 + 4q_x\int_{-\infty}^{x} dy\,q, \quad \lambda_{(-)} = 4\zeta^2. \tag{28b}$$

The corresponding temporal evolution is constructed from Eqs. (22) and (23) by

$$\frac{\partial}{\partial t}\Phi = \mathbf{A}_{(-)}\Phi, \tag{29a}$$

with

$$\mathbf{A}_{(-)} = -\frac{\partial^3}{\partial x^3} - 6q^2\frac{\partial}{\partial x} - 12qq_x. \tag{29b}$$

Here, it should be observed that Eq. (29b) is nothing but a straightforward linearization of the modified Korteweg–de Vries equation.

IV. CONCLUDING DISCUSSION

In a series of analyses of structure of Lax-pair operators, we have shown that, at least within the scheme of the Ablowitz–Kaup–Newell–Segur inverse scattering transformation, the squared-sum and the squared-difference eigenfunctions possess their own Lax-pair operators. In particular, the temporal evolution operators of the squared-sum or the squared-difference eigenfunctions are identified by taking an infinitesinal transformation of the given nonlinear evolution equations. Accordingly, we have established an interrelationship between the Chen–Lee–Liu algorithm and the Ablowitz–Kaup–Newell–Segur scheme.

However, our preliminary analysis of the inverse scattering transformation scheme of Wadati–Konno–Ichikawa[7] has failed to yield Lax-pair operators for their squared eigenfunctions. Hence, we conjecture that the Chen–Lee–Liu algorithm is not able to construct Lax-pair operators for those highly nonlinear evolution equations which are shown to be integrable by the Wadati–Konno–Ichikawa inverse scattering transformation. This is not surprising, because the Chen–Lee–Liu algorithm is based on perturbational analysis of nonlinear mode–mode coupling in constructing consistent eigenvalue operators.

[1]K. H. Iino, Y. H. Ichikawa, and M. Wadati, J. Phys. Soc. Jpn. **51**, 3724 (1982).
[2]K. H. Iino and Y. H. Ichikawa, J. Phys. Soc. Jpn. **51**, 4091 (1982).
[3]H. H. Chen, Y. C. Lee, and C. S. Liu, Physica Scr. **20**, 490 (1979).
[4]M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[5]D. J. Kaup, J. Math. Phys. **25**, 2467 (1984).
[6]Y. H. Ichikawa, *Advances in Nonlinear Waves*, edited by L. Debnath (Pitman, New York, 1984), Vol. II.
[7]Y. H. Ichikawa, K. Konno, and M. Wadati, in *Long-Time Prediction in Dynamics*, edited by C. W. Horton, Jr., L. E. Reichl, and V. G. Szebehely (Wiley, New York, 1983), p. 345.

On the convergence of the Rytov approximation for the reduced wave equation

V. H. Weston
Division of Mathematical Science, Purdue University, West Lafayette, Indiana 47907

(Received 24 April 1984; accepted for publication 19 April 1985)

The reduced wave equation $\Delta u + k^2 n^2(x) u = 0$ is treated where $n(x)$ fluctuates about unity in a compact domain D, and is equal to unity in the region exterior to D. In the Rytov approximation the total field $u(x)$ generated by an incident field $u^i(x)$ has the form $u(x) \sim u^i(x) \exp \phi(x)$ where $\phi(x) u^i(x) = (1/4\pi) \int_D (e^{ik|x-y|}/|x-y|) k^2 (n^2 - 1) u^i(y) d\tau_y$. It is shown that under suitable conditions on $n(x)$ and restrictions on $u^i(x)$, the approximation is a leading term of a convergent expansion holding for all x in D. This is in contrast to previous theory, which treated the Rytov approximation as an asymptotic expansion valid in the forward-scattered direction.

I. INTRODUCTION

The starting point for the analysis is the reduced wave equation

$$\Delta u + k^2 n^2(x) u = 0, \quad x \in R^3, \tag{1}$$

where the associated time dependence $e^{-i\omega t}$ of the quantities is suppressed. The index of refraction $n(x)$, a real bounded function, will be taken to be unity exterior to a simply connected bounded region D. The region D contains the scattering object.

The total field $u(x)$ consists of the sum of an incident wave $u^i(x)$ and a scattered wave $u^s(x)$,

$$u(x) = u^i(x) + u^s(x). \tag{2}$$

The incident field is produced by sources exterior to D, and therefore satisfies the nonhomogeneous equation

$$\Delta u^i + k^2 u^i = \rho(x),$$

where the source term $\rho(x)$ vanishes in $\overline{D}$. For the case of plane wave incidence, $\rho(x)$ vanishes everywhere. The scattered field $u^s(x)$ satisfies the radiation condition for $|x| = r \to \infty$,

$$\lim_{r \to \infty} \left(\frac{\partial u^s}{\partial r} - iku^s \right) = 0.$$

The Rytov approximation[1,2] is obtained in the following manner. The total field is decomposed in terms of a complex amplitude term and a complex phase,

$$u = u^i + u^s = \psi e^{\phi}, \quad \text{where} \quad \psi = u^i. \tag{3}$$

Upon insertion of ψe^{ϕ} into Eq. (1), the following equation is obtained[3] for ϕ:

$$(\Delta + k^2)(\psi\phi) = -[\nabla\phi \cdot \nabla\phi + k^2(n^2 - 1)]\psi. \tag{4}$$

This holds everywhere in R^3 except where the source term $\rho(x)$ is present. The Rytov approximation is given by neglecting the first term on the right-hand side of Eq. (4), yielding

$$(\Delta + k^2)(\psi\phi) \sim -k^2(n^2 - 1)\psi,$$

and then taking as the solution

$$\phi(x) \sim \frac{k^2}{4\pi} \int_D \frac{e^{ik|x-y|}}{|x-y|} [n^2(y) - 1] \frac{u^i(y)}{u^i(x)} d\tau_y. \tag{5}$$

A formal "asymptotic" expansion is often presented by just formally iterating Eq. (4), starting with relation (5) as the leading term. It is justified by assuming that $|\nabla\phi| \ll 2\pi/\lambda$ and $|n - 1| \ll 1$. However, there have been questions on the validity of the expansion. It has been argued that the domain of validity is no greater than the domain specified by the Born approximation.

For comparison the Born approximation is given as follows. The scattering problem is formulated in terms of the well-known integral equation for $u^s(x)$,

$$u^s(x) = u^s_B(x) + k^2 \int_D G_0(x,y)[n^2(y) - 1] u^s(y) d\tau_y, \tag{6}$$

where $G_0(x,y)$ is the Green's function

$$G_0(x,y) = (1/4\pi)(e^{ik|x-y|}/|x-y|), \tag{7}$$

and $u^s_B(x)$ is given by

$$u^s_B(x) = k^2 \int G_0(x,y)[n^2(y) - 1] u^i(y) d\tau_y. \tag{8}$$

If the norm of the operator in Eq. (6) is less than unity,

$$\sup_u \frac{\| \int_D G_0(x,y) k^2 [n^2(y) - 1] u(y) d\tau_y \|}{\|u\|} < 1, \tag{9}$$

where the uniform norm $\|u\| = \operatorname{Max}_{x \in D} |u(x)|$ is taken, then the integral equation (6) can be solved by iteration (Neumann series), with convergence in the sense of uniform convergence. If the quantity on the left-hand side of inequality (9) is much less than unity, then the solution is given by the Born approximation $u^s(x) \sim u^s_B(x)$.

As opposed to the Born approximation, which has a rigorous mathematical justification, there is no rigorous justification of the Rytov approximation. What analysis there is, is asymptotic in nature.[4,5] However Brown[4] does not indicate that the usual assumptions $|\nabla\phi| \ll 2\pi/\lambda$, $|n - 1| \ll 1$, are not sufficient.

The question as to the validity of the Rytov expansion has become important recently, since both the Born and Rytov approximations have become tools for the theory of diffraction tomography.[6-8] Both approximations yield a linearized version of the reduced wave equation, which provide models that are used in the reconstruction process to obtain $n(x)$ from measured data.

In this paper, the Rytov approximation is examined in detail. In order to get sharper results and better insights the one-dimensional case will be treated first. This will be followed by a treatment of the much more complicated three-

 0022-2488/85/081979-07$02.50

dimensional case. Finally a comparison is given between the domains of validity of the Rytov and Born approximations.

II. ONE-DIMENSIONAL CASE

The one-dimensional case will be investigated in detail here since better estimates can be made for the conditions under which the Rytov approximation is the leading term of a convergent expansion.

The region D will be taken to be the interval $0<x<l$. The incident wave will be a plane wave propagating in the direction of the positive x axis,

$$u^i = \psi = e^{ikx}. \tag{10}$$

For $x \le 0$, the scattered wave is just the reflected wave

$$u^s = Re^{-ikx}, \quad x \leqslant 0. \tag{11}$$

For $x \ge l$, the total field is just the transmitted wave

$$u = Te^{ikx}.$$

In the one-dimensional case, Eq. (4) takes the form

$$\left(\frac{d^2}{dx^2} + k^2\right)\psi\phi = -\psi\left[\left(\frac{d\phi}{dx}\right)^2 + k^2(n^2-1)\right],$$
$$0<x<l, \tag{12}$$

From Eqs. (10) and (11) it can be shown that

$$\phi'(0)\exp\phi(0) = -2ikR, \quad \exp\phi(0) = 1 + R,$$

yielding the boundary condition at $x=0$,

$$\phi'(0) = 2ik(\exp[-\phi(0)] - 1).$$

In a similar manner it can be shown that

$$\phi'(l) = 0.$$

Using the Green's function

$$g(x,y) = (i/2k)\exp ik|x-y|, \tag{13}$$

Eq. (12) can be placed in the form

$$\phi(x) = \int_0^l g(x,y)e^{ik(y-x)}\left[\left(\frac{d\phi}{dy}\right)^2 + k^2(n^2(y)-1)\right]dy$$
$$+\phi(0) + e^{-\phi(0)} - 1. \tag{14}$$

Differentiate with respect to x and set

$$\frac{d\phi}{dx} = v(x). \tag{15}$$

The resulting equation is a quadratic integral equation involving the single unknown $v(x)$,

$$v(x) = +\int_x^l e^{2ik(y-x)}[v^2(y) + k^2(n^2(y)-1)]dy. \tag{16}$$

Note that $\phi(0)$ can be recovered from knowledge of $v(0)$ by

$$\phi(0) = -\ln[1 + v(0)/2ik]. \tag{17}$$

Equation (16) will now be investigated. It has the general form

$$v = F(v). \tag{18}$$

Define the norm

$$\|v\| = \operatorname*{Max}_{0<x<l}|v(x)|.$$

It is seen that for $v_1(x)$ and $v_2(x)$ in the ball $\|v\| \leqslant r$,

$$\|F(v_1) - F(v_2)\| \leqslant \Theta\,\|v_1 - v_2\|, \tag{19}$$

where

$$\Theta = 2lr. \tag{20}$$

Hence if $r < 1/2l$, the operator F is a contraction mapping[9,10] of the closed ball $\overline{U}(0,r)$. By known results[9,10] the method of successive approximations

$$v^{n+1} = F(v^n), \quad v^0 = 0 \tag{21}$$

will converge to a unique fixed point v_* in the ball $\overline{U}(0,r_0)$ if

$$r \geqslant [1/(1-\Theta)]\,\|F(0)\|. \tag{22}$$

Here

$$\|F(0)\| = \operatorname*{Max}_{0<x<l}\left|\int_x^l e^{2ik(y-x)}k^2(n^2(y)-1)dy\right|. \tag{23}$$

Combining Eqs. (20) and (22) it can be shown that for maximum results, r should be chosen so that $\Theta = \frac{1}{2}$, yielding the condition

$$\operatorname*{Max}_{0<x<l} k^2 l\left|\int_x^l e^{2iky}[n^2(y)-1]dy\right| < \frac{1}{8}. \tag{24}$$

Using the fact that the integral operator in Eq. (16) is Volterra one can get sharper results essentially reducing the numerical factor on the right-hand side of inequality (24). Briefly, it can be shown that the nth iterate of $F(v)$, namely $F^n(v)$, satisfies the relation

$$\|F^n(v_1) - F^n(v_2)\| \leqslant [(2rl)^n/n!]\,\|v_1 - v_2\|,$$

for v_1, v_2 in the ball $\overline{U}(0,r)$. From Theorem 12.5 of Rall,[9] it can be deduced that the method of successive approximations (21) converges to a unique fixed point in the ball $\overline{U}(0,r_0)$ if

$$2rl \geqslant e^{2rl}2l\,\|F(0)\| = 2r_0 l.$$

This is satisfied by taking $2rl = 1$, and $2l\,\|F(0)\| < 1/e$. This yields the condition

$$\operatorname*{Max}_{0<x<l} k^2 l\left|\int_x^l e^{2iky}[n^2(y)-1]dy\right| < \frac{1}{2e}. \tag{24'}$$

[Note, for comparison, Newton's iterative scheme starting from $v^0 = 0$ can be shown to converge for slightly less restrictive conditions, namely with the right-hand side of inequality (24) replaced by $\frac{1}{4}$.] However, of main interest is the form of the quantity on the left-hand side in equality (24) or (24').

The corresponding condition for the Born approximation is

$$\sup_u \frac{\|(k/2)\int_0^l \exp[ik(x_< - x_>)](n^2(y)-1)u(y)dy\|}{\|u\|} < 1, \tag{25a}$$

which yields the more useful sufficient (but *not* necessary) condition

$$kl\int_0^l \frac{|n^2-1|}{l}\,dy < 2. \tag{25b}$$

A comparison of the Rytov condition (24) and Born conditions (25a) and (25b) will be detailed in a later section.

Summing up, the expression for the complex phase term $\phi(x)$ corresponding to the nth iterate of Eq. (16) is

$$\phi^n(x) = \phi^0(x) + \int_0^l g(x,y)e^{ik(y-x)}[v^n(y)]^2\,dy$$
$$+\alpha^n(0), \tag{26}$$

where

$$\alpha^n(0) = -\ln[1 + v^n(0)/2ik] + v^n(0)/2ik,$$

$$\phi^0(x) = \int_0^l g(x,y)e^{ik(y-x)}k^2[n^2(y) - 1]dy.$$

It is interesting to note that the corresponding value of the reflection coefficient is given by

$$1 + R = -[1 + v^n(0)/2ik]^{-1}. \tag{27}$$

III. RYTOV APPROXIMATION IN R^3

The boundary S of the bounded connected open region D will be taken to be a smooth surface. It will be assumed that S is at least C^2, although this could be made less restrictive.[11]

The incident field ψ will be restricted by requiring that $|\psi(x)| > \text{const} > 0$ for all x in $\overline{D}$. Since the region $\overline{D}$ is assumed source-free, both $\psi(x)$ and $1/\psi(x)$ will be $C^\infty(D)$. Physically the restriction that $|\psi(x)|$ be bounded away from zero implies that there must be no nulls in the radiation pattern in the direction of the region $\overline{D}$, such as would be produced by a dipole or higher multiple type source.

For a point x in D, Eq. (4) can be transferred to the integral representation

$$\begin{aligned}\psi(x)\phi(x) &= \int_D G_0(x,y)[\nabla\phi\cdot\nabla\phi + k^2(n^2-1)]\psi(y)d\tau_y \\ &\quad + \int_S \Big[G_0(x,y)\frac{\partial}{\partial n_y}(\psi(y)\phi(y)) \\ &\quad - \psi(y)\phi(y)\frac{\partial}{\partial n_y}G_0(x,y)\Big]d\sigma_y,\end{aligned} \tag{28}$$

where $\partial/\partial n$ is the outward normal derivative and $d\sigma$ is the element of the surface area.

Equation (28) cannot be used by itself to solve for ϕ. Unlike the one-dimensional case the boundary terms will not disappear. The equation must be modified to take into account the behavior of ϕ in the region exterior to $\overline{D}$ and especially the behavior for $|x| \to \infty$. Here we use the fact that the scattered field $u^s(x)$, given by

$$u^s(x) = \psi(e^\phi - 1), \tag{29}$$

satisfies the Helmholtz equation in the region exterior to $\overline{D}$

$$\Delta u^s + k^2u^s = 0, \quad x \in R^3\backslash\overline{D},$$

and satisfies the radiation condition for $|x| \to \infty$. Using this result and Green's theorem for the exterior region one has

$$\begin{aligned}&\int_S \Big[G_0(x,y)\frac{\partial u^s(y)}{\partial n_y} - u^s(y)\frac{\partial G_0(x,y)}{\partial n_y}\Big]d\sigma_y \\ &= \begin{cases} -u^s(x), & \text{for } x \in R^3\backslash\overline{D}, \\ 0, & \text{for } x \in D. \end{cases}\end{aligned} \tag{30}$$

Thus combining Eqs. (28) and (30), and using relation (29) the following can be obtained for $x \in D$:

$$\begin{aligned}\phi(x) &= \frac{1}{\psi(x)}\Big\{\mathscr{K}([\nabla\phi\cdot\nabla\phi + k^2(n^2-1)]\psi) \\ &\quad + \mathscr{V}\Big(\frac{\partial}{\partial n}[\psi\mu(\phi)]\Big) - \mathscr{W}(\psi\mu(\phi))\Big\},\end{aligned} \tag{31}$$

where

$$\mu(\phi) = \phi + 1 - e^\phi. \tag{32}$$

The operators are defined as follows:

$$\mathscr{K}(u) = \int_D G_0(x,y)u(y)d\tau_y, \tag{33}$$

$$\mathscr{V}(v) = \int_S G_0(x,y)v(y)d\sigma_y, \tag{34}$$

$$\mathscr{W}(v) = \int_S \frac{\partial G_0(x,y)}{\partial n_y}\cdot v(y)d\sigma_y. \tag{35}$$

The operators $\mathscr{V}$ and $\mathscr{W}$ are, respectively, the generalized single- and double-layer potentials. The properties of all these operators will be examined later.

Equation (31) will be studied in detail. The advantage of Eq. (31) over Eq. (28) is that the boundary terms behave like ϕ^2 for small ϕ. Because the boundary terms are included, Eq. (31) will have to be considered for $x \in \overline{D}$, with the appropriate limits taken for the single- and double-layer potentials as the point x approaches a point on S from the interior of D.

Equation (31) has the general form for $x \in \overline{D}$,

$$\phi = F(\phi). \tag{36}$$

It will be shown later that the method of successive approximations

$$\phi^0 = 0, \quad \phi^{n+1} = F(\phi^n) \tag{37}$$

will converge to a unique solution under suitable restrictions on the quantity $(n^2 - 1)$.

It is interesting to note that the iterate ϕ^1 corresponds to the Rytov approximation,

$$\phi^1(x) = [1/\psi(x)]\mathscr{K}[k^2(n^2-1)\psi]. \tag{38}$$

The next correction to the Rytov approximation

$$\begin{aligned}\phi^2(x) &= \phi^1(x) + \frac{1}{\psi(x)}\Big\{\mathscr{K}(\psi\nabla\phi^1\cdot\nabla\phi^1) \\ &\quad + \mathscr{V}\Big(\frac{\partial}{\partial n}[\psi\mu(\phi^1)]\Big) \\ &\quad - \mathscr{W}(\psi\mu(\phi^1))\Big\}\end{aligned} \tag{39}$$

includes boundary terms.

In order to study the conditions for which the method of successive approximations (37) converges, some terminology involving Hölder continuous functions needs to be introduced.

First the uniform norm of a continuous function over a region T is defined by

$$\|u\|_T = \operatorname*{Max}_{x\in T}|u(x)|, \tag{40}$$

and the Hölder coefficients $U_{0\lambda}(T)$ and $U_{1\lambda}(T)$ are given by

$$\begin{aligned}&|u(x) - u(y)| \leqslant U_{0\lambda}(T)|x-y|^\lambda, \\ &\sup_{i=1,2,3}\left|\frac{\partial u(x)/\partial x_i - \partial u(y)/\partial y_i}{|x-y|^\lambda}\right| \leqslant U_{1\lambda}(T),\end{aligned} \tag{41}$$

where the inequalities hold for all $x, y \in T$. The Hölder exponent λ will be taken $0 < \lambda < 1$.

The Banach space $C^{1,\lambda}(\overline{D})$ of Hölder differentiable functions over $\overline{D}$ has norm[10]

$$\|u\|_{C^{1,\lambda}(\overline{D})} = \|u\|_{\overline{D}} + \sum_{i=1}^{3} \left\| \frac{\partial u}{\partial x_i} \right\|_{\overline{D}} + U_{1\lambda}(\overline{D}) \tag{42}$$

and the Banach space $C^{0,\lambda}(\overline{D})$ of Hölder continuous functions over $\overline{D}$ has norm

$$\|u\|_{C^{0,\lambda}(\overline{D})} = \|u\|_{\overline{D}} + U_{0\lambda}(\overline{D}). \tag{43}$$

Similar definitions are valid for the Hölder continuous functions over S and the Banach spaces $C^{1,\lambda}(S)$ and $C^{0,\lambda}(S)$.

The properties of the operators can now be stated. From Miranda,[11] the operator $\mathscr{K}$ maps $u \in C^{(0,\mu)}(D)$ into $C^2(D)$. However, the following weaker result[12] is all that is needed. If $u(x)$ is bounded and integrable on $\overline{D}$ then $\mathscr{K}u$ belongs to $C^{(1,\lambda)}(R^3)$, hence $C^{(1,\lambda)}(\overline{D})$. Because of the restrictions imposed on the incident field ψ, both ψ and $1/\psi$ will be at least $C^2(\overline{D})$, and an operator

$$\mathscr{K}_\psi u = (1/\psi)\mathscr{K}(\psi u) \tag{44}$$

can be defined which maps bounded $u(x)$ into $C^{1,\lambda}(\overline{D})$. Hence the following norm will be defined:

$$\|\mathscr{K}_\psi\| = \sup_u \|\mathscr{K}_\psi u\|_{C^{1,\lambda}(\overline{D})} / \|u\|_{\overline{D}} \,. \tag{45}$$

The operator $\mathscr{V}$ maps functions $v \in C^{0,\lambda}(S)$ into functions in $C^{1,\lambda}(D)$, which can be extended by continuity to $\overline{D}$ so that $\mathscr{V}v \in C^{1,\lambda}(\overline{D})$.[11–14] Define the operator

$$\mathscr{V}_\psi(\eta) = \frac{1}{\psi}\mathscr{V}\left(\frac{\partial}{\partial n}(\psi\eta)\right). \tag{46}$$

Since the surface is at least C^2, the direction cosines of the normal belong to $C^{(1,\lambda)}(S)$, hence since ψ is at least $C^2(S)$, the mapping

$$\frac{\partial}{\partial n}(\psi\eta) = v$$

maps $\eta \in C^{(1,\lambda)}(\overline{D})$ into $v \in C^{(0,\lambda)}(S)$. Thus the operator $\mathscr{V}_\psi$ maps $\eta \in C^{(1,\lambda)}(\overline{D})$ into $C^{1,\lambda}(\overline{D})$. Consequently the following norm can be introduced:

$$\|\mathscr{V}_\psi\| = \sup_\eta \|\mathscr{V}_\psi(\eta)\|_{C^{1,\lambda}(\overline{D})} / \|\eta\|_{C^{1,\lambda}(\overline{D})} \,. \tag{47}$$

From Refs. 11 and 13, the operator $\mathscr{W}$ maps $v \in C^{1,\lambda}(S)$ into $C^{1,\lambda}(D)$, which can be extended by continuity to $C^{1,\lambda}(\overline{D})$. The operator $\mathscr{W}_\psi$, defined by

$$\mathscr{W}_\psi(\eta) = (1/\psi)\mathscr{W}(\psi\eta), \tag{48}$$

has the same property. Hence the following norm can be given:

$$\|\mathscr{W}_\psi\| = \sup_\eta \|\mathscr{W}_\psi \eta\|_{C^{1,\lambda}(\overline{D})} / \|\eta\|_{C^{1,\lambda}(S)}. \tag{49}$$

The conditions for which Eq. (36) can be solved by successive approximations will now be examined. It must first be established that for some appropriate r, $F(\phi)$ is a contraction mapping for ϕ in a ball $U(0,r)$ (given explicitly by $\|\phi\|_{C^{1,\lambda}(\overline{D})} \leqslant r$).

It is seen that for ϕ_1, ϕ_2 in the ball $U(0,r)$,

$$\begin{aligned}
&\|F(\phi_1) - F(\phi_2)\|_{C^{1,\lambda}(\overline{D})} \\
&\quad \leqslant \|\mathscr{K}_\psi(\nabla\phi_1 \cdot \nabla\phi_1 - \nabla\phi_2 \cdot \nabla\phi_2)\| \\
&\qquad + \|\mathscr{V}_\psi(\mu(\phi_1) - \mu(\phi_2))\| \\
&\qquad + \|\mathscr{W}_\psi(\mu(\phi_1) - \mu(\phi_2))\| \\
&\quad \leqslant \|\mathscr{K}_\psi\| \, \|(\nabla\phi_1 \cdot \nabla\phi_1 - \nabla\phi_2 \cdot \nabla\phi_2)\|_{\overline{D}} \\
&\qquad + \|\mathscr{V}_\psi\| \, \|\mu(\phi_1) - \mu(\phi_2)\|_{C^{1,\lambda}(S)} \\
&\qquad + \|\mathscr{W}_\psi\| \, \|\mu(\phi_1) - \mu(\phi_2)\|_{C^{1,\lambda}(S)} \,.
\end{aligned}$$

It may be shown that

$$\|\nabla\phi_1 \cdot \nabla\phi_1 - \nabla\phi_2 \cdot \nabla\phi_2\|_D \leqslant 6r\|\phi_1 - \phi_2\|_{C^{1,\lambda}(\overline{D})} \,,$$

$$\begin{aligned}
&\|\mu(\phi_1) - \mu(\phi_2)\|_{C^{1,\lambda}(S)} \\
&\quad \leqslant [e^r - 1 + 2re^r + re^{3r}(1+r)] \|\phi_1 - \phi_2\|_{C^{1,\lambda}(\overline{D})} \,.
\end{aligned}$$

It thus follows that

$$\|F(\phi_1) - F(\phi_2)\|_{C^{1,\lambda}(\overline{D})} \leqslant \Theta \|\phi_1 - \phi_2\|_{C^{1,\lambda}(\overline{D})} \,, \tag{50}$$

where

$$\begin{aligned}
\Theta = {}& 6r\|\mathscr{K}_\psi\| + [e^r - 1 + 2re^r + re^{3r}(1+r)] \\
&\times [\|\mathscr{V}_\psi\| + \|\mathscr{W}_\psi\|].
\end{aligned} \tag{51}$$

Since for small r, Θ behaves as

$$\Theta \sim r\{6\|\mathscr{K}_\psi\| + 4\|\mathscr{V}_\phi\| + 4\|\mathscr{W}_\phi\|\}, \tag{52}$$

it is seen that r can be chosen so that $\Theta < 1$. For convenience r will be chosen so that

$$\Theta(r) = \tfrac{1}{2}. \tag{53}$$

This fixes the ball $U(0;r)$. From inequality (50) it is now seen that $F(\phi)$ is a contraction mapping of the closed ball $\overline{U}(0;r)$.

For the method of successive approximations [Eq. (37)] to converge to a fixed point in the ball, it is required that

$$\|F(0)\|_{C^{1,\lambda}(\overline{D})} \leqslant (1 - \Theta)r = \tfrac{1}{2}r.$$

This places the following restriction on $n^2 - 1$:

$$\|\mathscr{K}_\psi(k^2(n^2-1))\|_{C^{1,\lambda}(\overline{D})} \leqslant \tfrac{1}{2}r. \tag{54}$$

These results can be summarized as follows.

Theorem: If (i) the incident wave is generated by sources exterior to $\overline{D}$ and has no nulls in $\overline{D}$, (ii) r is chosen to satisfy Eqs. (51) and (53), and (iii) $n(x)$ is restricted by inequality (54), then the method of successive approximations starting from $\phi^0 = 0$, applied to Eq. (31) for $x \in \overline{D}$, converges to a unique solution in the ball $\overline{U}(O;r)$. Furthermore, the first iterate ϕ^1 given by Eq. (38) is the Rytov approximation.

The Rytov inequality (54) is a generalization of the one-dimensional inequality (24).

To further quantify the condition given by inequality (54) one would like to know how r [prescribed by Eq. (53)] depends upon the size of the domain D. This will require estimates for the norms of the operators $\mathscr{K}_\psi$, $\mathscr{V}_\psi$, and $\mathscr{W}_\psi$. Here, estimates for the norm of $\mathscr{K}_\psi$ only will be obtained. To make it more explicit consider the case in which the incident wave is the plane wave $u^i(x) = \exp(ikx_3)$, and set

$$h(x,y) = G_0(x,y)e^{ik(y_3 - x_3)}. \tag{55}$$

Now the operator $\mathscr{K}_\psi$ becomes

$$\mathscr{K}_\psi u = \int_D h(x,z)u(z)d\tau_z \,. \tag{56}$$

Using the inequalities

$$|G_0(x,z)| \leqslant (1/4\pi)(1/|x-z|),$$

$$\left|\frac{\partial G_0(x,z)}{\partial x_i}\right| \leqslant \frac{1}{4\pi}\left(\frac{k}{|x-z|} + \frac{1}{|x-z|^2}\right),$$

it is seen that

$$\|\mathscr{K}_\psi u\|_{\bar D} < \operatorname*{Max}_{x\in\bar D} \frac{1}{4\pi}\int_D \frac{1}{|x-z|}\,d\tau_z \|u\|_{\bar D},$$

$$\left|\left|\frac{\partial}{\partial x_i}\mathscr{K}_\psi u\right|\right|_{\bar D}$$

$$< \operatorname*{Max}_{x\in\bar D} \frac{1}{4\pi}\int_D \left(\frac{k(1+\delta_{i3})}{|x-z|} + \frac{1}{|x-z|^2}\right) d\tau_z \|u\|_{\bar D},$$

where δ_{i3} is the Kronecker delta.

Now use the following Lemma (the proof of which is omitted).

Lemma: If l is the radius of the smallest sphere enclosing $\bar D$, then for $0<\mu<3$,

$$\operatorname*{Max}_{x\in\bar D} \frac{1}{4\pi}\int_D \frac{1}{|x-z|^\mu}\,d\tau_z < \frac{l^{3-\mu}}{(3-\mu)}. \tag{57}$$

It follows that

$$\|\mathscr{K}_\psi u\|_{\bar D} < (l^2/2)\|u\|_{\bar D},$$

$$\left|\left|\frac{\partial \mathscr{K}_\psi u}{\partial x_i}\right|\right|_{\bar D} < \left[\frac{kl^2}{2}(1+\delta_{i3}) + l\right]\|u\|_{\bar D}.$$

From Eq. (A13) it is seen that

$$\sup_{i=1,2,3}\left|\frac{\partial}{\partial x_i}\mathscr{K}_\psi u - \frac{\partial}{\partial y_i}\mathscr{K}_\psi u\right| < H|x-y|^\lambda,$$

where

$$H = 4m_\lambda k^\lambda[kl^2 + l] + (3+2^{1-\lambda}) \times\left(\frac{kl^{2-\lambda}}{2-\lambda} + \frac{l^{1-\lambda}}{1-\lambda}\right) \tag{58}$$

and m_λ is defined by Eq. (A2).

Combining these results, and using the definitions given by Eqs. (42) and (45), one can obtain

$$\|\mathscr{K}_\psi\| = l^2/2 + (3l + 2kl^2) + H.$$

From Eqs. (52) and (53) it is seen thar r is inversely proportional to terms which are at least the order of l^2, kl^2. Hence, as the size of the domain D becomes larger and larger, condition (54) becomes more restrictive. This implies that the iterative process applied to an unbounded domain $D = R^3$ does not converge.

IV. COMPARISON OF BORN AND RYTOV DOMAINS OF CONVERGENCE

There are two main quantities of interest that one would want to consider in comparing the Born and Rytov expansions. They are (i) the domain of convergence [values of the parameters $n^2(x)$ and k for which the series converges] and (ii) the degree of approximation (the number of terms required in the expansion to obtain a desired accuracy). For the latter quantity the series need only be asymptotic if one wants a good approximation given by only one or two terms in the expansion. Here only the domains of convergence will be compared since no one else has proved convergence of the Rytov expansion. But before doing so, a few comments will be made about the second quantity of interest.

Keller[5] in a formal asymptotic analysis has indicated that the Rytov expansion tends to give a better approximation than the Born. Hadden and Mintzer[15] in an extensive analysis compared the first two iterates of the Born and Rytov expressions to the exact solution for the Epstein problem and found, in general, not too much difference between the two sets of approximations, although in one case they found that the Rytov is superior. [However, to obtain the domain of convergence of the respective series, one has to consider all the terms in the series, and not just for one particular choice of $n(x)$ but for all possible choices.]

The comparison of the domains of convergence will be restricted to the one-dimensional case for which the results are more precise. The condition for convergence of the Born series is given by the operator norm condition (25a). For comparison of domains it is easier to use the more stringent condition (25b) given by

$$k\int_0^l |n^2(y) - 1|\,dy < 2.$$

A sufficient condition indicating convergence of the Rytov series is the function norm inequality given by (24′),

$$\operatorname*{Max}_{0<x<l}\left|k^2 l\int_x^l e^{2iky}[n^2(y)-1]\,dy\right| < \frac{1}{2e}.$$

It should be remembered that this condition guarantees convergence of the Rytov series if it is satisfied. However, because the numerical factor on the right-hand side is a lower bound estimate, the series may still converge if the left-hand side is slightly larger than $1/2e$.

Apart from the numerical factors, the important difference between the two expansions is that the absolute sign is outside one integral and inside the other. This implies that a big difference in the respective domains will occur for a rapidly fluctuating medium, i.e., when $n(x)$ rapidly fluctuates about unity in a distance of a wavelength. This is easily seen by taking the simple example where $n^2(x) - 1 = \delta\sin(\pi px/l)$, with p being a large positive integer, such that $\pi p \gg kl$. The Born inequality yields the condition $kl\delta < \pi$. [Note that it can be shown that the more precise operator norm inequality (25a) applied to this case, yields the condition $kl\delta < 4$.] Since

$$\left|\int_x^l e^{i2ky}\sin\left(\frac{\pi py}{l}\right)dy\right| < \frac{2l}{\pi p}\left[1 + O\left(\frac{kl}{p\pi}\right)\right],$$

the Rytov inequality yields

$$4e(k\delta l) < (\pi p/kl)[1 + O(kl/p\pi)].$$

Since $\pi p \gg kl$, it is seen that the Rytov condition is less stringent than the Born.

For the opposite case, where $n(x)$ is a slowly varying (differentiable) function in comparison to $\exp(2ikx)$, it can be shown that the Rytov inequality (24′) is similar to the Born inequality (25b), but is slightly more restrictive.

The parameter domain, where the second iterate of the Rytov expansion is much, much less than the first term, has been investigated by various authors, among them Keller[4] and Yura *et al.*[16] Although this parameter domain is *not* the domain of convergence, it is important. In many applications, such as diffraction tomography where one desires a linear relationship between phase and the quantity $(n^2(x-1))$, one requires that not only the second term be small compared to the first, but also that the contribution from the remaining terms be negligible too.

ACKNOWLEDGMENT

The research was supported under ONR Contract No. N00014-83-K-6038.

APPENDIX: ESTIMATE OF THE HÖLDER COEFFICIENT FOR $(\partial/\partial x_i)\mathscr{K}_\psi u$

This Appendix presents an analysis for estimating the Hölder coefficient of $(\partial/\partial x_i)\mathscr{K}_\psi u$, where the operator $\mathscr{K}_\psi$ has kernel $h(x,z)$ given by Eq. (55). As a preliminary the following quantities will be defined:

$$r_1 = |x-z|, \quad r_2 = |y-z|,$$
$$r_< = \mathrm{Min}[r_1,r_2], \quad r_> = \mathrm{Max}[r_1,r_2],$$
$$G_0(x,z) = G_0(r_1),$$
$$G_0'(r) = (e^{ikr}/4\pi r)(ik - 1/r),$$

From this it follows that

$$\frac{\partial}{\partial x_i} G_0(x,z) = \frac{G_0'(r_1)(x_i - z_i)}{r_1}. \tag{A1}$$

Furthermore, define

$$m_\lambda = \underset{\alpha>0}{\mathrm{Max}}(\sin\alpha/2^\lambda\alpha^\lambda), \quad 0<\lambda<1. \tag{A2}$$

The following inequalities will also be employed:

$$|r_1 - r_2| \leqslant |x-y|, \quad |r_1 - r_2| \leqslant r_> .$$

From these, the additional inequalities

$$\frac{|r_1-r_2|}{(r_1r_2)} \leqslant \frac{|x-y|^\lambda r_>^{1-\lambda}}{(r_1r_2)} \leqslant \frac{|x-y|^\lambda}{r_<^{1+\lambda}}, \tag{A3}$$

$$\begin{aligned}|e^{ikr_1} - e^{ikr_2}| &\leqslant 2|\sin(k/2)(r_1-r_2)|\\ &\leqslant 2(|\sin\alpha|/|\alpha|^\lambda)(k/2)^\lambda|x-y|^\lambda\\ &\leqslant 2m_\lambda k^\lambda|x-y|^\lambda\end{aligned} \tag{A4}$$

can be deduced, and similarly

$$\begin{aligned}|e^{ik(z_3-x_3)} - e^{ik(z_3-y_3)}| &\leqslant 2|\sin(k/2)(x_3-y_3)|\\ &\leqslant 2m_\lambda k^\lambda|x-y|^\lambda.\end{aligned} \tag{A5}$$

In order to get the Hölder coefficient of composite functions the following general inequality will be employed:

$$\begin{aligned}&|a(x)b(x) - a(y)b(y)|\\ &\quad\leqslant |a(x)-a(y)|\,|b(x)| + |b(x)-b(y)|\,|a(y)|.\end{aligned} \tag{A6}$$

Using the above inequality one can obtain, using relations (A3) and (A4),

$$\begin{aligned}|G_0(r_1) - G_0(r_2)| &\leqslant \frac{1}{4\pi r_1}|e^{ikr_1} - e^{ikr_2}| + \frac{1}{4\pi}\left|\frac{1}{r_1} - \frac{1}{r_2}\right|\\ &\leqslant (|x-y|^\lambda/4\pi)[2m_\lambda k^\lambda/r_1 + r_<^{-1-\lambda}].\end{aligned} \tag{A7}$$

Similarly

$$\begin{aligned}|G_0'(r_1) - G_0'(r_2)| &\leqslant \frac{1}{4\pi}|e^{ikr_1} - e^{ikr_2}|\left|\frac{ik}{r_1} - \frac{1}{r_1^2}\right|\\ &\quad + \frac{1}{4\pi}\left|\frac{1}{r_1} - \frac{1}{r_2}\right|\left|ik - \frac{1}{r_1} - \frac{1}{r_2}\right|\\ &\leqslant \frac{1}{2\pi}|x-y|^\lambda m_\lambda k^\lambda\left|\frac{k}{r_1} + \frac{1}{r_1^2}\right|\\ &\quad + \frac{|x-y|^\lambda}{4\pi}\left|k + \frac{1}{r_1} + \frac{1}{r_2}\right| r_<^{-1-\lambda}\\ &\leqslant \frac{|x-y|}{4\pi}\left[2m_\lambda k^\lambda\left(\frac{k}{r_1} + \frac{1}{r_1^2}\right)\right.\\ &\quad\left. + \frac{k}{r_<^{1+\lambda}} + \frac{2}{r_<^{2+\lambda}}\right].\end{aligned} \tag{A8}$$

The inequality

$$\begin{aligned}&\left|\frac{(x_i-z_i)}{r_1} - \frac{(y_i-z_i)}{r_2}\right|\\ &\quad\leqslant \frac{|(x_i-z_i)-(y_i-z_i)|}{r_1} + |x_i - z_i|\left|\frac{1}{r_1} - \frac{1}{r_2}\right|\\ &\quad\leqslant \frac{|x_i-y_i|^\lambda}{r_1}(|x_i-z_i| + |y_i-z_i|)^{1-\lambda}\\ &\qquad + |x_i-z_i|\frac{|x-y|^\lambda}{r_<^{1+\lambda}}\\ &\quad\leqslant |x-y|^\lambda r_>[2^{1-\lambda}+1]\,r_<^{-1-\lambda}\end{aligned} \tag{A9}$$

can be obtained where the relations $|x_i - y_i| \leqslant |x-y|$, $|x_i - z_i| \leqslant r_>$, and $|y_i - z_i| \leqslant r_>$ are employed. Inequality (A6) will be applied to the composite function (A1), setting $b(x) = G_0'(r_1)$, $|a(y)| = |y_i - z_i|/r_2 \leqslant 1$, when $r_1 \geqslant r_2$, and $a(y) = G_0'(r_2)$, $|b(x)| = |x_i - z_i|/r_1 \leqslant 1$, when $r_2 \geqslant r_1$. This results in

$$\begin{aligned}\left|\frac{\partial G_0(r_1)}{\partial x_i} - \frac{\partial G_0(r_2)}{\partial y_i}\right| &\leqslant |G_0'(r_>)|\left|\frac{(x_i-z_i)}{r_1} - \frac{(y_i-z_i)}{r_2}\right|\\ &\quad + |G_0'(r_1) - G_0'(r_2)|\\ &\leqslant \frac{|x-y|^\lambda}{4\pi}\left\{2m_\lambda k^\lambda\left(\frac{k}{r_<} + \frac{1}{r_<^2}\right)\right.\\ &\quad\left. + \frac{k(2+2^{1-\lambda})}{r_<^{1+\lambda}} + \frac{(3+2^{1-\lambda})}{r_<^{2+\lambda}}\right\},\end{aligned} \tag{A10}$$

where use is made of the inequalities $(r_1)^{-1} \leqslant (r_<)^{-1}$ and $(r_>)^{-1} \leqslant (r_<)^{-1}$.

Finally the Hölder coefficient of the derivatives of the kernel $(\partial/\partial x_i)h(x,z)$ may be obtained by applying inequality (A6) to the differentiated product of terms yielding

$$\begin{aligned}\left|\frac{\partial}{\partial x_i}h(x,z) - \frac{\partial}{\partial y_i}h(y,z)\right| &\leqslant k\delta_{i3}|G_0(r_1) - G_0(r_2)|\\ &\quad + \left|\frac{\partial G_0(r_1)}{\partial x_i} - \frac{\partial G_0(r_2)}{\partial y_i}\right|\\ &\quad + |e^{ik(z_3-x_3)} - e^{ik(z_3-y_3)}|\\ &\quad \times\left(k\delta_{i3}|G_0(r_1)| + \left|\frac{\partial G_0(r_1)}{\partial x_i}\right|\right),\end{aligned} \tag{A11}$$

where δ_{i3} is the Kronecker delta. The right-hand side of Eq. (A11) may be reduced using previous results (A5), (A7), and (A10) to give

$$\left|\frac{\partial}{\partial x_i}h(x,z) - \frac{\partial}{\partial y_i}h(y,z)\right| \leqslant \frac{|x-y|^\lambda}{4\pi}E,$$

with

$$E = 4m_\lambda k^\lambda \left[\frac{(1+\delta_{i3})k}{r_<} + \frac{1}{r_<^2} \right] + \frac{k\,[2+\delta_{i3}+2^{1-\lambda}]}{r_<^{1+\lambda}} + \frac{(3+2^{1-\lambda})}{r_<^{2+\lambda}}. \tag{A12}$$

Finally, using the estimates for the integrals given by Eq. (57), the following result may be deduced:

$$\left| \frac{\partial}{\partial x_i} \mathscr{K}_\psi u - \frac{\partial}{\partial y_i} \mathscr{K}_\psi u \right| < \int_D \left| \frac{\partial}{\partial x_i} h(x,z) - \frac{\partial}{\partial y_i} h(y,z) \right| |u(z)| d\tau_z < (x-y)^\lambda H_i \|u\|_{\bar{D}}, \tag{A13}$$

where

$$H_i = 4m_\lambda k^\lambda [(kl^2/2)(1+\delta_{i3}) + l] + (2+\delta_{i3}+2^{1-\lambda}) \frac{kl^{2-\lambda}}{(2-\lambda)} + \frac{(3+2^{1-\lambda})l^{1-\lambda}}{(1-\lambda)}.$$

[1]L. A. Chernov, *Wave Propagation in a Random Medium* (McGraw–Hill, New York, 1960).
[2]V. I. Tatarski, *Wave Propagation in a Turbulent Medium* (McGraw–Hill, New York, 1961).
[3]A. Ishimaru, *Wave Propagation and Scattering in Random Media* (Academic, New York, 1978).
[4]W. P. Brown, "Validity of the Rytov approximation in optical propagation calculations," J. Opt. Soc. Am. **56**, 1045 (1966).
[5]J. B. Keller, "Accuracy and validity of the Born and Rytov approximations," J. Opt. Soc. Am. **59**, 1003 (1969).
[6]A. J. Devaney, "Inverse-scattering theory within the Rytov approximation," Opt. Lett. **6**, 374 (1981).
[7]K. Iwata and R. Nagata, Jpn. J. Appl. Phys. **14**, Suppl. 14-1, 379 (1975).
[8]M. Slaney and A. C. Kak, "Diffraction tomography," Proc. SPIE **413**, 2 (1983).
[9]L. Rall, *Computational Solution of Non-Linear Operator Equations* (Krieger, Huntington, NY, 1979).
[10]M. S. Berger, *Nonlinearity and Functional Analysis* (Academic, New York, 1977).
[11]C. Miranda, *Partial Differential Equations of Elliptic Type* (Springer, New York, 1970).
[12]N. M. Günter, *Potential Theory* (Ungar, New York, 1967).
[13]G. Giraud, "Sur certains problèmes non linéares de Neumann and sur certains problemes non lineares mixtes," Ann. Ec. N. Sup. **49**, 1 (1932).
[14]W. Pogorzelski, "Propriétés des dérivées tangentielles d'une integral de l'équation elliptique," Ann. Polon. Math. **7**, 321 (1960).
[15]W. J. Hadden and D. Mintzer, "Test of the Born and Rytov approximations using the Epstein problem," J. Acoust. Soc. Am. **63**, 1279 (1978).
[16]H. T. Yura, C. C. Sung, S. F. Clifford, and R. J. Hill, "Second-order Rytov approximation," J. Opt. Soc. Am. **73**, 500 (1983).

Bilinear phase-plane distribution functions and positivity

A. J. E. M. Janssen
Philips Research Laboratories, P. O. Box 80.000, 5600 JA Eindhoven, The Netherlands

(Received 17 January 1985; accepted for publication 5 April 1985)

There is a theorem of Wigner that states that phase-plane distribution functions involving the state bilinearly and having correct marginals must take negative values for certain states. The purpose of this paper is to support the statement that these phase-plane distribution functions are for hardly any state everywhere non-negative. In particular, it is shown that for certain generalized Wigner distribution functions there are no smooth states (except the Gaussians for the Wigner distribution function itself) whose distribution function is everywhere non-negative. This class of Wigner-type distribution functions contains the Margenau–Hill distribution. Furthermore, the argument used in the proof of Wigner's theorem is augmented to show that under mild conditions one can find for any two states f, g with non-negative distribution functions a linear combination h of f and g whose distribution function takes negative values, unless f and g are proportional.

I. INTRODUCTION AND PRELIMINARIES

The formulation of quantum mechanics by means of phase-plane distribution functions involving the states bilinearly allows one to exhibit quantum mechanical expectation values as averages over the phase-plane of classical observables. Given a bilinear map[1] $(f,g)\to C_{f,g}$ mapping pairs of states onto functions of position q and momentum p, one can formulate a correspondence principle between bounded self-adjoint linear operators T of $L^2(\mathbb{R})$ and functions $a(q,p)$ as follows: T and a are said to correspond to each other when

$$(Tf,g) = \iint a(q,p)C_{f,g}(q,p)dq\,dp\,, \qquad (1)$$

for all $f\in L^2(\mathbb{R}), g\in L^2(\mathbb{R})$. Here $(\ ,\ )$ denotes the usual inner product in $L^2(\mathbb{R})$. Of course, in order that to any T (or a) there is a unique $a=a_T$ (or $T=T_a$) such that (1) holds, the mapping $(f,g)\to C_{f,g}$ should satisfy certain properties. When one takes $f=g$ in (1) the left-hand side equals the expectation of T in the state f while the right-hand side equals an average of the classical observable a corresponding to T, where $C_{f,f}$ is used as the weight function.

In view of the interpretation of $C_{f,f}$ as a distribution function, one would like the following properties to be satisfied: (a) correct marginals, i.e., for all states f one has

$$\int C_{f,f}(q,p)dp = |f(q)|^2, \quad q\in\mathbb{R}\,, \qquad (2)$$

$$\int C_{f,f}(q,p)dq = |F(p)|^2, \quad p\in\mathbb{R}\,, \qquad (3)$$

where F is the Fourier transform[2] of f, given by

$$F(p) = \int e^{-2\pi iqp} f(q)dq, \quad p\in\mathbb{R}\,; \qquad (4)$$

(b) positivity, i.e., for all states f one has

$$C_{f,f}(q,p)\geqslant 0, \quad q\in\mathbb{R}, \quad p\in\mathbb{R}\,. \qquad (5)$$

It has been shown by Wigner[3] that the requirements (a) and (b) (together with the bilinearity) are incompatible. What one can distill from the arguments in his proofs is this: when f_1 and f_2 are two compactly supported states with $f_1\not\equiv 0$, $f_2\not\equiv 0$, $f_1(q)f_2(q)=0$ for all q while the requirements (a) and (b) are met, then $C_{f,f}(q,p)$ takes negative values, where $f=f_1+f_2$. Hence, there is an abundance of states whose distribution functions take negative values. A particular strong result of this type, known as Hudson's theorem,[4] holds for the Wigner distribution[5] [See (11) below with $\alpha=0$]: the only square-integrable states for which the Wigner distribution is everywhere non-negative are the Gaussians.

It is the purpose of this paper to give generalizations on both Wigner's theorem and Hudson's theorem. The starting point we take in this paper differs slightly from the one taken by Wigner in Ref. 3. Wigner assumes the existence of self-adjoint operators $M(q,p)$ of $L^2(\mathbb{R})$ such that for all states f,

$$C_{f,f}(q,p) = (f,M(q,p)f), \quad q\in\mathbb{R}, \quad p\in\mathbb{R}\,. \qquad (6)$$

We consider in this paper Cohen's class of phase-plane distribution functions.[6–8] This class is parametrized by means of a function Φ of two variables as follows: for any state f one defines

$$C^{(\Phi)}_{f,f}(q,p) = \iiint \exp(-2\pi i(\theta q+\tau p-\theta u))$$
$$\times\Phi(\theta,\tau)f(u+\tfrac12\tau)\overline{f(u-\tfrac12\tau)}d\theta\,d\tau\,du\,,$$
$$q\in\mathbb{R}, \quad p\in\mathbb{R}\,. \qquad (7)$$

The distributions in Cohen's class all have the shift properties

$$\begin{aligned} C^{(\Phi)}_{T_af,T_af}(q,p) &= C^{(\Phi)}_{f,f}(q+a,p), \quad q\in\mathbb{R}, \quad p\in\mathbb{R}\,,\\ C^{(\Phi)}_{R_bf,R_bf}(q,p) &= C^{(\Phi)}_{f,f}(q,p+b), \quad q\in\mathbb{R}, \quad p\in\mathbb{R}\,, \end{aligned} \qquad (8)$$

where for all $a\in\mathbb{R}$, $b\in\mathbb{R}$ and all states $f\in L^2(\mathbb{R})$,

$$(T_af)(q) = f(q+a), (R_bf)(q) = e^{-2\pi ibq}f(q), \quad q\in\mathbb{R}\,. \qquad (9)$$

In this paper we pay particular attention to the choice

$$\Phi_\alpha(\theta,\tau) = \exp(2\pi i\alpha\theta\tau), \quad \theta\in\mathbb{R}, \quad \tau\in\mathbb{R}\,, \qquad (10)$$

where $\alpha\in\mathbb{R}$. This yields what may be called generalized Wigner distributions $C^{(\alpha)}_{f,f}$ that can be brought into the form

 0022-2488/85/081986-09$02.50

$$C_{f,f}^{(\alpha)}(q,p) \equiv C_{f,f}^{(\Phi_\alpha)}(q,p)$$
$$= \int e^{-2\pi i p t} f\left(q + \left(\frac{1}{2} - \alpha\right)t\right) \times \overline{f\left(q - \left(\frac{1}{2} + \alpha\right)t\right)} dt. \quad (11)$$

The choice $\alpha = 0$ yields the Wigner distribution $W_{f,f}$, and the choice $\alpha = -\frac{1}{2}$ yields what is known in signal analysis as Rihaczek's distribution[9] $R_{f,f}$. The latter distribution can be written as

$$R_{f,f}(q,p) = e^{2\pi i q p}\, \overline{f(q)} F(p), \quad q \in \mathbb{R}, \quad p \in \mathbb{R}, \quad (12)$$

with F given in (4). The real part of Rihaczek's distribution was considered by Margenau and Hill.[10] We observe here that $R_{f,f}(q,p)$ cannot be brought in the form (6) with a bounded linear operator $M(q,p)$ of $L^2(\mathbb{R})$.

As is well known[11] every $C_{f,f}$ can be expressed in terms of $W_{f,f}$ as

$$C_{f,f}^{(\varphi)}(q,p) = \iint \varphi(q-a, p-b) W_{f,f}(a,b)\, da\, db, \quad q \in \mathbb{R}, \quad p \in \mathbb{R}, \quad (13)$$

with

$$\varphi(q,p) = \iint e^{-2\pi i(\theta q + \tau p)} \Phi(\theta,\tau)\, d\theta\, d\tau, \quad q \in \mathbb{R}, \quad p \in \mathbb{R}. \quad (14)$$

As an aside we note that this φ can be used to relate the operators $M(q,p)$ of Wigner's approach in (6) and the function Φ in Cohen's approach. To that end we assume that Φ is such that $|C_{f,f}^{(\varphi)}(0,0)| \leqslant M \|f\|^2$ for some $M > 0$. Then the operator $M(0,0)$, determined by

$$(f, M(0,0)g) = C_{f,g}^{(\varphi)}(0,0), \quad f,g \in L^2(\mathbb{R}), \quad (15)$$

is bounded, and we have, according to (13),

$$(f, M(0,0)f) = \iint \varphi(-a,-b) W_{f,f}(a,b)\, da\, db, \quad f \in L^2(\mathbb{R}). \quad (16)$$

Hence, $\varphi(-a,-b)$ is what is called the Weyl symbol[12] of the operator $M(0,0)$, i.e., $\varphi(-a,-b)$ and $M(0,0)$ correspond to each other as in (1) when we take $C_{f,g} = W_{f,g}$ (Weyl correspondence[13]). Observe that, because of the shift properties, we have

$$M(q,p) = R_{-p}\, T_{-q}\, M(0,0) T_q\, R_p, \quad q \in \mathbb{R}, \quad p \in \mathbb{R}. \quad (17)$$

We shall restrict ourselves in this paper to functions Φ that are bounded. The boundedness of Φ ensures that $C_{f,f}^{(\Phi)} \in L^2(\mathbb{R}^2)$ for all $f \in L^2(\mathbb{R}^2)$. In fact we have the estimate[14]

$$\iint |C_{f,f}^{(\varphi)}(q,p)|^2\, dq\, dp \leqslant \|f\|^4 \sup |\Phi|^2, \quad (18)$$

for all $f \in L_2(\mathbb{R})$. It does not follow from boundedness of Φ that $C_{f,f}^{(\varphi)}(0,0)$ can be expressed as in (15) with the aid of a bounded operator $M(0,0)$. As a counterexample we have $R_{f,f}(0,0)$ in (12). However, when (15) does hold with a bounded $M(0,0)$ it can be shown from (16) and the fact that $C_{f,f}^{(\varphi)}(q,p) = (f, M(q,p)f)$ that all $C_{f,f}^{(\varphi)}$'s are bounded, continuous functions[15] of (q,p).

A second restriction is that we want real $C_{f,f}^{(\Phi)}$'s only. This is guaranteed when the φ in (14) is real valued[16] (possibly as a generalized function). A third restriction is that we want the $C_{f,f}^{(\Phi)}$'s to have correct marginals. It is well known[17] that $C_{f,f}^{(\Phi)}$ has correct marginals for all $f \in L^2(\mathbb{R})$ if and only if

$$\Phi(0,\tau) = \Phi(\theta,0) = 1, \quad \theta \in \mathbb{R}, \quad \tau \in \mathbb{R}. \quad (19)$$

We now summarize the results of this paper. It is shown in Sec. II, under a regularity condition on Φ, that for any two smooth functions f, g with $C_{f,f}^{(\Phi)} > 0$, $C_{g,g}^{(\Phi)} > 0$ everywhere we can find an $a \in \mathbb{C}$ such that $C_{f+ag,f+ag}^{(\Phi)}$ takes negative values, unless f and g are proportional. In Sec. III we consider smooth states f for which $\operatorname{Re} C_{f,f}^{(\alpha)}(q,p) \geqslant 0$ in (q,p) sets of the form $(a,b) \times \mathbb{R}$. It is shown, for example, that for such an f we have $|f(q)| = \exp(\psi(q))$ with ψ concave on (a,b) [it is assumed here that $f(q) \neq 0$ for $q \in (a,b)$]. We also show that, if $\alpha \neq 0$, there is no smooth state $f \not\equiv 0$ such that $\operatorname{Re} C_{f,f}^{(\alpha)}(q,p) \geqslant 0$ for all $q \in \mathbb{R}, p \in \mathbb{R}$. The restriction to smooth states is not entirely necessary but makes the proofs run smoothly; a class of functions which are sufficiently smooth is the set $\mathscr{S}$ of Schwartz [it is, however, sufficient to require the states to be sufficiently often differentiable and to decay as rapidly as $(1+q^2)^{-k}$ with k sufficiently large]. For $|\alpha| = \frac{1}{2}$ we have a stronger result, viz. there is no $f \in L^2(\mathbb{R})$, $f \not\equiv 0$ such that $\operatorname{Re} R_{f,f} \geqslant 0$ almost everywhere [see (12)]. A remarkable phenomenon here is that there do exist generalized functions $f \not\equiv 0$ with $\operatorname{Re} R_{f,f} \geqslant 0$ (in generalized sense). This is remarkable since the sets of smooth and generalized functions f for which $W_{f,f} = C_{f,f}^{(0)} \geqslant 0$ everywhere are essentially the same, i.e., (degenerate) Gaussians. We conjecture stronger results than the ones proved in this paper. This is based on the fact that we have not been able to find any Φ (other than $\Phi \equiv 1$, Wigner distribution case) and any $f \in L^2(\mathbb{R})$ (other than Gaussians) for which $C_{f,f}^{(\Phi)} \geqslant 0$ everywhere.

II. A RESULT FOR GENERAL BILINEAR PHASE-PLANE DISTRIBUTION FUNCTIONS

We consider in this section bounded Φ's for which $C_{f,f}^{(\Phi)}$ is real valued and has correct marginals for all $f \in L^2(\mathbb{R})$. The arguments used in the proof of the theorem below are somewhat similar to those used by Wigner in Ref. 1.

Theorem 1: Assume that the set $\{(\theta,\tau) | \Phi(\theta,\tau) \neq 0\}$ is dense in $\mathbb{R}^2$, and let $f \not\equiv 0$, $g \not\equiv 0$ be smooth states such that $C_{f,f}^{(\Phi)} > 0$, $C_{g,g}^{(\Phi)} > 0$ everywhere. Then there is an $a \in \mathbb{C}$, $b \in \mathbb{C}$ such that $C_{af+bg,af+bg}^{(\Phi)}$ takes negative values, unless f and g are proportional.

Proof: Suppose that

$$C_{af+bg,af+bg}^{(\Phi)}(q,p) \geqslant 0, \quad q \in \mathbb{R}, \quad p \in \mathbb{R}, \quad a \in \mathbb{C}, \quad b \in \mathbb{C}. \quad (20)$$

We shall show that f and g are proportional by using the following steps.

(1) We show that

$$|f(q)|^2 C_{g,g}^{(\Phi)}(q,p) = |g(q)|^2 C_{f,f}^{(\Phi)}(q,p), \quad q \in \mathbb{R}, \quad p \in \mathbb{R}. \quad (21)$$

Let $q \in \mathbb{R}$, and take $a_0 \in \mathbb{C}$, $b_0 \in \mathbb{C}$ such that $a_0 f(q) + b_0 g(q) = 0$. Then

$$C_{a_0 f + b_0 g, a_0 f + b_0 g}^{(\Phi)}(q,p) = 0, \quad p \in \mathbb{R}, \quad (22)$$

because of (20) and the correct marginals condition (2) applied to $a_0 f + b_0 g$. Hence we have for all $a \in \mathbb{C}$, $b \in \mathbb{C}$, $p \in \mathbb{R}$ (see Ref. 18),

$$|a|^2C^{(\Phi)}_{f,f}(q,p)+|b|^2C^{(\Phi)}_{g,g}(q,p)+2\,\mathrm{Re}\,a\bar{b}C^{(\Phi)}_{f,g}(q,p)\geqslant 0\,, \tag{23}$$

with equality when $a=a_0$, $b=b_0$. Let $p\in\mathbb{R}$, $a_0=-g(q)/f(q)$, $b=1$, and write $A=C^{(\Phi)}_{f,f}(q,p)$, $B=C^{(\Phi)}_{g,g}(q,p)$, $C=C^{(\Phi)}_{f,g}(q,p)$, $a=x+iy$, $a_0=x_0+iy_0$. Now (23) can be written as

$$P(x,y):=A(x^2+y^2)+B+2x\,\mathrm{Re}\,C-2y\,\mathrm{Im}\,C\geqslant 0\,, \tag{24}$$

with equality when $x=x_0$, $y=y_0$. Since for all $x\in\mathbb{R}$, $y\in\mathbb{R}$,

$$P(x,y)=A(x+A^{-1}\,\mathrm{Re}\,C)^2+A(y-A^{-1}\,\mathrm{Im}\,C)^2+B-A^{-1}|C|^2\geqslant 0\,, \tag{25}$$

and $P(x_0,y_0)=0$ it follows that $AB=|C|^2$, $C=-A(x_0-iy_0)$, i.e.,

$$|C^{(\Phi)}_{f,g}(q,p)|^2=C^{(\Phi)}_{f,f}(q,p)C^{(\Phi)}_{g,g}(q,p)\,,$$
$$C^{(\Phi)}_{f,g}(q,p)=\frac{\overline{g(q)}}{\overline{f(q)}}\,C^{(\Phi)}_{f,f}(q,p)\,. \tag{26}$$

Now elimination of $C_{f,g}(q,p)$ gives (21).

(2) We have

$$|F(p)|^2C^{(\Phi)}_{g,g}(q,p)=|G(p)|^2C^{(\Phi)}_{f,f}(q,p),\quad q\in\mathbb{R},\quad p\in\mathbb{R}\,. \tag{27}$$

This is proved in a similar way as (21).

(3) There is a constant $D>0$ such that

$$C^{(\Phi)}_{g,g}(q,p)=DC^{(\Phi)}_{f,f}(q,p),\quad q\in\mathbb{R},\quad p\in\mathbb{R}\,. \tag{28}$$

Indeed from (21) and (27) we get

$$\frac{C^{(\Phi)}_{g,g}(q,p)}{C^{(\Phi)}_{f,f}(q,p)}=\frac{|G(p)|^2}{|F(p)|^2}=\frac{|g(q)|^2}{|f(q)|^2},\quad q\in\mathbb{R},\quad p\in\mathbb{R}\,, \tag{29}$$

and (28) follows.

(4) The proof is completed as follows. We see from (13) and (28) that $\varphi*(W_{g,g}-DW_{f,f})=0$. Here $*$ denotes two-dimensional convolution. Performing the inverse double Fourier transform and using the convolution theorem we get by (14)

$$\Phi(\theta,\tau)[A_{g,g}(\theta,\tau)-DA_{f,f}(\theta,\tau)]=0,\quad \theta\in\mathbb{R},\quad \tau\in\mathbb{R}\,. \tag{30}$$

Here

$$A_{g,g}(\theta,\tau)=\int e^{2\pi i\theta t}g\Big(t+\frac{1}{2}\tau\Big)\overline{g\Big(t-\frac{1}{2}\tau\Big)}dt,\quad \theta\in\mathbb{R},\quad \tau\in\mathbb{R}\,, \tag{31}$$

and $A_{f,f}(\theta,\tau)$ is defined similarly. Since $A_{g,g}, A_{f,f}$ are continuous functions and $\Phi(\theta,\tau)\neq 0$ in a set of (θ,τ), which is dense in $\mathbb{R}^2$, we conclude that $A_{g,g}(\theta,\tau)=DA_{f,f}(\theta,\tau)$ for all $(\theta,\tau)\in\mathbb{R}^2$. It follows easily that g is a multiple of f.

Notes: (1) When we have a Φ such that Moyal's formula

$$\iint C^{(\Phi)}_{f,f}(q,p)C^{(\Phi)}_{g,g}(q,p)dq\,dp=|(f,g)|^2 \tag{32}$$

holds for all $f\in L^2(\mathbb{R})$, $g\in L^2(\mathbb{R})$, then $|\Phi(\theta,\tau)|=1$ (see Ref. 19).

(2) When we take $\Phi(\theta,\tau)=\cos\pi\alpha\theta\tau$, so that $C^{(\Phi)}_{f,f}=\mathrm{Re}\,C^{(\alpha)}_{f,f}$ [cf. (11)], the condition that $\{(\theta,\tau)|\ \Phi(\theta,\tau)\neq 0\}$ is dense in $\mathbb{R}^2$ is satisfied. We shall show in Sec. III that there are no smooth states for which $\mathrm{Re}\,C^{(\alpha)}_{f,f}\geqslant 0$ everywhere.

(3) The condition of having correct marginals forces $\Phi(\theta,0)=\Phi(0,\tau)=1\neq 0$ for $\theta\in\mathbb{R}$, $\tau\in\mathbb{R}$. The proof of Theorem 1 shows that

$$|g(q)|^2=D|f(q)|^2,\quad |G(p)|^2=D|F(p)|^2,\quad q\in\mathbb{R},\quad p\in\mathbb{R}\,, \tag{33}$$

when $C^{(\Phi)}_{af+bg,af+bg}\geqslant 0$ everywhere for all $a\in\mathbb{C}$, $b\in\mathbb{C}$. It does not follow, however, from (33) that f and g are proportional. [As a counterexample, take an $f\in L^2(\mathbb{R})$ with $|f(q)|=|f(-q)|$ for all $q\in\mathbb{R}$, and let $g(q)=\overline{f(-q)}$. Then $|g(q)|=|f(q)|$ for all $q\in\mathbb{R}$ and $G(p)=\overline{F(p)}$ so that $|G(p)|=|F(p)|$ for all $p\in\mathbb{R}$.]

Corollary: The condition $C^{(\Phi)}_{f,f}>0$, $C^{(\Phi)}_{g,g}>0$ everywhere can be replaced by the condition $C^{(\Phi)}_{f,f}\geqslant 0$, $C^{(\Phi)}_{g,g}\geqslant 0$ everywhere when at least one of the functions $f\cdot g$ and $F\cdot G$ does not vanish identically. To show this, we only have to provide a new proof of formula (28). To this end we let I and J be two open intervals where f and F, respectively, have no zeros. We claim that there is a constant $D_{I,J}$ such that $C_{g,g}=D_{I,J}C_{f,f}$ in the set $S_{I,J}=I\times\mathbb{R}\cup\mathbb{R}\times J$, while $|g/f|^2=D_{I,J}$ in I, $|G/F|^2=D_{I,J}$ in J. To see this we let $(q,p)\in S_{I,J}$ such that $C^{(\Phi)}_{f,f}(q,p)>0$. Since $C_{f,f}$ is continuous, there are open intervals $I_q\subset I$, $J_p\subset J$ such that $C^{(\Phi)}_{f,f}(q,p)>0$ in $I_q\times J_p$. Now formulas (21) and (27) show that there is a $D_{q,p}\geqslant 0$ such that $C^{(\Phi)}_{g,g}(q,p)=D_{q,p}C^{(\Phi)}_{f,f}(q,p)$ in $I_q\times J_p$, while $|g/f|^2=D_{q,p}$ in I_q, $|G/F|^2=D_{q,p}$ in J_p. Hence, $|g/f|^2$ is a continuous function on I and for every $q\in I$ there is an open interval I_q containing q where $|g/f|^2$ is constant. Hence $|g/f|^2$ is constant on I. Similarly, $|G/F|^2$ is constant on J, and our initial claim follows. When I, J and K,L are four open intervals such that f and F have no zeros in I,K and J,L, respectively, we find four constants $D_{I,J}$, $D_{I,L}$, $D_{K,J}$, $D_{K,L}$. These constants must all be equal since, e.g., $D_{I,J}=D_{I,L}=|g(q)/f(q)|^2$, when $q\in I$. It thus follows that there is a constant $D\geqslant 0$ such that $C^{(\Phi)}_{g,g}(q,p)=DC^{(\Phi)}_{f,f}(q,p)$ for all $(q,p)\in\mathbb{R}^2$ with $C^{(\Phi)}_{f,f}(q,p)>0$. Similarly, there is a constant $E\geqslant 0$ such that $C^{(\Phi)}_{f,f}(q,p)=EC^{(\Phi)}_{g,g}(q,p)$ for all $(q,p)\in\mathbb{R}^2$ with $C^{(\Phi)}_{g,g}(q,p)>0$. Now when there is a point q with $f(q)g(q)\neq 0$ or a point p with $F(p)G(p)\neq 0$ we see that $D\neq 0\neq E$ and $E=D^{-1}$. And then $C^{(\Phi)}_{g,g}=DC^{(\Phi)}_{f,f}$ everywhere, as was required to prove. We note that there exist smooth $f\in L^2(\mathbb{R})$, $g\in L^2(\mathbb{R})$ with $f\cdot g=0$, $F\cdot G=0$. These examples can be found by properly smoothing, multiplying, and shifting the generalized function $f_0=\Sigma_{n=-\infty}^{\infty}\delta_n$, whose Fourier transform equals $F_0=\Sigma_{m=-\infty}^{\infty}\delta_m$.

III. WIGNER-TYPE PHASE-PLANE DISTRIBUTION FUNCTIONS THAT ARE NON-NEGATIVE

We consider in this section phase-plane distribution functions $\mathrm{Re}\,C^{(\alpha)}_{f,f}$ with $\alpha\in\mathbb{R}$ and $f\in L^2(\mathbb{R})$ a smooth state. We are particularly interested in consequences for the state f of non-negativity of $\mathrm{Re}\,C^{(\alpha)}_{f,f}$ in certain strips in the phase plane. Our main result is that there are no smooth states f with $\mathrm{Re}\,C^{(\alpha)}_{f,f}\geqslant 0$ everywhere, except when $\alpha=0$. This section is divided into four subsections. In Sec. III A we examine the consequence of non-negativity of $\mathrm{Re}\,C^{(\alpha)}_{f,f}$ in strips for the (smooth) state f. The results of this subsection are based on the inversion formulas

$$h(q,s) := f(q + (\tfrac{1}{2} - \alpha)s)\overline{f(q - (\tfrac{1}{2} + \alpha)s)}$$
$$= \int e^{2\pi i s p}\, C^{(\alpha)}_{f,f}(q,p)dp\,, \tag{34}$$
$$H(p,s) := F(p - (\tfrac{1}{2} + \alpha)s)\overline{F(p + (\tfrac{1}{2} - \alpha)s)}$$
$$= \int e^{2\pi i s q}\, C^{(\alpha)}_{f,f}(q,p)dq\,, \tag{35}$$
or rather
$$g(q,s) := \tfrac{1}{2}h(q,s) + \tfrac{1}{2}\overline{h(q,-s)}$$
$$= \int e^{2\pi i s p}\, \mathrm{Re}\, C^{(\alpha)}_{f,f}(q,p)dp\,, \tag{36}$$
$$G(p,s) := \tfrac{1}{2}H(p,s) + \tfrac{1}{2}\overline{H(p,-s)}$$
$$= \int e^{2\pi i s q}\, \mathrm{Re}\, C^{(\alpha)}_{f,f}(q,p)dq\,. \tag{37}$$
Another important fact is $C^{(\alpha)}_{F,F}(q,p) = C^{(\alpha)}_{f,f}(-p,q)$, $q \in \mathbb{R}$, $p \in \mathbb{R}$. In Sec. III B we give the proof of our main result for smooth functions and $0 \neq |\alpha| \neq \frac{1}{2}$, and in Sec. III C we treat the case of $f \in L^2(\mathbb{R})$, $\alpha = \pm \frac{1}{2}$. The case $\alpha = 0$ has been considered in Refs. 4 and 20 and does not need a proof here. Finally, Sec. III D contains some examples of generalized functions f and numbers α such that $\mathrm{Re}\, C^{(\alpha)}_{f,f} \geqslant 0$ everywhere.

A. States with $\mathrm{Re}\, C^{(\alpha)}_{f,f} \geqslant 0$ in a strip

We assume in this subsection that both f and F are smooth functions of q and p respectively so that all manipulations below can be justified (it is not hard to give more specific conditions that guarantee this).

Theorem 2: Let $-\infty < a < b < \infty$, and assume that $f(q) \neq 0$, $\mathrm{Re}\, C^{(\alpha)}_{f,f}(q,p) \geqslant 0$ for $q \in (a,b)$, $p \in \mathbb{R}$. Write $|f(q)| = \exp(\psi(q))$ with ψ smooth. Then
$$(\tfrac{1}{4} + \alpha^2)\,\psi''(q) + 2\alpha^2(\psi'(q))^2 \leqslant 0, \quad q \in (a,b)\,. \tag{38}$$
Proof: Write $f(q) = \exp(\psi(q) + i\varphi(q))$ with φ smooth on (a,b), and insert the expansions
$$\varphi(q+u) = \varphi(q) + u\varphi'(q) + \tfrac{1}{2}u^2\varphi''(q) + o(u^2),$$
$$u \to 0\,, \tag{39}$$
$$\psi(q+u) = \psi(q) + u\psi'(q) + \tfrac{1}{2}u^2\psi''(q) + o(u^2),$$
$$u \to 0\,, \tag{40}$$
into the definition (34) of $h(q,s)$ and $\overline{h(q,-s)}$. We get
$$\begin{aligned} h(q,s) &= \exp(2\psi(q) - 2\alpha s\psi'(q) + (\tfrac{1}{4} + \alpha^2)s^2\psi''(q) \\ &\quad + i[s\varphi'(q) - \alpha s^2\varphi''(q)] + o(s^2)), \\ \overline{h(q,-s)} &= \exp(2\psi(q) + 2\alpha s\psi'(q) + (\tfrac{1}{4} + \alpha^2)s^2\psi''(q) \\ &\quad + i[s\varphi'(q) + \alpha s^2\varphi''(q)] + o(s^2)), \quad s \to 0\,. \end{aligned} \tag{41}$$
Hence
$$\begin{aligned} g(q,s) &= \tfrac{1}{2}h(q,s) + \tfrac{1}{2}\overline{h(q,-s)} \\ &= \tfrac{1}{2}\exp(2\psi(q) + (\tfrac{1}{4} + \alpha^2)s^2\psi''(q) + is\varphi'(q)) \\ &\quad \times \{\exp(-2\alpha s\psi'(q) - i\alpha s^2\varphi''(q)) \\ &\quad + \exp(2\alpha s\psi'(q) + i\alpha s^2\varphi''(q))\}(1 + o(s^2)) \\ &= \exp(2\psi(q) + (\tfrac{1}{4} + \alpha^2)s^2\psi''(q) + is\varphi'(q)) \\ &\quad \times \{\cosh(2\alpha s\psi'(q) + i\alpha s^2\varphi''(q))\}(1 + o(s^2)), \quad s \to 0\,. \end{aligned} \tag{42}$$
Now
$$\cosh(2\alpha s\psi'(q) + i\alpha s^2\varphi''(q))$$
$$= 1 + 2\alpha^2 s^2(\psi'(q))^2 + o(s^2), \quad s \to 0\,, \tag{43}$$
so that
$$\begin{aligned} |g(q,s)| &= \exp(2\psi(q) + (\tfrac{1}{4} + \alpha^2)s^2\psi''(q)) \\ &\quad \times \{1 + 2\alpha^2 s^2(\psi'(q))^2 + o(s^2)\}(1 + o(s^2)) \\ &= \exp(2\psi(q))\{1 + (\tfrac{1}{4} + \alpha^2)s^2\psi''(q) \\ &\quad + 2\alpha^2 s^2(\psi'(q))^2 + o(s^2)\}(1 + o(s^2))\,, \quad s \to 0\,. \end{aligned} \tag{44}$$
Since $\mathrm{Re}\, C^{(\alpha)}_{f,f}(q,p) \geqslant 0$, $p \in \mathbb{R}$, we see from (36) that
$$|g(q,s)| \leqslant g(q,0) = |f(q)|^2 = \exp(2\psi(q)), \quad s \in \mathbb{R}\,. \tag{45}$$
This can only be true when (38) holds, and the proof is complete.

Notes: (1) When $-\infty < c < d < \infty$ and $F(p) \neq 0$, $\mathrm{Re}\, C^{(\alpha)}_{f,f}(q,p) \geqslant 0$ for $q \in \mathbb{R}$, $p \in (c,d)$, we have
$$(\tfrac{1}{4} + \alpha^2)\Psi''(p) + 2\alpha^2(\Psi'(p))^2 \leqslant 0, \quad p \in (c,d)\,, \tag{46}$$
where $|F(p)| = \exp(\Psi(p))$ and Ψ is smooth on (c,d).

(2) The condition (38) [and (46)] is weakest for $\alpha = 0$. When $\alpha = 0$ we see that (38) [(46)] means that $\psi(\Psi)$ is concave on (a,b) [(c,d)].

We next study the behavior of an f with $\mathrm{Re}\, C^{(\alpha)}_{f,f}(q,p) \geqslant 0$, where q is a zero of f.

Theorem 3: Let $|\alpha| \neq \frac{1}{2}$, $q \in \mathbb{R}$, and assume that $f(q) = 0$, $\mathrm{Re}\, C^{(\alpha)}_{f,f}(q,p) \geqslant 0$, $p \in \mathbb{R}$. Then $f^{(n)}(q) = 0$, $n = 1,2,\ldots$.

Proof: We have as in the proof of Theorem 2 that
$$|g(q,s)| = |\tfrac{1}{2}h(q,s) + \tfrac{1}{2}\overline{h(q,-s)}|$$
$$\leqslant |g(q,0)| = |f(q)|^2 = 0\,. \tag{47}$$
Let $n \in \mathbb{N}$ be the smallest number with $f^{(n)}(q) \neq 0$. Then
$$\begin{aligned} h(q,s) &= f(q + (\tfrac{1}{2} - \alpha)s)\overline{f(q - (\tfrac{1}{2} + \alpha)s)} \\ &= [(\tfrac{1}{2} - \alpha)^n(s^n/n!)f^{(n)}(q) + o(s^n)] \\ &\quad \times [(-(\tfrac{1}{2} + \alpha))^n(s^n/n!)\overline{f^{(n)}(q)} + o(s^n)] \\ &= [(-1)^n(\tfrac{1}{4} - \alpha^2)^n/(n!)^2]\,|f^{(n)}(q)|^2 s^{2n} \\ &\quad + o(s^{2n}), \quad s \to 0\,. \end{aligned} \tag{48}$$
Similarly,
$$\overline{h(q,-s)} = [(-1)^n(\tfrac{1}{4} - \alpha^2)^n/(n!)^2]\,|f^{(n)}(q)|^2 s^{2n}$$
$$+ o(s^{2n}), \quad s \to 0\,. \tag{49}$$
Hence
$$g(q,s) = [(-1)^n(\tfrac{1}{4} - \alpha^2)^n/(n!)^2]\,|f^{(n)}(q)|^2 s^{2n}$$
$$+ o(s^{2n}), \quad s \to 0\,. \tag{50}$$
This contradicts (47), and therefore $f^{(n)}(q) = 0$.

Theorem 4: Let $\alpha \neq 0$, $|\alpha| \neq \frac{1}{2}$. Assume that $-\infty < a < b < \infty$, that $\mathrm{Re}\, C^{(\alpha)}_{f,f}(q,p) \geqslant 0$ for $q \in (a,b)$, $p \in \mathbb{R}$, and that $f(a) \neq 0 \neq f(b)$. Then the set $\{q \mid f(q) \neq 0\}$ is dense in (a,b).

Proof: Suppose we can find an interval $(c_0,d_0) \subset (a,b)$ such that $f(q) = 0$ for $q \in (c_0,d_0)$. We can assume that both c_0

and d_0 are accumulation points of the set $\{q\,|\,f(q)\neq 0\}$. When $q\in(c_0,d_0)$ we have for $s\in\mathbb{R}$

$$f(q+(\tfrac12-\alpha)s)\overline{f(q-(\tfrac12+\alpha)s)} + \overline{f(q-(\tfrac12-\alpha)s)}f(q+(\tfrac12+\alpha)s)=0\,. \tag{51}$$

We consider the following cases.

(1) $0<\alpha<\frac12$. We can find $c_n<c_0$, $d_n>d_0$ with $f(c_n)\neq 0\neq f(d_n)$, $c_n\uparrow c_0$, $d_n\downarrow d_0$. Now if we let

$$q_n=\tfrac12(c_n+d_n)+\alpha(d_n-c_n),\quad s_n=d_n-c_n,\quad n=1,2,\ldots, \tag{52}$$

then we have $q_n+(\frac12-\alpha)s_n=d_n$, $q_n-(\frac12+\alpha)s_n=c_n$ while $q_n-(\frac12-\alpha)s_n\in(c_0,d_0)$ for large n. We conclude that the left-hand side of (51) with $q=q_n$, $s=s_n$ equals $f(d_n)\overline{f(c_n)}\neq 0$ for large n. We thus have a contradiction.

(2) $-\frac12<\alpha<0$. This case is similar to the previous one.

(3) $\alpha>\frac12$. Let $c_n<c_0$ be such that $c_n\uparrow c_0, f(c_n)\neq 0$. Now if we let

$$q_{n,m}=\tfrac12(c_n+c_m)+\alpha(c_m-c_n),\quad s_{n,m}=c_m-c_n,\quad n,m=1,2,\ldots, \tag{53}$$

we have $q_{n,m}+(\frac12-\alpha)s_{n,m}=c_m$, $q_{n,m}-(\frac12+\alpha)s_{n,m}=c_n$. We can take n so large that $q_{n,\infty}+(\frac12+\alpha)s_{n,\infty}\in(c_0,d_0)$. Here $q_{n,\infty}=\frac12(c_n+c_0)+\alpha(c_0-c_n)$, $s_{n,\infty}=c_0-c_n$. Observe that $q_{n,\infty}>c_0$, and that $q_{n,m}<q_{n,\infty}$, $s_{n,m}<s_{n,\infty}$, $q_{n,m}\uparrow q_{n,\infty}$ when $m\to\infty$. Hence, when m is large enough, $q_{n,m}>c_0$ while $q_{n,m}+(\frac12+\alpha)s_{n,m}<d_0$. Therefore, the left-hand side of (51), with $q=q_{n,m}$, $s=s_{n,m}$, equals $f(c_m)\overline{f(c_n)}\neq 0$ when m is large enough. We thus have a contradiction.

(4) $\alpha<-\frac12$. This case is similar to the case $\alpha>\frac12$.

B. Smooth states $f\neq 0$ with Re $C^{(\alpha)}_{f,f}\geqslant 0$ everywhere do not exist: case $0\neq|\alpha|\neq\frac12$

We now start the proof of the statement that there are no smooth states f with Re $C^{(\alpha)}_{f,f}\geqslant 0$ everywhere for the case $0\neq|\alpha|\neq\frac12$; the case $\alpha=0$ is covered by Hudson's theorem and the case $|\alpha|=\frac12$ will be considered in Sec. III C. The proof is lengthy and consists of several steps; it can be outlined as follows. Suppose that we have a smooth $f\not\equiv 0$ with Re $C^{(\alpha)}_{f,f}\geqslant 0$ everywhere. It will be shown in Lemma 1 below that the zero set of f must be unbounded above and below; at the same time it will be shown that we must have $|\alpha|<\frac12$. In Lemma 2 below it will be shown that f cannot vanish identically on any interval. The remainder of the proof consists of a careful analysis of f around its zeros. When $f(a)=0$, we have

$$f(a+(\tfrac12-\alpha)s)\,\overline{f(a-(\tfrac12+\alpha)s)} = -\overline{f(a-(\tfrac12-\alpha)s)}f(a+(\tfrac12+\alpha)s),\quad s\in\mathbb{R}\,. \tag{54}$$

This identity can be used to show the following (Lemma 3): Let a be a zero of f for which there is a $\delta>0$ such that $f(q)\neq 0$ for $q\in(a-\delta,a)$. Then there is a smooth function k_a: $(0,\infty)\to\mathbb{C}$ such that

$$k_a(q)=-k_a([(\tfrac12+\alpha)/(\tfrac12-\alpha)]q),\quad q>0\,, \tag{55}$$

while

$$f(a+q)=k_a(q)f(a-q),\quad q>0\,. \tag{56}$$

Then it can be shown that the zero set of f has the form $\{a+bl\,|\,l\in\mathbb{Z}\}$, where $a\in\mathbb{R}$, $b\in\mathbb{R}$. We finally derive a contradiction by showing the identity

$$k_a(q)=k_{a+b}(q-b)\,\overline{k_a(2b-q)}k_{a-b}(q),\quad b<q<2b\,, \tag{57}$$

in which the left-hand side behaves smoothly as $q\downarrow b$ whereas the right-hand side oscillates violently as $q\downarrow b$ because of the factor $k_{a+b}(q-b)$ and (55).

Lemma 1: (i) Let $0<|\alpha|<\frac12$, and assume Re $C^{(\alpha)}_{f,f}\geqslant 0$ everywhere. Then the zero set of f is unbounded below and above.

(ii) Let $|\alpha|>\frac12$. Then Re $C^{(\alpha)}_{f,f}$ takes negative values.

Proof: (i) Suppose there is a $b\in\mathbb{R}$ with $f(q)\neq 0$ for $q<b$. Then we can write $|f(q)|=\exp(\psi(q))$, where ψ is smooth and satisfies (38) for $q<b$. In particular ψ is concave on $(-\infty,b)$. Since $\int|f(q)|^2\,dq<\infty$ we must have that $\psi(q)\to-\infty$ as $q\to-\infty$. Furthermore we can find a $q_0<b$ such that $\psi'(q_0)>0$. Now $\psi'(q)\geqslant\psi'(q_0)$ for $q<q_0$, and, according to (38),

$$1/\psi'(q)-1/\psi'(q_0)=\int_q^{q_0}\frac{\psi''(r)}{(\psi'(r))^2}\,dr\leqslant -c_\alpha(q_0-q),\quad q\leqslant q_0\,, \tag{58}$$

where $c=2\alpha^2/(\frac14+\alpha^2)>0$. Therefore

$$1/\psi'(q)\leqslant 1/\psi'(q_0)-c_\alpha(q_0-q),\quad q\leqslant q_0\,. \tag{59}$$

The left-hand side of (59) is positive for $q\leqslant q_0$ whereas the right-hand side of (59) tends to $-\infty$ when $q\to-\infty$. Contradiction. Hence, f must have zeros in $(-\infty,b)$. In a similar way we conclude that f has zeros in any interval (a,∞).

(ii) Suppose Re $C^{(\alpha)}_{f,f}\geqslant 0$ everywhere and $|\alpha|>\frac12$. In the proof of (i) it was not used that $0<|\alpha|<\frac12$. Since $f\not\equiv 0$ we can find a $q_0\in\mathbb{R}$ with $f(q_0)\neq 0$. Now let $a:=\inf\{q_1\,|\,f(q)\neq 0$ for $q\in(q_1,q_0)\}$. Then $a>-\infty$, $f(a)=0$, and there is a $b>a$ such that $f(q)\neq 0$ for $q\in(a,b)$. As $f(q)\neq 0$ for $q\in(a,b)$ we have, according to (38),

$$|f|''(q)=[\psi''(q)+(\psi'(q))^2]\exp(\psi(q)) \leqslant -\frac{\alpha^2-\frac14}{\alpha^2+\frac14}(\psi'(q))^2\exp(\psi(q))\leqslant 0,\quad q\in(a,b)\,. \tag{60}$$

Hence $|f|$ is concave on (a,b) and $\lim_{q\downarrow a}|f|'(q)>0$ (this limit may be $+\infty$). But

$$\begin{aligned}|f|'(q)&=\exp(-i\arg f(q))[f'(q)-i(\arg f)'(q)f(q)]\\&=\exp(-i\arg f(q))f'(q)\\&\quad -i|f(q)|(\arg f)'(q),\quad q\in(a,b)\,,\end{aligned} \tag{61}$$

and the real part of the right-hand side tends to 0 on account of Theorem 3. This results in a contradiction, and the proof is complete.

We assume from now on that $0\neq|\alpha|<\frac12$.

Lemma 2: The set $\{q\,|\,f(q)\neq 0\}$ is dense in $\mathbb{R}$.

Proof: We proceed according to the following steps.

(1) f cannot have compact support. For otherwise F is an analytic function. According to Lemma 1, F has zeros, and at a zero, p, of F we have according to Theorem 3 that $F^{(n)}(p)=0$. This is impossible in view of the analyticity of F and the fact that $f\not\equiv 0$.

(2) f cannot vanish identically on a semi-infinite interval. For suppose that $f(q)=0$ for $q\leqslant a$, where $a:=\max\{q_1\,|\,f(q)=0,\ q\leqslant q_1\}$. We know from step (2) that

the support of f is unbounded above. Hence, in view of Theorem 4 there are no intervals $[c,d]$ with $d>c>a$ such that f vanishes on $[c,d]$. By Lemma 1 we can find a $q>a$ with $f(q)=0$. When $\alpha<0$, we see from (54) that $\overline{f(q+(\frac{1}{2}-\alpha)s)} \times f(q-(\frac{1}{2}+\alpha)s)=0$ whenever $s\geqslant(\frac{1}{2}-\alpha)^{-1}(q-a)$. However, when $(\frac{1}{2}-\alpha)^{-1}(q-a)<s<(\frac{1}{2}+\alpha)^{-1}(q-a)$ we have $q+(\frac{1}{2}-\alpha)s>a$, $q-(\frac{1}{2}+\alpha)s>a$, so that neither $f(q+(\frac{1}{2}-\alpha)s)$ nor $f(q-(\frac{1}{2}+\alpha)s)$ can vanish identically for s in any subinterval of $(\frac{1}{2}-\alpha)^{-1}(q-a)$, $(\frac{1}{2}+\alpha)^{-1}(q-a)$. This contradicts the continuity of f. When $\alpha>0$ we can derive a contradiction in a similar way. It follows therefore that the support of f cannot be bounded below, and in a similar way it can be shown that the support of f is not bounded above. This completes the proof of step (2).

The proof of the lemma is now easily completed by using Theorem 4.

We assume now that $-\frac{1}{2}<\alpha<0$.

Lemma 3: Let a be a zero of f for which there is a $\delta>0$ such that $f(q)\neq 0$ for $q\in(a-\delta,a)$. Then there is a smooth function $k_a:(0,\infty)\mapsto\mathbb{C}$ such that $k_a(q)=-k_a(\beta q)$, $f(a+q)=k_a(q)\overline{f(a-q)}$, $q>0$. Here $\beta=(\frac{1}{2}+\alpha)/(\frac{1}{2}-\alpha)<1$.

Proof: With β as given above we can write (54) as

$$f(a+q)\overline{f(a-\beta q)}=-\overline{f(a-q)}f(a+\beta q),\quad q\in\mathbb{R}. \tag{62}$$

Therefore, when $f(a-q)\neq 0\neq f(a-\beta q)$ (in particular when $0<q<\delta$)

$$f(a+q)/\overline{f(a-q)}=-f(a+\beta q)/\overline{f(a-\beta q)}. \tag{63}$$

Define

$$k_a(q):=f(a+q)/\overline{f(a-q)},\quad 0<q<\delta, \tag{64}$$

and extend the domain of k_a to $(0,\infty)$ by setting for $q\geqslant\delta$

$$k_a(q):=(-1)^n k_a(q\beta^n), \tag{65}$$

where $n=1,2,\ldots$ is such that $q\beta^n\in[\beta\delta,\delta)$. We claim that this k_a satisfies the requirements. Indeed, we have the identity $k_a(q)=-k_a(\beta q)$ for $0<q<\delta$ because of (63), and the extension of k_a according to (65) is such that this identity remains valid for $q\geqslant\delta$. It is also clear that k_a is smooth. Finally, let $q_0>0$. We want to show that $f(a+q_0)=k_a(q_0)\times\overline{f(a-q_0)}$. When $0<q_0<\delta$ this follows at once from the definition (64). When $q_0\geqslant\delta$ we take $n=1,2,\ldots$ such that $q_0\beta^n\in[\beta\delta,\delta)$. The set $\{q\,|\,f(q)\neq 0\}$ is open and dense in $\mathbb{R}$ according to Lemma 2. Hence we can find a sequence $(q_k)_{k=1,2,\ldots}$ with $q_k\beta^n\in[\beta\delta,\delta)$, $q_k\to q_0$, and $f(a-q_k)\neq 0$, $f(a-\beta q_k)\neq 0,\ldots,f(a-\beta^{n-1}q_k)\neq 0$. It follows from (63)–(65) that $f(a+q_k)=k_a(q_k)\overline{f(a-q_k)}$ for all $k=1,2,\ldots$. By taking the limit $k\to\infty$ and using continuity of f and k_a we conclude that $f(a+q_0)=k_a(q_0)\overline{f(a-q_0)}$, and the proof is complete.

Corollaries: (1) Let a be a zero of f for which there is a $\delta>0$ such that $f(q)\neq 0$ for $q\in(a,a+\delta)$. Then there is a smooth function l_a: $(0,\infty)\mapsto\mathbb{C}$ such that $l_a(q)=-l_a(\beta q)$, $\overline{f(a-q)}=l_a(q)f(a+q)$, $q>0$. Here $\beta=(\frac{1}{2}+\alpha)/(\frac{1}{2}-\alpha)<1$.

(2) Let a be a zero of f for which there is a $\delta>0$ such that $f(q)\neq 0$ for $q\in(a,a+\delta)\cup(a-\delta,a)$. Then the function k_a of Lemma 3 has no zeros. Indeed, $k_a(q)\neq 0$ for $0<q<\delta$, and hence, by (65), $k_a(q)\neq 0$ for $q>0$.

Lemma 4: The zero set of f is of the form $\{a+lb\,|\,l\in\mathbb{Z}\}$ for some $a\in\mathbb{R}$, $b\in\mathbb{R}$.

Proof: We can find an $a\in\mathbb{R}$, $b>0$ such that $f(a)=0=f(a+b)$ while $f(q)\neq 0$ for $q\in(a,a+b)$. Assume the interval $(a-b,a)$ contains a zero $a-c_0$ of f with $0<c_0<b$. With l_a as in Corollary 1 we see that $l_a(c_0)=0$, as $f(a+c_0)\neq 0$. Since $l_a(q)=-l_a(\beta q)$ for $q>0$ we get that $f(a-c_0\beta^n)=0$ for $n=1,2,\ldots$. Hence a is an accumulation point of the zeros of f less than a. We can find zeros c and d of f with $a-\frac{1}{2}b<c<d<a$ such that the interval (c,d) is zero-free [otherwise f would vanish identically on a subinterval of $(a-\frac{1}{2}b,a)$, which cannot happen by Theorem 4]. According to Lemma 3 we have $f(d+q)=k_d(q)\overline{f(d-q)}$ for all $q>0$. Since $f(d+q)\neq 0$ for $q\in(a-d,a+b-d)$ we see that $f(d-q)\neq 0$ for $q\in(a-d,a+b-d)$. In particular $f(a-b)=f(d-(d-(a-b)))\neq 0$ since $d-(a-b)\in(a-d,a+b-d)$. However, by Corollary 1, $\overline{f(a-b)}=l_a(b)f(a+b)=0$. Thus, we have a contradiction. Hence, f has no zeros in $(a-b,a)$ while $f(a-b)=0$. It follows now by induction that $f(a-2b)=f(a-3b)=\cdots=0$, while $f(c)\neq 0$ for $c<a, c\neq a-b, a-2b,\ldots$. Similarly, $f(a+2b)=f(a+3b)=\cdots=0$ while $f(c)\neq 0$ for $c>a, c\neq a+b, a+2b, a+3b,\ldots$, and the proof is complete.

We shall now finish the proof of the main result of this subsection. We assume for convenience that $a=0$, $b=1$. According to Corollary 2 to Lemma 3 the functions k_l with integer l have no zeros, are smooth and satisfy $k_l(q)=-k_l(\beta q)$, $f(l+q)=k_l(q)\overline{f(l-q)}$ for $q>0$. Therefore, $k_l(q)$ is bounded away from 0 and oscillates violently when $q\downarrow 0$. We shall show that

$$k_1(q-1)\overline{k_0(2-q)}k_{-1}(q)=k_0(q),\quad 1<q<2. \tag{66}$$

Indeed, we have for $1<q<2$ that $f(q)=k_0(q)\overline{f(-q)}$, and at the same time

$$\begin{aligned} f(q)&=k_1(q-1)\overline{f(1-(q-1))}=k_1(q-1)\overline{f(2-q)}\\ &=k_1(q-1)\overline{k_0(2-q)\overline{f(q-2)}}\\ &=k_1(q-1)\overline{k_0(2-q)}f(-1+(q-1))\\ &=k_1(q-1)\overline{k_0(2-q)}k_{-1}(q-1)\overline{f(-q)}. \end{aligned} \tag{67}$$

Since f has no zeros in $(-2,-1)$ we conclude that (66) holds by equating the two expressions for $f(q)$. From (66) we easily derive a contradiction: $\lim_{q\downarrow 1}k_0(q)$, $\lim_{q\downarrow 1}k_0(2-q)$, $\lim_{q\downarrow 1}k_{-1}(q)$ exist and are unequal to 0 whereas $\lim_{q\downarrow 1}k_1(q-1)$ does not exist.

This completes the proof of our main result for the case $-\frac{1}{2}<\alpha<0$. The proof for the case $0<\alpha<\frac{1}{2}$ is practically the same.

Note: We can ask ourselves how close we can get to proving Hudson's theorem for the Wigner distribution by employing the same techniques as in this section. It is not hard to show that any smooth f with $W_{f,f}\geqslant 0$ everywhere has no zeros, and that $|f(q)|=\exp(\psi(q))$, $|F(p)|=\exp(\Psi(p))$, where ψ and Ψ are concave functions defined on $\mathbb{R}$ with $\psi(q)\to-\infty$ as $q\to\pm\infty$, $\Psi(p)\to-\infty$ as $p\to\pm\infty$. We have not been able to find examples f [other than Gaussians $f(q)=\exp(-\pi\gamma q^2+2\pi\delta q-\pi\epsilon)$ with $\operatorname{Re}\gamma>0$, $\delta\in\mathbb{C}$, $\epsilon\in\mathbb{C}$]

such that both log $|f|$ and log $|F|$ are well defined and concave everywhere on $\mathbb{R}$.

C. Square-integrable states f with Re $C_{f,f}^{(\alpha)} \geqslant 0$ everywhere do not exist: case $|\alpha| = \frac{1}{2}$

In this subsection we shall show that there is no $f \in L^2(\mathbb{R})$ such that Re $C_{f,f}^{(1/2)} \geqslant 0$ almost everywhere (unless $f = 0$ almost everywhere). The reason why the proof in Sec. III B for (smooth) f's fails when $\alpha = \pm \frac{1}{2}$ is that formula (54) does not provide useful information when $f(a) = 0$. [The result of Theorem 2 is still valid and shows that $\psi''(q) + (\psi'(q))^2 \leqslant 0$ on any interval where f has no zeros; this Theorem 2 implies that $|f|$ is concave when f is smooth on such an interval.] We therefore have to resort to entirely different methods. We shall show that the main result in this case also holds for all $f \in L^2(\mathbb{R})$, which are not necessarily smooth. Furthermore, the proof is more constructive in the sense that one can more explicitly indicate the regions where Re $C_{f,f}^{(\pm 1/2)}$ takes negative values.

Since $C_{f,f}^{(1/2)}(q,p) = \overline{C_{f,f}^{(-1/2)}(q,p)}$ it is sufficient to consider the case $\alpha = \frac{1}{2}$ only. We have

$$\mathrm{Re}\, C_{f,f}^{(1/2)}(q,p) = \mathrm{Re}[e^{2\pi iqp}\, \overline{f(q)} F(p)] \,. \tag{68}$$

The proof can be outlined as follows. Assume for a while that $f(q)$ is positive on an interval $[a,b]$. Then non-negativity of (68) for all q and p implies that $\mathrm{Re}[e^{2\pi iqp} F(p)] \geqslant 0$ for $q \in [a,b]$, $p \in \mathbb{R}$. When $|p|$ is sufficiently large, $\{e^{2\pi iqp} \mid q \in [a,b]\} = \{z \mid |z| = 1\}$, leaving for $F(p)$ no other possibility than $F(p) = 0$. Hence $F(p) = 0$ for $|p|$ sufficiently large. When f is continuous and nonzero in an interval a slightly more sophisticated argument gives the same conclusion. When F is continuous and nonzero in an interval as well, we can argue in a similar fashion that $f(q) = 0$ for $|q|$ sufficiently large. This then gives a contradiction since f and F cannot both be compactly supported. The reasoning is essentially the same but gets more technical when $f \in L^2(\mathbb{R})$ since one has now to consider Lebesgue points[21] of f and F, instead of continuity points.

We proceed with the proof according to the following steps, where we denote the sets of Lebesgue points of f and F by LP_f and LP_F, respectively. We suppose that $\mathrm{Re}[e^{2\pi iqp}\, \overline{f(q)} F(p)] \geqslant 0$ almost everywhere, where $f \not\equiv 0$.

Step 1: When $q_0 \in \mathrm{LP}_f$, $p_0 \in \mathrm{LP}_F$, and $f(q_0) \neq 0 \neq F(p_0)$, then $\mathrm{Re}[e^{2\pi iq_0p_0}\, \overline{f(q_0)} F(p_0)] \geqslant 0$. Indeed if this is not true, we can find a $\delta > 0$ such that $\mathrm{Re}[e^{2\pi iqp} \bar{z} w] < 0$ for all $z \in \mathbb{C}$, $w \in \mathbb{C}$, $q \in \mathbb{R}$, $p \in \mathbb{R}$ with $|z - f(q_0)| < \delta$, $|w - F(p_0)| < \delta$, $|q - q_0| < \delta$, $|p - p_0| < \delta$. Since $q_0 \in \mathrm{LP}_f$ we can find[21] an ϵ_1, $0 < \epsilon_1 < \delta$, with $\mu(\{q \in [q_0 - \epsilon_1, q_0 + \epsilon_1] \mid |f(q) - f(q_0)| < \delta\}) > \epsilon_1$. Similarly, we can find an ϵ_2, $0 < \epsilon_2 < \delta$ such that $\mu(\{p \in [p_0 - \epsilon_2, p_0 + \epsilon_2] \mid |F(p) - F(p_0)| < \delta\}) > \epsilon_2$. With $\epsilon := \min(\epsilon_1, \epsilon_2)$ we get $\mu(\{(q,p) \mid \mathrm{Re}[e^{2\pi iqp}\, \overline{f(q)} F(p)] < 0\}) > \epsilon^2$, and this contradicts the assumption that $\mathrm{Re}[e^{2\pi iqp}\, \overline{f(q)} F(p)] \geqslant 0$ almost everywhere.

Step 2: Let $q \in \mathrm{LP}_f$ with $q > 0, f(q) \neq 0$, let C be the conic set $\{z \mid \arg z \in (\pi a, \pi b)\}$ with $b - a < 1$, and let $F^{\leftarrow}(C) = \{p \in \mathbb{R} \mid F(p) \in C\}$. Then we have, for $n = 0,1,\ldots$,

$$I_n(q) \cap \mathrm{LP}_F \cap F^{\leftarrow}(C) = \phi \,, \tag{69}$$

where $I_n(q)$ is the interval

$$I_n(q) = \left[\frac{2n + \frac{1}{2} + \pi^{-1} \arg f(q) - a}{2q}, \frac{2n + \frac{3}{2} + \pi^{-1} \arg f(q) - b}{2q} \right], \tag{70}$$

whose midpoint $(2n + 1 + (1/\pi) \arg f(q) - \frac{1}{2}(a + b))/2q$ and length $[1 - (b - a)]/2q$ are denoted by $m_n(q)$ and $l(q)$, respectively. Indeed when $p \in I_n(q) \cap \mathrm{LP}_F \cap F^{\leftarrow}(C)$ we have $\mathrm{Re}[e^{2\pi iqp}\, \overline{f(q)} F(p)] < 0$, and this contradicts the result of step 1.

Step 3: Let $q_0 \in \mathrm{LP}_f$ with $q_0 > 0, f(q_0) \neq 0$, and let

$$V_{\delta,\epsilon} := \{q \in (q_0 - \epsilon, q_0 + \epsilon) \mid |f(q) - f(q_0)| < \delta\} \cap \mathrm{LP}_f \tag{71}$$

for $\delta > 0$, $\epsilon > 0$, $\epsilon < \frac{1}{2} q_0$. Furthermore let $J_n(q)$ be the closed interval with midpoint $(2n/2q) + m_0(q_0)$ and length $\frac{1}{4} l(q_0)$. We can find $\delta > 0$, $\epsilon > 0$ such that $J_n(q) \subset I_n(q)$ for $q \in V_{\delta,\epsilon}$. Indeed this is achieved when δ and ϵ are so small that $|m_0(q) - m_0(q_0)| < \frac{1}{8} l(q_0)$.

Step 4: Let $\epsilon(n) = 4q_0/(2n + 1)$. Then the set of midpoints of $J_n(q)$ with $|q - q_0| < \frac{1}{2} \epsilon(n)$ equals $(m_n(q_0) - (1/q_0) \times [2n/(2n + 3)], m_n(q_0) + (1/q_0)[2n/(2n - 1)])$.

Step 5: We have[21] $\lim_{\epsilon \downarrow 0} \mu(V_{\delta,\epsilon})/2\epsilon = 1$ for every $\delta > 0$. We can take n so large that $(q_0 - \epsilon(n), q_0 + \epsilon(n)) \setminus V_{\delta,\epsilon(n)}$ contains no intervals of length $\geqslant \frac{1}{2} q_0\, l(q_0)\, \epsilon(n)$. The latter number is $< \frac{1}{2}\epsilon(n)$, and when q_1, $q_2 \in (q_0 - \epsilon(n), q_0 + \epsilon(n))$, $|q_1 - q_2| < \frac{1}{2} q_0 l(q_0) \epsilon(n)$, then $J_n(q_1) \cap J_n(q_2) \neq \phi$. Hence,

$$S_n(q_0) := \bigcup_{q \in V_{\delta,\epsilon(n)}} J_n(q) \supset \bigcup_{|q - q_0| < \frac{1}{2}\epsilon(n)} J_n(q) \,, \tag{72}$$

and this set contains the interval of length $= 4q_0^{-1} \times (2n + 3)^{-1} n$ with midpoint $m_n(q_0)$. Hence the $S_n(q_0)$'s overlap when n is large enough.

Step 6: When $p \in S_n(q_0) \cap \mathrm{LP}_F$, we have $F(p) \notin C$. For otherwise $p \in I_n(q) \cap \mathrm{LP}_F \cap F^{\leftarrow}(C)$ for some $q \in V_{\delta,\epsilon(n)}$, which contradicts (69). Hence $F(p) \notin C$ when $p > 0$ is sufficiently large.

Step 7: By taking three different conic sets C_1, C_2, C_3 as in step 2 with $C_1 \cup C_2 \cup C_3 = \mathbb{C} \setminus \{0\}$, we see that $F(p) = 0$ when $p > 0$ is sufficiently large. Similarly, $F(p) = 0$ when $-p > 0$ is sufficiently large.

Step 8: By interchanging q and p we see that $f(q) = 0$ when $|q|$ is sufficiently large. The proof is now completed by noting that f and F cannot both be compactly supported.

Note: As the proof shows, it is not necessary to require $f \in L^2(\mathbb{R})$; the requirement that f, F can be identified with locally integrable functions is sufficient. For instance, when $f \in L^1(\mathbb{R})$ we have that F is continuous and bounded, and $\mathrm{Re}[e^{2\pi iqp}\, \overline{f(q)} F(p)]$ takes negative values.

D. Examples of generalized functions f with Re $C_{f,f}^{(\alpha)} \geqslant 0$ everywhere

In this subsection we give examples of generalized functions f and $\alpha \in \mathbb{R}$ such that Re $C_{f,f}^{(\alpha)} \geqslant 0$ everywhere (in the generalized sense). It was already noted in Sec. I that the existence of such f's and α's is remarkable in view of the fact that the sets of smooth and generalized functions f with $W_{f,f} = C_{f,f}^{(0)} \geqslant 0$ everywhere are essentially the same. In constructing examples of f's with Re $C_{f,f}^{(\alpha)} \geqslant 0$ we are led by what may be called the interference formula

$$|C^{(\alpha)}_{f,f}(q,p)|^2 = \left(\frac{1}{2}-\alpha\right)^2 \iint W_{f,f}\left(\frac{q}{\frac{1}{2}+\alpha}-a, \frac{p}{\frac{1}{2}-\alpha}-b\right) \times W_{f,f}\left(\frac{\frac{1}{2}+\alpha}{\frac{1}{2}-\alpha}a, \frac{\frac{1}{2}-\alpha}{\frac{1}{2}+\alpha}b\right) da\, db. \tag{73}$$

The way this formula is used is as follows. Assume we have an f, which can be thought of as a sum of functions each of which is "coherent" in the sense that its Wigner distribution is concentrated (and positive) in a rather small region of the phase-plane. Due to the presence of cross terms in $C^{(\alpha)}_{f,f}$, each pair of components of f will produce what is called a ghost. Although averages of these cross terms over sufficiently large regions are small, provided that the regions to which the Wigner distributions of the components are confined are more or less disjoint, the amplitude is not. Hence, negative values of Re $C^{(\alpha)}_{f,f}$ are likely to be found in the regions where the ghosts appear. What formula (73) tells us is how the regions where we can expect ghosts vary with α. For example, when $f = f_1 + f_2$ and W_{f_1,f_1}, W_{f_2,f_2} is concentrated around (q_1, p_1) and (q_2, p_2), respectively, one can show from (73) that $C^{(\alpha)}_{f,f}$ has a ghost around the point

$$(\tfrac{1}{2}(q_1+q_2), \tfrac{1}{2}(p_1+p_2)) + \alpha(q_1-q_2, p_2-p_1). \tag{74}$$

Hence, if one wants to construct examples $f = \Sigma_n f_n$, with each f_n "coherent" in the above sense, such that Re $C^{(\alpha)}_{f,f} \geq 0$ everywhere, one should take care that for each pair f_n, f_m producing their ghost according to (74), there is an f_k whose Wigner distribution is positive in the region where the ghost appears. It should be noted that all Re $C^{(\alpha)}_{f_k,f_k}$ tend to be concentrated and non-negative in approximately the same regions as W_{f_k,f_k}; this can be seen from formula (13), with $\Phi(\theta,\tau) = \cos \pi\alpha\theta\tau, \varphi(q,p) = \alpha^{-1}\cos \pi\alpha^{-1}qp$, which exhibits Re $C^{(\alpha)}_{f_k,f_k}$ as the convolution of W_{f_k,f_k} and a function φ, which is positive and slowly varying near (0,0) and rapidly oscillating far away from (0,0). In this way we achieve that the negative values of Re $C^{(\alpha)}_{f_n,f_m}$ are masked by the positive values of Re $C^{(\alpha)}_{f_k,f_k}$.

We start with the case $\alpha = \frac{1}{2}$. Although formula (73) degenerates in this case, the above noted principles are still valid. As is strongly suggested by the proof in Sec. III C, we should look for f's for which either f or F is supported by very small sets. We consider below f's that are supported by discrete sets.

Example 1: Let $f = \delta_a$, where $a \in \mathbb{R}$. Define for $b \in \mathbb{R}$

$$e_b(p) := \exp(-2\pi i b p), \quad p \in \mathbb{R}. \tag{75}$$

Then $F(p) = e_a(p)$, and

$$C^{(1/2)}_{f,f} = \delta_a \otimes e_0 \geq 0. \tag{76}$$

Example 2: Let $f = \delta_a + \delta_b$, where $a \in \mathbb{R}, b \in \mathbb{R}, a \neq b$. Then $F(p) = e_a(p) + e_b(p)$, and

$$\text{Re}[C^{(1/2)}_{f,f}(q,p)] = (1 + \cos 2\pi(a-b)p)(\delta_a(q) + \delta_b(q)), \quad q \in \mathbb{R}, \quad p \in \mathbb{R}, \tag{77}$$

which is non-negative everywhere. Note that the ghosts of δ_a and δ_b (whose Wigner distributions are $\delta_a \otimes e_0$ and $\delta_b \otimes e_0$, respectively) appear on the lines $q = a$ and $q = b$.

Example 3: Let $f = \Sigma^{\infty}_{n=-\infty} \delta_n$. Then $F(p) = \Sigma^{\infty}_{m=-\infty} \delta_m$, and thus

$$C^{(1/2)}_{f,f} = \sum_{n,m=-\infty}^{\infty} \delta_n \otimes \delta_m \geq 0. \tag{78}$$

Notes: (1) The following can be shown. When $f = \Sigma^{\infty}_{n=-\infty} c_n\delta_n$ and Re $C^{(1/2)}_{f,f} \geq 0$, then either (a) infinitely many of the c's are $\neq 0$ (b) only one c is $\neq 0$, or (c) only two c's $\neq 0$. In case (c) the two c's that are $\neq 0$ have equal modules. If F is smooth, only the last two options can occur.

(2) The only square-integrable states that have non-negative Wigner distributions are the Gaussians. When one passes from square integrable to generalized states, the situation remains the same, except that one has to allow certain degeneracies (delta functions and exponentials, cf. Ref. 20). Such a thing does not hold for the distributions Re $C^{(1/2)}_{f,f}$. A second deviation is found when one considers the behavior of the distributions under smoothing by means of Gaussians. It has been shown in Ref. 20 that a (generalized) function f for which $G_\gamma * W_{f,f} \geq 0$ everywhere must be a (degenerate) Gaussian when $\gamma > 1$. Here

$$G_\gamma(q,p) = \exp(-2\pi\gamma(q^2+p^2)), \quad q \in \mathbb{R}, \quad p \in \mathbb{R}. \tag{79}$$

When we consider as an example $f = \delta_a + z\delta_b$, where $a \in \mathbb{R}$, $b \in \mathbb{R}, z \in \mathbb{R}, z > 1$, we have

$$\text{Re}[C^{(1/2)}_{f,f}(q,p)] = \delta_a(q)(1 + z\cos 2\pi(a-b)p) + z\delta_b(q)(z + \cos 2\pi(a-b)p). \tag{80}$$

The second term at the right-hand side of (80) is non-negative; the convolution of the first term with G equals

$$\frac{1}{\sqrt{2\gamma}}\exp(-2\pi\gamma q^2) \times\left(1 + z\exp\left(-\frac{\pi}{2\gamma}(a-b)^2\right)\cos 2\pi(a-b)p\right) \tag{81}$$

and is non-negative everywhere when $z\exp(-(\pi/2\gamma)(a-b)^2) \leq 1$. Hence, when $|b-a|$ is sufficiently large, a small amount of smoothing will turn Re $C^{(1/2)}_{f,f}$ into an everywhere non-negative distribution.

We finally consider some examples with $\alpha = -k + \frac{1}{2}$, where k is an integer $\neq 0,1$ and f is a sum of delta functions. We note that the ghosts of δ_a and δ_b manifest themselves in $C^{(\alpha)}_{f,f}$ on the lines $q = a + k(b-a), q = b - k(b-a)$. Hence, we must consider sums consisting of either one or infinitely many terms.

Example 4: Let $f = \delta_a$, where $a \in \mathbb{R}$. Then

$$C^{(\alpha)}_{f,f} = \delta_a \otimes e_0 \geq 0. \tag{82}$$

Example 5: When $f = \Sigma^{\infty}_{n=-\infty} \delta_n$, then

$$C^{(\alpha)}_{f,f} = \sum_{n,m=-\infty}^{\infty} \delta_n \otimes \delta_m \geq 0. \tag{83}$$

Example 6: When $f = \Sigma^{\infty}_{n=-\infty}(\delta_{nk} + \delta_{nk+1})$, we have

$$C^{(\alpha)}_{f,f}(q,p) = k^{-1}\sum_{n,m=-\infty}^{\infty}[\delta_{nk^{-1}}(p)\delta_{mk}(q)(1+e^{2\pi i q}) + \delta_{nk^{-1}}(p)\delta_{mk+1}(q)(1+e^{-2\pi i q})], \tag{84}$$

and the real part of this distribution equals

$$k^{-1}\sum_{n,m=-\infty}^{\infty}\delta_{nk^{-1}}(p)(1+\cos 2\pi q) \times[\delta_{mk}(q) + \delta_{mk+1}(q)] \geq 0. \tag{85}$$

Note that when $V := \{nk + l \,|\, l = 0,1; n \in \mathbb{Z}\}$, we have $a \in V$,

$b \in V$, $a + k(b-a) \in V$, $b - k(b-a) \in V$. It can furthermore be shown that $\operatorname{Re} C_{g,g}^{(\alpha)}$ takes negative values when $g = \Sigma_{n=-\infty}^{\infty} (\delta_{nk-1} + \delta_{nk} + \delta_{nk+1})$.

ACKNOWLEDGMENT

The author thanks T. A. C. M. Claasen for stimulating discussions on the subject.

[1]We have $C_{af_1+bf_2, cg_1+dg_2} = a\bar{c}C_{f_1,g_1} + a\bar{d}C_{f_1,g_2} + b\bar{c}C_{f_2,g_1} + b\bar{d}C_{f_1,g_2}$.

[2]We designate states by lowercase symbols and their Fourier transforms by the corresponding capitals.

[3]E. P. Wigner, in *Perspectives in Quantum Theory*, edited by W. Yourgrau and A. van der Merwe (Dover, New York, 1979), Chap. 4.

[4]R. L. Hudson, Rep. Math. Phys. **6**, 249 (1974).

[5]E. Wigner, Phys. Rev. **40**, 749 (1932).

[6]L. Cohen, J. Math. Phys. 7, 781 (1966).

[7]T. A. C. M. Claasen and W. F. G. Mecklenbräuker, Philips J. Res. **35**, 217, 276, 372 (1980).

[8]A. J. E. M. Janssen, J. Math. Phys. **25**, 2240 (1984). We write in the present paper $C_{f,f}^{(\varphi)}$ instead of $C_f^{(\varphi)}$ to accommodate some of our proofs notationally, and to better emphasize the bilinear dependence on the state f.

[9]A. W. Rihaczek, IEEE Trans. Inf. Theory **IT-14**, 369 (1968).

[10]H. Margenau and R. Hill, Prog. Theor. Phys. (Kyoto) **26**, 722 (1961).

[11]See Refs. 3, 7, and 8 and L. Cohen, J. Math. Phys. **17**, 1863 (1976).

[12]A. J. E. M. Janssen and S. Zelditch, Trans. Am. Math. Soc. **280**, 563 (1983).

[13]N. G. de Bruijn, Nieuw Arch. Wiskunde **21**, 205 (1973).

[14]This can be deduced from the equality of the left-hand sides of (21) and (24) in A. J. E. M. Janssen, Philips J. Res. **37**, 79 (1982), Sec. 3. Also, see J. G. Krüger and A. Poffyn, Physica A **85**, 84 (1976), Sec. 12.

[15]This follows from the fact that for any $f \in L^2(\mathbf{R})$ and any $\epsilon > 0$ one can find a step function s with $\|f - s\| < \epsilon$. Now $C_{s,s}^{(\varphi)}(q,p) = (s, M(q,p)s)$ is uniformly continuous by (17) and boundedness of $M(0,0)$ and $|C_{s,s}^{(\varphi)}(q,p) - C_{f,f}^{(\varphi)}(q,p)| < \|M(0,0)\|\epsilon(\|f\| + \|s\|)$ for all $q \in \mathbf{R}$, $p \in \mathbf{R}$. Here $\|\ \|$ denotes the $L^2(\mathbf{R})$ norm.

[16]It can be shown that $C_{f,f}^{(\varphi)}$ is real valued for all $f \in L^2(\mathbf{R})$ if and only if φ is real. This is done by choosing $f(q) = 2^{1/4}\exp(-\pi q^2)$ so that $W_{f,f}(q,p) = 2\exp(-2\pi(q^2+p^2))$ and observing that $\psi * W_{f,f} = 0 \Leftrightarrow \psi = 0$. The distributions $C_{f,f}^{(\alpha)}$ with $\alpha \neq 0$ are, in general, complex valued. We consider therefore $\operatorname{Re} C_{f,f}^{(\alpha)}$, which can be brought into the form (7) by taking $\Phi(\theta,\tau) = \cos \pi\alpha\theta\tau$, i.e., $\varphi(q,p) = \alpha^{-1}\cos \pi\alpha^{-1}qp$.

[17]See, e.g., Refs. 7–9.

[18]We use here that $C_{f,g}^{(\varphi)}(q,p) = \overline{C_{g,f}^{(\varphi)}(q,p)}$. This follows from the fact that $W_{f,g}(q,p) = \overline{W_{g,f}(q,p)}$, formula (13), and the fact that φ is real by assumption.

[19]See the paper quoted in Ref. 14.

[20]A. J. E. M. Janssen, SIAM J. Math. Anal. **15**, 170 (1984).

[21]When $f: \mathbf{R} \to \mathbf{C}$ is locally integrable we say that $q \in \mathbf{R}$ is a Lebesgue point of f when

$$\lim_{\epsilon \downarrow 0} \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} |f(q+u) - f(q)|\, du = 0\,.$$

The set of $q \in \mathbf{R}$ that are not Lebesgue points of f has zero measure. When q is a Lebesgue point of f we have

$$\lim_{\epsilon \downarrow 0} \frac{1}{2\epsilon} \mu(\{u \in [-\epsilon,\epsilon] \,|\, |f(q+u) - f(q)| < \delta\}) = 1$$

for every $\delta > 0$. Here μ is ordinary Lebesgue measure.

Proof of the Levinson theorem by the Sturm–Liouville theorem

Zhong-Qi Ma
Institute for Theoretical Physics, State University of New York at Stony Brook, Stony Brook, New York 11794

(Received 19 January 1985; accepted for publication 12 April 1985)

The Levinson theorem is proved by the Sturm–Liouville theorem in this paper. For the potential $\int_0^1 r|V(r)|dr < \infty, V(r) \to b/r^2$ when $r \to \infty$, the modified Levinson theorem is derived as $n_l = (1/\pi)\delta_l(0) + (a-l)/2 - \frac{1}{2}\sin^2\{\delta_l(0) + [(a-l)/2]\pi\}$, if $a(a+1) \equiv b + l(l+1) > \frac{3}{4}$ or $a = 0$. Two examples which violate the Levinson theorem and satisfy the modified Levinson theorem are discussed.

I. INTRODUCTION

Consider the Schrödinger equation ($2m = 1, \hbar = 1$)

$$-\frac{d^2u(r)}{dr^2} = [E - V(r) - l(l+1)/r^2]u(r), \tag{1}$$

with the cutoff potential

$$\int_0^1 r|V(r)|\mathrm{d}r < \infty, \quad V(r) = 0, \quad \text{when } r \geqslant r_0. \tag{2}$$

If $E \leqslant 0$, physically admissible solutions should be vanishing at the origin and at spatial infinity. One can solve Eq. (1) in the two regions $0 \leqslant r \leqslant r_0$ and $r_0 \leqslant r < \infty$, respectively, and can obtain the two logarithmic derivatives of wave functions, u'/u, at $r = r_0$, from the different sides. If these two logarithmic derivatives match at a definite energy $E < 0$, there is a bound state with this energy. The Sturm–Liouville theorem shows that the logarithmic derivative is monotonic with respect to the energy, therefore, the relation between two logarithmic derivatives at zero energy determines whether there are bound states or not. On the other hand, the logarithmic derivative for $E \gtrsim 0$ determines the phase shift at zero energy $\delta_l(0)$. The Levinson theorem[1] reveals the relation between $\delta_l(0)$ and the number n_l of bound states, which, obviously, relates with the Sturm–Liouville theorem closely. Even though the Levinson theorem has been proved in different methods[2] and generalized to different potentials and problems,[3] the proof method of the Levinson theorem by the Sturm–Liouville theorem is very simple, intuitive, rigorous, and easy to generalize.

The previous proof of Levinson theorem requires the potential $V(r)$ to satisfy the following conditions:

$$\int_0^1 r|V(r)|dr < \infty, \tag{3a}$$

$$\int_1^\infty r^2|V(r)|dr < \infty. \tag{3b}$$

The first one is necessary for the nice behavior of the wave function at the origin, and the second one is necessary for the previous proof method. Newton[4] pointed out two examples where the Levinson theorem is violated because the second condition is not satisfied.

It is proved in terms of the Sturm–Liouville theorem in this paper that for the potential $V(r)$

$$\int_0^1 r|V(r)|\mathrm{d}r < \infty, \quad V(r) \to \frac{b}{r^2}, \quad \text{when } r \to \infty, \tag{4}$$

and

$$b + l(l+1) = a(a+1) > \tfrac{3}{4} \quad \text{or } a = 0, \tag{5}$$

the modified Levinson theorem is as follows:

$$n_l = \frac{1}{\pi}\delta_l(0) + \frac{a-l}{2} - \tfrac{1}{2}\sin^2\left(\delta_l(0) + \frac{a-l}{2}\pi\right), \tag{6}$$

where n_l is the number of bound states with angular momentum l, and $\delta_l(0)$ is the phase shift of zero energy. Newton's examples satisfy the modified Levinson theorem.

II. PROOF OF THE LEVINSON THEOREM BY THE STURM–LIOUVILLE THEOREM

When $V = 0$, the solution for the Schrödinger equation (1) with $E > 0$ is

$$u = \cos\delta_l \hat{j}_l(kr) - \sin\delta_l \hat{n}_l(kr), \ k = \sqrt{E}, \tag{7}$$

where δ_l is an integral constant and $\hat{j}_l$ and $\hat{n}_l$ are the spherical Bessel functions

$$\hat{j}_l(x) = \sqrt{\pi x/2}J_{l+1/2}(x), \ \hat{n}_l(x) = \sqrt{\pi x/2}N_{l+1/2}(x), \tag{8}$$

with the asymptotic behaviors

$$\hat{j}_l(x) \to \begin{cases} x^{l+1}/(2l+1)!!, & \text{when } x \to 0, \\ \sin(x - l\pi/2), & \text{when } x \to \infty, \end{cases} \tag{9a}$$

$$\hat{n}_l(x) \to \begin{cases} -(2l-1)!!/x^l, & \text{when } x \to 0, \\ -\cos(x - l\pi/2), & \text{when } x \to \infty. \end{cases} \tag{9b}$$

Therefore, the solution with $\delta_l = 0$ is vanishing at the origin and admissible in physics. The logarithmic derivative at $r = r_0$ for this solution is

$$\frac{u'(r_0)}{u(r_0)} = k\frac{\hat{j}_l'(kr_0)}{\hat{j}_l(kr_0)} \xrightarrow{k\to 0} \frac{l+1}{r_0}. \tag{10}$$

On the other hand, when k goes to zero, the finite solution r^{-l} at $r \geqslant r_0$ is the limit of $\hat{n}_l(kr)$, and its logarithmic derivative at $r = r_0$ is

$$k\,[\hat{n}_l'(kr_0)/\hat{n}_l(kr_0)] \xrightarrow{k\to 0} -l/r_0. \tag{11}$$

Now, introduce a parameter λ for a given cutoff potential (2)

$$V(r,\lambda) = \begin{cases} \lambda V(r), & \text{if } V(r) > 0 \text{ and } 0 \leqslant \lambda \leqslant 1, \\ V(r), & \text{if } V(r) > 0 \text{ and } 1 \leqslant \lambda \leqslant 2, \\ 0, & \text{if } V(r) \leqslant 0 \text{ and } 0 \leqslant \lambda \leqslant 1, \\ (\lambda - 1)V(r), & \text{if } V(r) \leqslant 0 \text{ and } 1 \leqslant \lambda \leqslant 2. \end{cases} \tag{12}$$

 0022-2488/85/081995-05$02.50

When $\lambda = 0$, $V(r,0) = 0$; when λ increases from 0 to 1, $V(r,\lambda)$ increases; when λ increases from 1 to 2, $V(r,\lambda)$ decreases and reaches to $V(r)$. Suppose that the nontrivial solution $u_{\lambda E}(r)$ in the region $0 \leqslant r \leqslant r_0$ is obtained for this potential $V(r,\lambda)$, and then, the logarithmic derivative can be calculated as

$$[u'_{\lambda E}(r)/u_{\lambda E}(r)]|_{r_0 -} = A(\lambda,E), \tag{13}$$

where A is the function of λ and E. When $E>0$, by the matching condition at $r = r_0$,

$$[u'(r)/u(r)]|_{r_0 -} = [u'(r)/u(r)]|_{r_0 +}, \tag{14}$$

the potential determines the constant δ_l through the logarithmic derivative A:

$$\tan \delta_l = \frac{\hat{j}_l(kr_0)}{\hat{n}_l(kr_0)} \frac{A - k\hat{j}'_l(kr_0)/\hat{j}_l(kr_0)}{A - k\hat{n}'_l(kr_0)/\hat{n}_l(kr_0)}. \tag{15a}$$

When k tends to zero ($E \gtrsim 0$),

$$\tan \delta_l \simeq - \frac{(kr_0)^{2l+1}}{(2l+1)!!(2l-1)!!} \times \frac{A - (l+1)/r_0}{A - [-l/r_0 + k^2 r_0/(2l-1)]}. \tag{15b}$$

Here, $\delta_l(k)$ is called the phase shift which describes the properties of the potential $V(r)$ (see Ref. 1). When $V=0$, $A = k\hat{j}'_l(kr_0)/\hat{j}_l(kr_0)$, so $\delta_l(k) = m\pi$. We use the convention in this paper that

$$\delta_l(k) = 0, \quad \text{when} \quad V(r) = 0. \tag{16}$$

It is shown[5] that in this convention the phase shift $\delta_l(\infty)$ for the infinity energy vanishes.

From Eq. (15) we get

$$\frac{\partial \delta_l}{\partial A} = - \frac{k \cos^2 \delta_l}{[A\hat{n}_l(kr_0) - k\hat{n}'_l(kr_0)]^2} < 0. \tag{17}$$

The phase shift increases monotonically as the logarithmic derivative A decreases.

When $E<0$, only one independent solution of Eq. (1) is vanishing at the spatial infinity

$$u = e^{i(\pi/2)(l+1)} \hat{h}_l(ik_1 r) \xrightarrow{r\to\infty} e^{-k_1 r}, \tag{18}$$

where $k_1 = \sqrt{-E}$ and $\hat{h}_l$ is the spherical Hankel function

$$\hat{h}_l(x) = \hat{j}_l(x) + i\hat{n}_l(x). \tag{19}$$

The logarithmic derivative of this solution at $r = r_0$ is

$$\frac{u'(r)}{u(r)}\Big|_{r_0 +} = ik_1 \frac{\hat{h}'_l(ik_1 r_0)}{\hat{h}_l(ik_1 r_0)} = \begin{cases} -l/r_0, & \text{when } k_1 + 0, \\ -k_1, & \text{when } k_1 \to \infty, \end{cases} \tag{20}$$

which decreases monotonically from $-l/r_0$ to negative infinity as energy decreases from zero to negative infinity.

In the region $0 \leqslant r \leqslant r_0$ the solution with $u(0) = 0$ for $E<0$ when $V(r) = 0$ is

$$u = e^{-i(\pi/2)(l+1)} \hat{j}_l(ik_1 r), \tag{21}$$

and

$$\frac{u'(r)}{u(r)}\Big|_{r_0 -} = ik_1 \frac{\hat{j}'_l(ik_1 r_0)}{\hat{j}_l(ik_1 r_0)} = \begin{cases} (l+1)/r_0, & \text{when } k_1 \to 0, \\ k_1, & \text{when } k_1 \to \infty. \end{cases} \tag{22}$$

The logarithmic derivative $A(0,E)$ increases monotonically from $(l+1)/r_0$ to positive infinity as energy decreases from zero to negative infinity.

Now, we turn to the Sturm–Liouville theorem. If u with $u(0) = 0$ is a nontrivial solution of Eq. (1) with $E - V(r) = \mu p(r) + q(r)$, where $p(r)$ does not change sign in the region $0 \leqslant r \leqslant r_0$ and μ may be understood as E or λ (it should be discussed separately when λ increases from 0 to 1 and from 1 to 2), then

$$u^2(r_0) \frac{d(u'/u)}{d\mu}\Big|_{r_0 -} = - \int_0^{r_0} p(r) u^2(r) dr. \tag{23}$$

If we fix the potential $V(r,\lambda)$, the logarithmic derivative $A(\lambda,E)$ increases monotonically as energy E decreases. Equation (22) is an example. If we fix energy E, $A(\lambda,E)$ increases monotonically as λ increases from 0 to 1, and decreases monotonically as λ increases from 1 to 2. If we apply the Sturm–Liouville theorem to the region $r_0 \leqslant r < \infty$, $(u'/u)|_{r_0 +}$ also changes monotonically with respect to the energy if $u(\infty) = 0$. Equation (20) gives an example.

There is another form of the Sturm–Liouville theorem called the comparison theorem. In our problem, for the nontrivial solution u of Eq. (1) with $u(0) = 0$, its zeros in the region $0 < r \leqslant r_0$ move away from the origin if E decreases or V increases (λ goes from 0 to 1), and move towards the origin if E increases or V decreases (λ goes from 1 to 2).

It is easy to see that $u(r_0) = 0$ will cause $A(\lambda,E)$ divergent. However, there are no zeros in the region $0 < r \leqslant r_0$ for the nontrivial solution u with $u(0) = 0$ if $E \leqslant 0$ and $V = 0$ [see Eq. (21)]. Then, according to the comparison theorem, there are no zeros in the region $0 < r \leqslant r_0$ for u with $u(0) = 0$ if $E \leqslant 0$ and $V \geqslant 0$ $(0 \leqslant \lambda \leqslant 1)$, namely, when $E \leqslant 0$ and λ increases from 0 to 1, $A(\lambda,E)$ increases but is finite. This conclusion also holds when $E>0$ but is small enough. In contrast, if one fixes E and increases λ from 1 to 2, $u(r_0)$ may be equal to zero and $A(\lambda,E)$ may decrease to negative infinity $[u(r_0) = 0]$, then jump to positive infinity and decrease again. Generally, $A(\lambda,E)$ may jump several times as λ increases from 1 to 2.

A potential is said to be repulsive dominated if $A(\lambda,0) \geqslant (l+1)/r_0$, $0 \leqslant \lambda \leqslant 2$. For a repulsive-dominated potential where

$$\frac{l+1}{r_0} \leqslant \frac{u'(r)}{u(r)}\Big|_{r_0 -} < \infty, \quad \text{when} - \infty < E \leqslant 0, \tag{24}$$

according to Eqs. (20) and (15b), there is no bound state in this case, and $\delta_l(0) = 0$. Therefore,

$$n_l = (1/\pi)\delta_l(0) = 0. \tag{25}$$

A potential is said to be attractive dominated if $A(\lambda,0)$ may be less than $(l+1)/r_0$ for some λ. If one fixes a very small k ($E \lesssim 0$) and increases λ from 1 to 2, the logarithmic derivative $A(\lambda,E)$ decreases monotonically from a positive value $\geqslant (l+1)/r_0$, and phase shift $\delta_l(k)$ increases. However, before $A(\lambda,E)$ reaches to $-l/r_0$, from Eq. (15), $\delta_l(k)$ is still near zero. If the potential is attractive dominated, $A(\lambda,E)$

may decrease through the value $-l/r_0$, and $\delta_l(k)$ increases rapidly through $\pi/2$ and goes to near π. Then, $\delta_l(n)$ increases slowly near π as $A(\lambda,E)$ decreases to negative infinity $[u(r_0)=0]$, jumps to positive infinity, and decreases again. At the same time, wave function u gets a zero in the region $0<r\leqslant r_0$. When $A(\lambda,E)$ decreases the second time through $-l/r_0$, $\delta_l(k)$ increases rapidly again, and goes through $3\pi/2$ to 2π. Generally, when λ increases from 1 to 2, $A(\lambda,E)$ decreases through m times of jumping from negative infinity to positive infinity and $n_l=m$ or $m+1$ times through the value $-l/r_0$, correspondently, $\delta_l(k)$ increases to near $n_l\pi$ and the wave function $u(r)$ has m zeros in the region $0<r\leqslant r_0$.

Now, we fix the potential $V(r)=V(r,2)$ and let the energy decrease from the small positive value. According to the Sturm–Liouville theorem all the zeros of the solution u in the region $0<r\leqslant r_0$ will move away from the origin and the logarithmic derivative $A(2,E)$ will increase monotonically. As long as a zero moves through r_0, $A(2,E)$ increases to positive infinity, jumps to negative infinity, and increases again. Therefore, when E decreases from zero to negative infinity, $A(2,E)$ increases from $A(2,0)$ through m times of jumping and n_l times through the value $-l/r_0$, and finally increases to positive infinity. Thus, from Eq. (20) $A(2,E)$ will match n_l times with the values of $[u'(r)/u(r)]|_{r_0+}$ when E decreases from zero to negative infinity, namely, there are n_l bound states

$$n_l=\delta_l(0)/\pi. \tag{26}$$

The above discussion is the essence of the Levinson theorem.

We should pay some attention to the case where $A(2,0)=-l/r_0$. In this case there is a solution with $E=0$ satisfying the matching condition (14). This solution is in proportion to r^{-l} in the region $r>r_0$, so it is a bound state except for $l=0$. It is interesting to know whether $\delta_l(0)$ increases an additional π when $A(\lambda,0)$ finally reaches the value $-l/r_0$. The crucial point is that when $\lambda=2$ and k is small enough, $\tan\delta_l(k)$ is positive or negative and tends to zero or infinity as k goes to zero. If $\tan\delta_l(k)$ is negative and tends to zero, $\delta_l(0)$ increases an additional π. If $\tan\delta_l(k)$ is positive and tends to zero, $\delta_l(0)$ does not increase a π. If $\tan\delta_l(k)$ tends to infinity, $\delta_l(0)$ increases $\pi/2$.

Since $u(r_0)\neq 0$, from Eq. (23) we obtain

$$\frac{\partial A(2,E)}{\partial E}=-\frac{1}{u^2(r_0)}\int_0^{r_0}u^2(r)dr<0,$$
$$A(2,E)\simeq -l/r_0-c^2k^2,\quad \text{when } E\gtrsim 0, \tag{27}$$

where $c^2>0$. Substituting it into Eq. (15b) we obtain

$$\tan\delta_l(k)\simeq -c'^2k^{2l-1}, \tag{28}$$

where $c'^2>0$. Therefore, when $l\neq 0$, Eq. (26) holds. When $l=0$, $\hat{j}_0(kr_0)=\sin kr_0$, $\hat{n}_0(kr_0)=-\cos kr_0$,

$$\tan\delta_0(k)=-\tan kr_0\frac{A-k\cot kr_0}{A+k\tan kr_0}\to\infty, \tag{29}$$

that is, $\delta_0(0)$ increases an additional $\pi/2$ when $A(\lambda,0)$ finally reaches to 0. Since this state for $E=0$ and $l=0$ is not a bound state, but, so called, a half-bound state, the additional $\pi/2$ should be subtracted from $\delta_0(0)$ in Eq. (26). Finally, the Levinson theorem should be read as

$$n_l=(1/\pi)\delta_l(0)-\tfrac{1}{2}\sin^2\delta_l(0). \tag{30}$$

This form of Levinson's theorem was first presented by Ni.[2]

This proof method of the Levinson theorem for the Schrödinger equation can easily be generalized into that for the Dirac equation, which will be discussed elsewhere, and obtain the same result as in Ref. 6.

III. MODIFIED LEVINSON THEOREM

Now, we turn to the case where the potential has a tail at $r\geqslant r_0$:

$$V(r)\sim br^{-m},\quad \text{when } r\to\infty. \tag{31}$$

Let r_0 be so large that only the leading term in $V(r)$ is concerned in the region $r\geqslant r_0$. Substituting it into Eq. (1) and changing the variable r to $\xi=kr$, we get

$$-\frac{d^2u(\xi)}{d\xi^2}=\left(1-\frac{b}{\xi^m}k^{m-2}-\frac{l(l+1)}{\xi^2}\right)u(\xi),\quad r\geqslant r_0. \tag{32}$$

As far as the Levinson theorem is concerned, only those solutions with small k are interesting to us. If $m=1$, the potential term becomes very large and, as is well known, the phase shift changes logarithmically. If $m\geqslant 3$, the potential term becomes too small to affect the phase shift. Thus, we are going to discuss the case where $m=2$. For convenience, we divide the potential into two parts:

$$V(r)=V_1(r)+V_2(r), \tag{33a}$$

$$V_1(r)=\begin{cases}V(r), & r\leqslant r_0,\\ 0, & r>r_0,\end{cases} \tag{33b}$$

$$V_2(r)=\begin{cases}0, & r<r_0,\\ V(r)\sim br^{-2}, & r\geqslant r_0.\end{cases} \tag{33c}$$

Substituting it into the Schrödinger equation (1), we get

$$-\frac{d^2u(r)}{dr^2}=\left(E-\frac{a(a+1)}{r^2}\right)u(r),\quad r>r_0, \tag{34}$$

where

$$a(a+1)\equiv b+l(l+1). \tag{35}$$

If $b+l(l+1)<-\frac{1}{4}$, this potential causes an infinite number of bound states, which is not interesting to us. For $a(a+1)>-\frac{1}{4}$, $a>-\frac{1}{2}$, the solution to Eq. (34) is as follows: When $E>0$, $k=\sqrt{E}$,

$$u=\cos\delta_a\hat{j}_a(kr)-\sin\delta_a\hat{n}_a(kr), \tag{36}$$

where $\delta_a(k)$ is related with the phase shift $\delta_l(k)$ by

$$\delta_a(k)=\delta_l(k)+[(a-l)/2]\pi; \tag{37}$$

When $E=0$,

$$u=r^{-a}. \tag{38}$$

When $E<0$, $k_1=\sqrt{-E}$,

$$u=e^{i(\pi/2)(a+1)}\hat{h}_a(ik_1r). \tag{39}$$

Like Eq. (15), we get

$$\tan\delta_a(k)=\frac{\hat{j}_a(kr_0)}{\hat{n}_a(kr_0)}\,\frac{A-k\hat{j}_a'(kr_0)/\hat{j}_a(kr_0)}{A-k\hat{n}_a'(kr_0)/\hat{n}_a(kr_0)}, \tag{40}$$

and

$$\frac{\partial \delta_a}{\partial A} = -\frac{k\cos^2\delta_a}{[A\hat{n}_a(kr_0) - k\hat{n}'_a(kr_0)]^2}. \tag{41}$$

Now, we can follow the proof for the Levinson theorem (30) in Sec. II to obtain the modified Levinson theorem (6) where the potential $V_1(r)$ plays the role of the potential (2). The only thing which should be discussed carefully is the case when $A(2,0) = -a/r_0$.

If $a > \frac{1}{2}$ and $A(2,0) = -a/r_0$, the solution for $E = 0$ is a bound state with the behavior $u(r) \sim r^{-a}$, when $r \to \infty$. When $A(2,0) = -a/r_0$, just like Eq. (28), we get

$$\tan\delta_a(k) \sim -C'^2 k^{2a-1}, \tag{42}$$

where $C'^2 > 0$. Thus, $\delta_a(0)$ increases an additional π, so we come to the modified Levinson theorem (6).

If $a = \frac{1}{2}$ and $A(2,0) = -a/r_0$, the solution for $E = 0$ is not a bound state because $\int_{r_0}^{\infty}[r^{-1/2}]^2\,dr$ is divergent. However,

$$\hat{j}_{1/2}(x) \sim \sqrt{\frac{\pi}{2}}\,\frac{x^{3/2}}{2},$$
$$\hat{n}_{1/2}(x) \sim -\sqrt{\frac{2}{\pi x}} + \frac{x^{3/2}}{\sqrt{2\pi}}\ln\frac{x}{2}.$$

Here, $\tan\delta_{1/2}(k)$ is negative when k is small enough and tends to zero as k goes to zero, that is, $\delta_{1/2}(0)$ increases an additional π, which violates the modified Levinson theorem (6).

If $a = 0$, the calculation is the same as that in Sec. II when $l = 0$, so Eq. (6) holds. If $-\frac{1}{2} < a < \frac{1}{2}$ but $a \neq 0$,

$$\hat{j}_a(kr_0) \sim \frac{\sqrt{\pi}}{\Gamma(\frac{3}{2}+a)}\left(\frac{kr_0}{2}\right)^{a+1},$$
$$\hat{n}_a(kr_0) \sim -\frac{\Gamma(\frac{1}{2}+a)}{\sqrt{\pi}}\left(\frac{2}{kr_0}\right)^{a} \times\left[1 - \left(\frac{kr_0}{2}\right)^{2a+1}\frac{\pi\cot(a+\frac{1}{2})\pi}{(\frac{1}{2}+a)[\Gamma(\frac{1}{2}+a)]^2}\right],$$
$$\tan\delta_a(k) \sim -\frac{\pi}{\Gamma(\frac{3}{2}+a)\Gamma(\frac{1}{2}+a)}\left(\frac{kr_0}{2}\right)^{2a+1} \times\frac{A-(a+1)/r_0}{A + a/r_0 + k(kr_0/2)^{2a}\pi\cot(a+\frac{1}{2})\pi/[\Gamma(a+\frac{1}{2})]^2}, \tag{43}$$

When $A(2,0) = -a/r_0$, $\tan\delta_a(k)$ tends to $\tan(a+\frac{1}{2})\pi$ as k goes to zero, that is, $\delta_a(0)$ increases an additional $(a+\frac{1}{2})\pi$, which violates the modified Levinson's theorem (6). For $\frac{1}{2} \geqslant a > -\frac{1}{2}$, the number of bound states n_l is equal to $\delta_l(0)/\pi + (a-l)/2$ except for the case with the "half-bound state," a nontrivial solution of zero energy with $u(0) = 0$ and $u(r) \sim r^{-a}$ at $r \to \infty$. In the latter case $n_l = \delta_l(0) - (a+l+1)/2$.

Now, we check the two examples given by Newton.[4]

Example 1:

$$V(r) = \frac{2c^2}{(1+cr)^2} \to \frac{2}{r^2}, \text{ when } r \to \infty. \tag{44}$$

The phase shift of the S wave is $\delta_0(0) = -\pi/2$ and there is no bound state for the S wave, $n_0 = 0$. However, since $a = 1$, Eq. (6) is satisfied.

Example 2:

$$V(r) = -6r\frac{2N^2 - r^3}{(N^2 + r^3)^2} \to \frac{6}{r^2}, \text{ when } r \to \infty, \tag{45}$$

where N is a constant. The S wave function is

$$u(k,r) = \frac{\sin kr}{k} - \frac{3r}{k^3(N^2+r^3)}(\sin kr - kr\cos kr) \xrightarrow{k\to 0} N^2 r/(N^2 + r^3). \tag{46}$$

The solution with $E = 0$ is a bound state; however, $\delta_0(0) = 0$. In this example, $a = 2$, $n_0 = 1$, and Eq. (6) is satisfied.

Finally, we would like to say some words about the condition (3a) of the potential at the origin which excludes the potential with the form

$$V(r) \sim b_1 r^{-m}, \; m \geqslant 2, \text{ when } r \to 0. \tag{47}$$

It is obvious that the origin is an irregular singular point if $m > 2$. How about $m = 2$ in which the origin is a regular singular point? In fact, the centrifugal potential has behavior like that.

It is necessary for the Sturm–Liouville theorem that

$$(uu_1' - u_1 u')|_{r=0} = 0. \tag{48}$$

If the potential has the behavior

$$V(r) \to b_1 r^{-2}, \text{ when } r \to 0, \tag{49}$$

at the origin, the indicial equation is

$$\alpha(\alpha - 1) = b_1 + l(l+1). \tag{50}$$

If $b_1 + l(l+1) < -\frac{1}{4}$, we have an oscillation solution, which should be excluded. If $b_1 + l(l+1) = a_1(a_1+1)$ and $a_1 \geqslant 0$, we have $\alpha = 1 + a_1 \geqslant 1$, so u tends to zero and u' is finite as r goes to zero. Therefore, the Levinson theorem holds for the potential

$$\lim r^2 V(r) = b_1 < \infty, \quad b_1 + l(l+1) \geqslant 0, \tag{51}$$

instead of condition (3a). If $-\frac{1}{2} < a_1 < 0$, u' is infinite, but uu_1' goes to zero as r goes to zero. The condition (48) excludes $a_1 = -\frac{1}{2}$. It seems to us that the Levinson theorem also holds for $-\frac{1}{2} < a_1 < 0$.

Note added in proof: After this paper was completed I read a new paper by Z. R. Iwinski, L. Rosenberg, and L. Spruch [Phys. Rev. A **31**, 1229 (1985)] in which Levinson's theorem for the Schrödinger equation with a short-range potential [$r^3 V(r) \to 0$ as $r \to \infty$] was also proved by the Sturm–Liouville theorem. In the present paper the case with the long-range potential $V(r) \to b/r^2$ as $r \to \infty$ is discussed. If a potential $c/(r^2 \ln r)$ is added to b/r^2, the phase shift will not be affected because this potential $|c|/(r^2\ln r)$ is less than $\epsilon(2a+1)/r^2$ with $\epsilon = |c|\{(2a+1)\ln r_0\}^{-1}$ at $r > r_0$ which changes the phase shift only by an infinitesimal value $\epsilon\pi/2$ for the large r_0. The author would like to thank Professor C. N. Yang for bringing his attention to this kind of potential with logarithm.

ACKNOWLEDGMENTS

The author would like to thank Professor Chen Ning Yang for suggesting looking into application of Sturm–Liouville theorem to Levinson's theorem, instructive discussions, and encouragement and Professor Roger G. Newton at Indiana University for discussions in private communication. It is a pleasure to thank Professor C. N. Yang, Professor H. T. Nieh, and the Institute for Theoretical Physics, SUNY at Stony Brook, for their warm hospitality.

The author is supported by a Fung King-Hey fellowship through the Committee for Educational Exchange with China at Stony Brook. This paper is supported in part by the National Science Foundation under Grant No. PHY 81-09110 A-03.

[1]N. Levinson, Kgl. Danske Videnskab. Selskab, Mat-fys Medd. **25**, No. 9 (1949).

[2]For example, J. M. Jauch, Helv. Phys. Acta **30**, 143 (1957); G. J. Ni, Phys. Energiae Fortis Phys. Nucl. **3**, 449 (1979).

[3]For example, for a noncentral potential, R. G. Newton, J. Math. Phys. **1**, 319 (1960); **18**, 1348 (1977); for nonlocal interaction, A. Martin, Nuovo Cimento **7**, 607 (1958); R. G. Newton, J. Math. Phys. **18**, 1582 (1977); for the three-body system, J. A. Wright, Phys. Rev. B **139**, 137 (1965); for Dirac particles, M-C. Barthélémy, Ann. Inst. H. Poincaré A **7**, 115 (1967); G. J. Ni, see Refs. 2 and 6.

[4]R. G. Newton, *Scattering Theory of Waves and Particles*, (Springer-Verlag, New York, 1982), 2nd. ed., p. 438–439.

[5]L. I. Schiff, *Quantum Mechanics* (McGraw–Hill, New York, 1968), 3rd ed, p. 349; F. Calogero, *Variable Phase Approach to Potential Scattering* (Academic, New York, 1967), p. 37.

[6]Z. Q. Ma and G. J. Ni, Phys. Rev. D **31**, 1482 (1985).

Tunneling through a singular potential barrier

J. Dittrich
Nuclear Physics Institute, Czechoslovak Academy of Sciences, 250 68 Řež, Czechoslovakia

P. Exner[a)]
Nuclear Centre, Charles University, 180 00 Prague, Czechoslovakia

(Received 31 August 1984; accepted for publication 1 February 1985)

Quantum tunneling of a nonrelativistic particle through a singular potential barrier V is studied on the line. The Hamiltonian is a self-adjoint extension of the operator $H_1 = -d^2/dx^2 + V(x)$. If H_1 is essentially self-adjoint on its natural domain, the tunneling is forbidden for a class of potentials that includes all semiclassically impenetrable barriers. If H_1 is not essentially self-adjoint, the Friedrichs extension of H_1 yields no tunneling for another class of potentials which again includes the semiclassically impenetrable ones. In general, the occurrence of tunneling is not excluded and depends on the self-adjoint extension we choose as the Hamiltonian of our problem. As an example, we evaluate the transmission coefficient for all self-adjoint extensions of the operator H_1 referring to $V(x) = gx^{-2}$ with $0 < g < \frac{3}{4}$.

I. INTRODUCTION

Quantum tunneling is a phenomenon with a vast range of occurrence in diverse branches of physics. One of the particularly interesting cases concerns the problem of conservation of topological charges in various field-theoretical models. It is believed on the basis of heuristic arguments[1] that transitions between states of different topological charges are tunnelings through an infinitely high energy barrier; rigorous proofs of this assertion were given[2] for the $(2+1)$-dimensional $O(3)$ σ model, the $(2+1)$-dimensional electrodynamics, and the $(3+1)$-dimensional Yang–Mills–Higgs theory. Such a tunneling is forbidden semiclassically since the Euclidean action along any path crossing the barrier is infinite. The question naturally arises, whether the transitions are forbidden exactly, too.

Motivated by this problem, we address ourselves in this paper to a considerably simpler question: we are going to study tunneling through an infinitely high potential barrier V in one-dimensional quantum mechanics, particularly in the situations when the transition is semiclassically forbidden. There is no tunneling, of course, when the infinitely high potential wall has a nonzero thickness. On the other hand, the answer is not *a priori* clear if $V(x)$ is finite almost everywhere with exception of one or more point singularities. For simplicity, we restrict our attention to the potentials that have just one point singularity placed at $x = 0$. The main result of the paper is two conditions, namely,

$$\int_{-c}^{c} V(x)\,dx = \infty \tag{1.1}$$

and

$$\int_{-c}^{c} x^2 V(x)^2\,dx = \infty \tag{1.2}$$

for some $c > 0$, which are sufficient for the absence of tunneling provided the dynamics is determined by the potential V alone, or more explicitly, provided the formal Hamiltonian

$$H_1 = -\frac{d^2}{dx^2} + V(x) \tag{1.3}$$

is essentially self-adjoint (e.s.a.) on its natural domain. (A more detailed specification will be given in Sec. II below.)

In the opposite case, one must choose a suitable self-adjoint extension of H_1 which is to play the role of the Hamiltonian. Needless to say, this choice should be based on additional physical considerations; roughly speaking, one must specify what happens when the particle reaches the center of repulsion. A particular interest concerns the case when H_F, the extension in the sense of quadratic forms, is chosen for the Hamiltonian. In that case, the tunneling is forbidden under the condition (1.1). It should be noted, that both the conditions (1.1) and (1.2) are fulfilled if V is a semiclassically impenetrable barrier, i.e.,

$$\int_{\mathbb{R}} (V(x) - E)_+^{1/2} = \infty, \quad E > 0. \tag{1.4}$$

For other self-adjoint extensions, however, conditions of this type might not ensure absence of the tunneling. In order to illustrate this fact, we discuss in detail the example of the barrier

$$V(x) = gx^{-2}, \quad g > 0. \tag{1.5}$$

If $g < \frac{3}{4}$, the corresponding operator H_1 is not essentially self-adjoint and has a family of self-adjoint extensions parametrized by 2×2 unitary matrices; they can be easily constructed by the standard von Neumann method.[3,4] We evaluate the transmission coefficient for each of the extensions; it appears that the tunneling is forbidden iff the corresponding matrix is diagonal.

There are many papers in which the motion in the field of a singular potential is treated (on various levels of mathematical rigor); see, e.g., Refs. 5–10 and further references given therein. Most of them, however, concern potentials in $\mathbb{R}^3$ with a singularity at the origin. The one-dimensional tunneling discussed here has not been, up to our knowledge, considered earlier, though there naturally are some similarities to other works, especially in the example treated in Secs. IV and V. It should be stressed also, that the assumption

[a)] Present address: Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, 141980 Dubna, USSR.

 0022-2488/85/082000-09$02.50

about just one singularity in the potential V is not vital for our main results. The methods used to prove absence of the tunneling under the conditions (1.1) and (1.2) employ substantially the local behavior of $V(\cdot)$ only around the singularity, and therefore they are expected to work for potentials V having a (finite or infinite) set of point singularities with no accumulation points.

II. FORMULATION OF THE PROBLEM

As mentioned above, we shall concentrate on the case of a non-negative barrier with one point singularity. Hence we adopt the following assumptions.

(a) V is Lebesgue measurable and $V(x) \geqslant 0$ a.e. in $\mathbb{R}$.

(b) $V(\cdot)$ is bounded a.e. in $\mathbb{R} \setminus [-\eta,\eta]$ for each $\eta > 0$.

The Hamiltonian of the problem is, of course, a suitable self-adjoint extension of the symmetric operator

$$H_1 = H_0 + V, \tag{2.1}$$

with the domain $D(H_1) = D(H_0) \cap D(V)$, where $H_0\psi = -\psi''$ with $D(H_0) = \mathrm{AC}^2[\mathbb{R}]$ (cf. Ref. 3, Sec. X.1) and $(V\psi)(x) = V(x)\psi(x)$. Recall that $\mathrm{AC}[M]$ denotes the set of functions $f \in L^2(M)$ that are absolutely continuous in (each compact subinterval of) M with $f' \in L^2(M)$ and $\mathrm{AC}^2[M] = \{f \in \mathrm{AC}[M] : f' \in \mathrm{AC}[M]\}$. We are particularly interested in the self-adjoint extension associated with the quadratic form

$$h : h(\psi) = \|\psi'\|^2 + \|V^{1/2}\psi\|^2, \tag{2.2}$$

defined on $Q(h) = \mathrm{AC}[\mathbb{R}] \cap D(V^{1/2})$. Since V is locally integrable in $\mathbb{R} \setminus \{0\}$ and non-negative, the form h is closed and semibounded (Ref. 11, Theorems 14.1.1–14.1.3). Consequently, there is a unique self-adjoint operator H_F associated with h, the Friedrichs extension of H_1.

In spite of possible singularity at $x = 0$, the scattering problem for H_F is well posed if only V decays rapidly enough at infinity (for general information about the rigorous scattering theory see Ref. 3, Chap. XI, or Ref. 12). To be specific, we assume the following.

(c) There are positive K, b, δ such that $V(x) \leqslant K|x|^{-1-\delta}$ for almost all $|x| > b$.

In such a case, the hypotheses of Theorem 14.2.1 in Ref. 11 are easily seen to be fulfilled, so the following assertion holds.

Proposition 2.1: Under the assumptions (a)–(c), the wave operators $W_\pm\ (H_F, H_0)$ exist and are complete. ■

Next one has to make an assumption concerning the singularity of V. Semiclassical impenetrability of the barrier demands

$$\int_{\mathbb{R}} [V(x) - E]_+^p \, dx = \infty, \tag{2.3}$$

with $p = \frac{1}{2}$ for a given energy E of the particle. For a greater generality, we shall consider this condition with other values of p, too. Alternatively, one may require

$$\int_{-c}^{c} V(x)^p \, dx = \infty, \tag{2.4}$$

for some $c > 0$, as the following assertion shows.

Proposition 2.2: Assume (a)–(c) and fix $p \in (0,1]$. Then the following conditions are equivalent: (i) (2.3) holds for some $E > 0$; (ii) (2.3) holds for all $E > 0$; (iii) (2.4) holds for some $c > 0$; and (iv) (2.4) holds for all $c > 0$. ■

Thus the assumption may be formulated as follows.

(d_p) For a given $p \in (0,1]$, any of the conditions (i)–(iv) hold.

Besides (d_p), we are going to consider one more condition of this type.

(e) The integral $\int_{-c}^{c} x^2 V(x)^2 \, dx$ is divergent for some (hence also for any) positive c.

Proposition 2.3: (d_p) implies (d_q) if $0 < p \leqslant q \leqslant 1$. If $p < \frac{2}{3}$, then (d_p) implies (e). On the other hand, the conditions (d_p) and (e) are independent for $\frac{2}{3} \leqslant p \leqslant 1$.

Proof: To a positive c, we denote $J_c^- = [-c,c] \cap V^{(-1)}([0,1))$.

Then

$$\int_{-c}^{c} V(x)^q \, dx \geqslant \int_{-c}^{c} V(x)^p \, dx + \int_{J_c^-} [V(x)^q - V(x)^p] dx \geqslant \int_{-c}^{c} V(x)^p \, dx - 2c,$$

so the left-hand side is infinite if (d_p) holds. Next we use the Hölder inequality

$$\int_{-c}^{c} V(x)^p \, dx \leqslant \left(\int_{-c}^{c} |x|^{-2p/(2-p)} \, dx \right)^{(2-p)/2} \times \left(\int_{-c}^{c} x^2 V(x)^2 \, dx \right)^{p/2}.$$

If $p < \frac{2}{3}$, the first integral on the right-hand side (rhs) is finite so the second assertion follows.

Finally, one can consider the following examples. For a powerlike potential, $V(x) = |x|^{-\alpha}$ with $p^{-1} \leqslant \alpha < \frac{3}{2}$, the condition ($d_p$) is fulfilled, while (e) is not. If $V(x) = -|x|^{-3/2} \times \ln^{-1}|x|$ for $0 < |x| < \frac{1}{2}$ and $V(x) = 0$ for $|x| \geqslant \frac{1}{2}$, the same is true with $p = \frac{2}{3}$. On the other hand, the potential

$$V(x) = \begin{cases} n^3, & x \in (1/n, 1/n + 1/n^5), \quad n = 1,2,3,\ldots, \\ 0, & \text{otherwise}, \end{cases}$$

gives

$$\int V(x)^p \, dx = \sum_n n^{3p-5}$$

and

$$\int x^2 V(x)^2 \, dx = \sum_n (n^{-1} + O(n^{-5})),$$

so (e) is valid, while (d_p) is fulfilled for no $p \in (0,1]$. ■

In the next section, we are going to discuss the implications of these conditions for tunneling through the barrier.

III. THE MAIN RESULTS

Our goal is to show that under the stated singularity conditions, a state localized initially on one of the half-lines $\mathbb{R}_\pm$ stays confined there if its evolution is governed by the operator H_F associated with (2.2). Specifically, we are going to prove the following.

Theorem 3.1: Assume (a)–(c) and (d_1); then H_F commutes with the projections $E_\pm$ on $L^2(\mathbb{R}_\pm)$.

In view of Proposition 2.3, it gives the following result.

Corollary 3.2: Let H_F be the Hamiltonian of the problem. If the conditions (a)–(c) and (d_p) for some $p\in(0,1]$ hold, then $\exp(-iH_F t)$ is reduced by $E_\pm$ for all $t\in\mathbb{R}$, so there is no tunneling. In particular, it is true for $p=\frac{1}{2}$, in which case the tunneling is semiclassically forbidden.

Proof of Theorem 3.1: We denote $Q_1=\{\psi\in AC[\mathbb{R}]:\psi(0)=0\}$. For a vector $\psi\in Q(h)$, the function $V^{1/2}\psi$ must be square integrable; in view of (b) it is sufficient to investigate its behavior in some neighborhood of the origin only. Suppose $\psi(0)\neq 0$. Since ψ is (absolutely) continuous, there is a positive c such that $|\psi(x)|\geqslant\frac{1}{2}|\psi(0)|$ for $|x|<c$. Then

$$\int_{-c}^{c} V(x)|\psi(x)|^2\,dx \geqslant \frac{1}{4}|\psi(0)|^2\int_{-c}^{c} V(x)\,dx,$$

so we have a contradiction with (d_1). Consequently, $Q(h)\subset Q_1$.

Next we shall use this fact to prove the required commutativity. We define the restricted forms $h_\pm : h_\pm(\psi)=\|\psi'\|_\pm^2+\|V^{1/2}\psi\|_\pm^2$, where $\|\psi\|_\pm := \int_{\mathbb{R}_\pm}|\psi(x)|^2\,dx$, with the domains

$$Q(h_\pm)=\{\psi\in AC[\mathbb{R}_\pm]:V^{1/2}\psi\in L^2(\mathbb{R}_\pm),$$
$$\lim_{x\to 0^\pm}\psi(x)=0\};$$

they are obviously non-negative. Let us check that h_+ is closed. We take an arbitrary sequence $\{\psi_n\}\subset Q(h_+)$ such that $h_+(\psi_n-\psi_m)+\|\psi_n-\psi_m\|_+^2\to 0$ as $n,m\to\infty$. By $\tilde\psi_n$, we denote the extension of ψ_n to $\mathbb{R}$ such that $\tilde\psi_n(x)=0$ for $x<0$. It is easy to see that $\tilde\psi_n$ is continuous and belongs to $Q(h)$ for $\psi_n\in Q(h_+)$, and $h(\tilde\psi_n)=h_+(\psi_n)$. Furthermore, $\|\tilde\psi_n\|=\|\psi_n\|_+$, so $h(\tilde\psi_n-\tilde\psi_m)+\|\tilde\psi_n-\tilde\psi_m\|^2\to 0$ as $n,m\to\infty$. Since the form h is closed, there is $\tilde\psi\in Q(h)$ which is the limit of the Cauchy sequence and $h(\tilde\psi-\tilde\psi_n)+\|\tilde\psi-\tilde\psi_n\|^2\to 0$ as $n\to\infty$. The function $\tilde\psi$ is continuous and $\tilde\psi(0)=0$ so $\psi:=\tilde\psi\restriction\mathbb{R}_+$ fulfills $\lim_{x\to 0^+}\psi(x)=0$. Obviously, $\psi\in Q(h_+)$ is the sought limit of the sequence $\{\psi_n\}$. Thus the form h_+ is closed, and by an analogous argument, h_- is closed.

Then there are unique self-adjoint operators $H_\pm$ on $L^2(\mathbb{R}_\pm)$ associated with the forms $h_\pm$. We construct the operator $\tilde H$ on $L^2(\mathbb{R})=L^2(\mathbb{R}_-)\oplus L^2(\mathbb{R}_+)$ as the orthogonal sum $\tilde H=H_-\oplus H_+$. This operator is self-adjoint and reduced by the projections $E_\pm$ by its definition. Furthermore, we define the quadratic form $\tilde h$ as $\tilde h=h_-\oplus h_+$, i.e., $\tilde h(\varphi_-\oplus\varphi_+)=h_-(\varphi_-)+h_+(\varphi_+)$ with $Q(\tilde h)=\{\varphi=\varphi_-\oplus\varphi_+:\varphi_\pm\in Q(h_\pm)\}$. According to the definition, $\tilde h$ is non-negative and closed, and $\tilde H$ is the self-adjoint operator associated with it.

Finally, we shall compare the forms h and $\tilde h$. First let $\varphi\in Q(h)$; we can write $\varphi=\tilde\varphi_-+\tilde\varphi_+$, where $\tilde\varphi_\pm=E_\pm\varphi$, or $\varphi=\varphi_-\oplus\varphi_+$ with $\varphi_\pm=\tilde\varphi_\pm\restriction\mathbb{R}_\pm$. In the same way as above, one may use $Q(h)\subset Q_1$ to check that $\varphi_\pm\in Q(h_\pm)$, i.e., $\varphi\in Q(\tilde h)$. On the other hand, consider a vector $\varphi=\varphi_-\oplus\varphi_+\in Q(\tilde h)$. Extending the functions $\varphi_\pm$, we can write $\varphi=\tilde\varphi_-+\tilde\varphi_+$. The two functions are absolutely continuous and $\lim_{x\to 0^\pm}\tilde\varphi_\pm(x)=0$ so φ is absolutely continuous in (any finite subinterval of) $\mathbb{R}$. Further, $\varphi'_\pm\in L^2(\mathbb{R}_\pm)$ implies $\varphi'\in L^2(\mathbb{R})$ and $V^{1/2}\varphi_\pm\in L^2(\mathbb{R}_\pm)$ implies $V^{1/2}\varphi\in L^2(\mathbb{R})$, so $\varphi\in Q(h)$ holds, too. Thus we have $Q(\tilde h)=Q(h)$ and the forms coincide as

$$\begin{aligned}h(\varphi)&=\|\varphi'\|_-^2+\|\varphi'\|_+^2+\|V^{1/2}\varphi\|_-^2+\|V^{1/2}\varphi\|_+^2\\&=h_-(\varphi_-)+h_+(\varphi_+)=\tilde h(\varphi),\end{aligned}$$

for $\varphi\in Q(h)$. However, the self-adjoint operator associated with h is unique, and therefore $H_F=\tilde H$. ■

Using an operator argument, the following weaker assertion can be proved.

Theorem 3.3: Under the assumptions (a)–(c) and (e), the operator $\overline{H_1}$ commutes with $E_\pm$. Then there is no tunneling if H_1 is essentially self-adjoint.

Proof: We denote $D_1=\{\psi\in AC^2[\mathbb{R}]:\psi(0)=\psi'(0)=0\}$. In order to determine $D(H_1)$, we must specify the behavior of $V\psi$ around $x=0$. For $\psi\in AC^2[\mathbb{R}]$, one may use the first-order Taylor expansion with integral remainder

$$\psi(x)=\psi(0)+\psi'(0)x+r_2(x),$$

where

$$r_2(x)=x^2\int_0^1(1-t)\psi''(tx)dt. \tag{3.1a}$$

Estimating the integral by the Schwarz inequality, we get

$$|r_2(x)|\leqslant 3^{-1/2}\left|\int_0^x|\psi''(y)|^2\,dy\right|^{1/2}|x|^{3/2}. \tag{3.1b}$$

In analogy with the preceding proof, (e) implies $\psi(0)=0$ for $\psi\in D(H_1)$. Suppose $\psi'(0)\neq 0$. In that case, (3.1b) gives $|\psi(x)|\geqslant\frac{1}{2}|\psi'(0)x|$ for all sufficiently small x, and therefore

$$\int_{-c}^{c}V(x)^2|\psi(x)|^2\,dx\geqslant\frac{1}{4}|\psi'(0)|^2\int_{-c}^{c}x^2V(x)^2\,dx,$$

for some $c>0$. Consequently, $\psi'(0)=0$ holds, too, and $D(H_1)\subset D_1$.

Now the commutativity of $E_\pm$ with H_1 verifies easily. Since a vector $\psi\in D(H_1)$ can be represented by a continuously differentiable function that fulfills $\psi(0)=\psi'(0)=0$, we have $E_\pm\psi\in D(H_0)$. At the same time, $V\psi\in L^2(\mathbb{R})$ implies $VE_\pm\psi\in L^2(\mathbb{R})$, so $E_\pm\psi\in D(H_1)$. Check of the equality $H_1E_\pm\psi=E_\pm H_1\psi$ for such ψ is straightforward; thus $E_\pm H_1\subset H_1E_\pm$. Finally, it is not difficult to see that if H_1 commutes with a bounded operator, the same is true for its closure $\overline{H_1}$. ■

Let us collect now some simple properties of the operator H_1 and its extensions. We denote $H_{\min}=H_1\restriction C_0^\infty(\mathbb{R}\setminus\{0\})$. An argument analogous to that presented in Ref. 3, Appendix to Sec. X.1, shows that

$$D(H^*_{\min})\subset D^*, \tag{3.2a}$$

where $D^*=\{\psi\in L^2(\mathbb{R}):\psi,\psi'$ absolutely continuous in $\mathbb{R}\setminus\{0\}$, $\psi''\in L^2_{\mathrm{loc}}(\mathbb{R}\setminus\{0\})$, $-\psi''+V\psi\in L^2(\mathbb{R})\}$ and

$$H^*_{\min}\psi=-\psi''+V\psi. \tag{3.2b}$$

We have the following chain of inclusions:

$$H^*_{\min}\supset H_1^*\supset H_F\supset\overline{H}_1\supset H_1\supset H_{\min}; \tag{3.3}$$

if H_1 is not e.s.a., then other self-adjoint extensions may stand in the place of H_F.

These relations are important because they allow us to find the deficiency subspaces of the operator H_1 by solving the ordinary differential equation

$$-\psi''(x)+V(x)\psi(x)=\pm i\psi(x); \tag{3.4}$$

in particular, they allow us to verify the essential self-adjointness of H_1 in the case that the last equation has no solution within D^*. In Sec. IV, we shall treat in this way the potential barrier (1.5).

IV. AN EXAMPLE: $V(x)=gx^{-2}$

Here we are going to illustrate that we may not generally replace the Friedrichs extension H_F of H_1 in Corollary 3.2 by another one. To this end, we shall treat in detail the particular barrier

$$V(x)=gx^{-2},\quad g>0. \tag{4.1}$$

This potential fulfills obviously the assumptions (a)–(c), as well as (e) and (d_p) for $p>\frac{1}{2}$; thus all conclusions of the preceding section apply.

Let us ask when $H_1=H_0+V$ is e.s.a. In view of relations (3.2)–(3.4), the equation specifying the deficiency subspaces

$$(H_1^*-i\lambda)\varphi=0,$$

for $\lambda=\pm 1$, is simply related to the Bessel equation if we set

$$\nu=(g+\tfrac{1}{4})^{1/2}; \tag{4.2}$$

its solutions in $\mathbb{R}_\pm$ are linear combinations of $(\pm x)^{1/2}H_\nu^{(k)}(e^{\pm i\pi\nu/4}x), k=1,2$. It is easy to see that none of them is square integrable if $\nu>1$, i.e., $g>\frac{3}{4}$, and therefore H_1 is e.s.a. in this case. On the other hand, if $g\in(0,\frac{3}{4}]$, or $\nu\in(\frac{1}{2},1)$, the deficiency subspaces $K_\pm$ are two dimensional and spanned by the vectors

$$\varphi_+^{(1)}:\quad \varphi_+^{(1)}(x)=\theta(x)x^{1/2}H_\nu^{(1)}(\epsilon x), \tag{4.3a}$$

$$\varphi_+^{(2)}:\quad \varphi_+^{(2)}(x)=\varphi_+^{(1)}(-x)$$
$$=-\theta(-x)(-x)^{1/2}\bar{\epsilon}^{4\nu}H_\nu^{(2)}(\epsilon x), \tag{4.3b}$$

and

$$\varphi_-^{(1)}:\quad \varphi_-^{(1)}(x)=\overline{\varphi_+^{(1)}(x)}$$
$$=\theta(x)x^{1/2}H_\nu^{(2)}(\bar{\epsilon}x), \tag{4.3c}$$

$$\varphi_-^{(2)}:\quad \varphi_-^{(2)}(x)=\overline{\varphi_+^{(2)}(x)}$$
$$=-\theta(-x)(-x)^{1/2}\epsilon^{4\nu}H_\nu^{(1)}(\bar{\epsilon}x). \tag{4.3d}$$

Here and further on, we abbreviate $\epsilon=e^{\pi i/4}$. The self-adjoint extensions of H_1 are then constructed in the standard way.[3,4] They are parametrized by the isometries $K_+\to K_-$, i.e., by 2×2 matrices U whose elements fulfill the unitarity condition

$$\overline{u_{1j}}u_{1k}+\overline{u_{2j}}u_{2k}=\delta_{jk},\quad j,k=1,2. \tag{4.4}$$

For a given U, we denote

$$\varphi_U^{(k)}=\varphi_+^{(k)}-u_{1k}\varphi_-^{(1)}-u_{2k}\varphi_-^{(2)},\quad k=1,2. \tag{4.5}$$

According to the second von Neumann formula, the domain of the extension H_U of H_1 consists of the vectors $\psi=\varphi+(I-U)\varphi_+$, where $\varphi_+=c_1\varphi_+^{(1)}+c_2\varphi_+^{(2)}$, i.e.,

$$\psi=\varphi+c_1\varphi_U^{(1)}+c_2\varphi_U^{(2)}, \tag{4.6a}$$

with $\varphi\in D(\overline{H}_1)$ and $c_1,c_2\in\mathbb{C}$. The operator H_U acts on them as $H_U\psi=\overline{H}_1\varphi+i(I+U)\varphi_+$; in view of (3.2b) and the inclusions $H_U\subset H_1^*\subset H_{\min}^*$, one has

$$H_U\psi=-\psi''+V\psi. \tag{4.6b}$$

Let us now look more closely at how the functions of $D(H_U)$ behave around $x=0$. We take $\psi\in D(H_U)$ and $\varphi\in D(H_U)$ with supp $\varphi\subset[-n,n]$; then

$$(\varphi,H_U\psi)=\lim_{\eta\to 0^+}\left(\int_{-n}^{-\eta}+\int_\eta^n\right)\overline{\varphi}(x)$$
$$\times(-\psi''(x)+gx^{-2}\psi(x))dx.$$

Since both ψ',φ' are absolutely continuous in any compact subinterval of $\mathbb{R}\setminus\{0\}$, one can integrate by parts, obtaining in this way

$$(\varphi,H_U\psi)=(H_U\varphi,\psi)+\lim_{\eta\to 0^+}\sum_{\alpha=\pm}\alpha(\overline{\varphi}(\alpha\eta)$$
$$\times\psi'(\alpha\eta)-\overline{\varphi}'(\alpha\eta)\psi(\alpha\eta)).$$

Furthermore, to any $\varphi\in D(H_U)$ one can always find a sequence $\{\varphi_n\}\subset D(H_U)$ of functions supported by $[-n,n]$ such that $\varphi_n\to\varphi, H_U\varphi_n\to H_U\varphi$, e.g., by imposing (sufficiently smooth) cutoffs on φ. Then the last equality holds for arbitrary $\varphi,\psi\in D(H_U)$, and therefore the second term on its rhs must be zero for each such pair of vectors. This requirement can be reformulated as the following continuity condition:

$$\lim_{x\to 0^+}j(\varphi,\psi;x)=\lim_{x\to 0^-}j(\varphi,\psi;x), \tag{4.7a}$$

for $\varphi,\psi\in D(H_U)$, where $j(\varphi,\psi;x)=\overline{\varphi}(x)\psi'(x)-\overline{\varphi}'(x)\psi(x)$. By a polarization-identity-type argument, it is further equivalent to

$$\lim_{x\to 0^+}j_\psi(x)=\lim_{x\to 0^-}j_\psi(x), \tag{4.7b}$$

for each $\psi\in D(H_U)$, where $j_\psi(x)=(2i)^{-1}j(\psi,\psi;x)=\text{Im}\,\overline{\psi}(x)\psi'(x)$. Hence the domain of every particular extension H_U is contained in the set of vectors for which the probability current is continuous at $x=0$.

A stronger assertion is valid for the vectors of $D(\overline{H_1})$. One can take φ from the domain of $\overline{H_1}=H_1^{**}$ and $\psi\in D(H_1^*)$, and repeat the above argument for $(\varphi,H_1^*\psi)$; it yields again the condition (4.7a). In the present case, however, one may always write $\psi=\psi_1+\psi_2$, where both ψ_1, ψ_2 belong to $D(H_1^*)$ and are supported by $(-\infty,0)$ and $(0,\infty)$, respectively. Consequently, one has

$$\lim_{x\to 0^\pm}j(\varphi,\psi;x)=0, \tag{4.8}$$

for $\varphi\in D(\overline{H_1})$ and $\psi\in D(H_1^*)$. Notice that the last relation is easily verified directly if $\varphi\in D(H_1)$ and $\psi\in K_\pm$, because then $|\varphi(x)|\leqslant K_\varphi|x|^{3/2}$ (cf. the proof of Theorem 3.3) and the functions (4.3) behave near $x=0$ as follows:

$$\varphi_+^{(1)}(x)=A\,[\bar{\epsilon}^\nu x^{1/2-\nu}-B\bar{\epsilon}^{3\nu}x^{1/2+\nu}$$
$$+O(x^{5/2-\nu})],\quad x>0, \tag{4.9a}$$

etc. (the remaining formulas are obtained by complex conjugation and/or replacement of $x\to -x$), where

$$A=-(i/\pi)2^\nu\Gamma(\nu)=-2^\nu i/\Gamma(1-\nu)\sin\nu\pi,$$

$$B=4^{-\nu}\Gamma(1-\nu)/\Gamma(1+\nu). \tag{4.9b}$$

On the other hand, (4.8) need not be true if neither φ nor ψ is contained in $D(\overline{H_1})$. In view of (4.6a), we are particularly interested in the case when φ,ψ are of the form (4.5). The limits can be calculated with the help of (4.9) and (4.4); they equal

$$\lim_{x\to 0^\pm} j(\varphi_U^{(1)},\varphi_U^{(1)};x) = (2i/\pi)|u_{21}|^2 \sec(\nu\pi/2), \tag{4.10a}$$

$$\lim_{x\to 0^\pm} j(\varphi_U^{(2)},\varphi_U^{(2)};x) = -(2i/\pi)|u_{12}|^2 \sec(\nu\pi/2), \tag{4.10b}$$

$$\lim_{x\to 0^\pm} j(\varphi_U^{(1)},\varphi_U^{(2)};x) = -(2i/\pi)\,\overline{u_{11}}u_{12} \sec(\nu\pi/2). \tag{4.10c}$$

We see that, in general, the probability current for $\psi\in D(H_U)$ need not vanish at $x=0$ unless U is diagonal. It indicates that the tunneling might occur in such cases. In the next section, we confirm this conjecture by evaluating the transmission coefficient. Notice that the matrix $U=U_F$ referring to the Friedrichs extension is diagonal: we have shown in the proof of Theorem 3.1 that $D(H_F)\subset Q(h)\subset Q_1$, and therefore one has to require $\lim_{x\to 0^\pm} \varphi_U^{(k)}(x)=0$. It yields easily

$$U_F = -\bar{\epsilon}^{2\nu}I. \tag{4.11}$$

Remark 4.1: In the above considerations, closedness of H_1 is not required. Before proceeding further, we would like to mention an elegant proof of this property which, however, works for $g>\frac{3}{4}$ only. It is based on the canonical commutation relations. We write $H_1=(P^2+\frac{3}{4}Q^{-2})+(g-\frac{3}{4})Q^{-2}$ and apply it on a vector ψ of a suitable domain, say, $C_0^\infty(\mathbb{R}\setminus\{0\})$. It gives

$$\|H_1\psi\|^2 = \|(P^2+\tfrac{3}{4}Q^{-2})\psi\|^2 + (g^2-\tfrac{9}{16})\|Q^{-2}\psi\|^2 + (g-\tfrac{3}{4})(\psi,(P^2Q^{-2}+Q^{-2}P^2)\psi).$$

Using the relation $[P,Q^{-1}]\psi = iQ^{-2}\psi$, one can rewrite the last term as follows:

$$(\psi,(P^2Q^{-2}+Q^{-2}P^2)\psi) = \tfrac{1}{2}\|(3PQ^{-1}-Q^{-1}P)\psi\|^2 - \tfrac{3}{2}\|Q^{-2}\psi\|^2.$$

Omitting the non-negative terms, we get the inequality

$$\|H_1\psi\|^2 \geqslant (g-\tfrac{3}{4})^2\|Q^{-2}\psi\|^2,\quad \psi\in C_0^\infty(\mathbb{R}\setminus\{0\}). \tag{4.12}$$

The remaining part of the argument is simple (cf. the analogous problem considered in Ref. 13, Proposition 1). Since $g>\frac{3}{4}$, $H_{\min}$ is e.s.a. and relations (3.3) show that $C_0^\infty(\mathbb{R}\setminus\{0\})$ is a core for H_1. To a vector $\psi\in D(\overline{H_1})$, we take a sequence $\{\psi_n\}\subset C_0^\infty(\mathbb{R}\setminus\{0\}),\psi_n\to\psi$: then $\{H_1\psi_n\}$ is Cauchy and the same is true for $\{Q^{-2}\psi_n\}$ due to (4.12), and for $\{P^2\psi_n\}$, too, since $P^2\psi_n = H_1\psi_n - gQ^{-2}\psi_n$. However, both P^2 and Q^{-2} are closed so $\psi\in D(P^2)\cap D(Q^{-2}) = D(H_1)$.

V. THE TRANSMISSION COEFFICIENT (EXAMPLE CONTINUES)

Now we shall discuss scattering on the barrier (4.1), restricting ourselves to the nontrivial case $0<g<\frac{3}{4}$ only. We use the time-independent setting, because it is simpler, and at the same time, it allows us to illustrate the main point, namely that the dynamics is determined by the self-adjoint extension H_U chosen to play the role of Hamiltonian. Hence we are going to work with the functions which obey all appropriate requirements *locally* but eventually do not exhibit the overall square integrability. In order to get the rigorous Hilbert-space (time-dependent) scattering theory, one should consider the motion of wave packets composed of the plane-wave solutions constructed below; but we are not going to pursue this task.

Given $E>0$, we are looking for solutions of the stationary Schrödinger equation

$$-\psi''(x) + gx^{-2}\psi(x) = E\psi(x), \tag{5.1}$$

assuming they are of the form analogous to (4.6a)

$$\psi = \psi_1 + c_1\varphi_U^{(1)} + c_2\varphi_U^{(2)}, \tag{5.2}$$

where ψ_1 belongs locally to $D(\overline{H_1})$. Equation (5.1) can then be rewritten as

$$-\psi_1''(x) + (gx^{-2}-E)\psi_1(x) = \chi(x), \tag{5.3a}$$

where χ is expressed through the functions (4.3) as

$$\chi(x) = \theta(x)[c_+^{(1)}\varphi_+^{(1)}(x) + c_-^{(1)}\varphi_-^{(1)}(x)] + \theta(-x)[c_+^{(2)}\varphi_+^{(2)}(x) + c_-^{(2)}\varphi_-^{(2)}(x)], \tag{5.3b}$$

with

$$c_+^{(k)} = c_k(E-i),\qquad c_-^{(k)} = -(c_1u_{k1}+c_2u_{k2})(E+i),\quad k=1,2. \tag{5.3c}$$

The function χ is C^∞ in $\mathbb{R}\setminus\{0\}$ so the same is true for ψ_1. It can be seen as follows: $\psi_1'' = (gx^{-2}-E)\psi_1 - \chi\in C^1(\mathbb{R}\setminus\{0\})$ because ψ_1 belongs locally to $D(\overline{H_1})\subset D^*$ [cf. (3.2a)]. Then one has to differentiate, successively, the last relation.

First we shall solve Eq. (5.3a) for $x>0$. We start with the related homogeneous equation whose solution is easily found to be $\psi_0 = \alpha_1\psi_{01}+\alpha_2\psi_{02}$, where

$$\psi_{0k}(x) = x^{1/2}H_\nu^{(k)}(\lambda x),\quad k=1,2, \tag{5.4a}$$

where $\lambda = E^{1/2}$. The Wronskian of these functions can be determined from their asymptotic behavior, either for $x\to\infty$ or for $x\to 0^+$

$$\psi_{01}(x) = \overline{\psi_{02}(x)} = A\,[E^{-\nu/2}x^{1/2-\nu} - B\bar{\epsilon}^{4\nu}E^{\nu/2}x^{1/2+\nu} + O(x^{5/2-\nu})]; \tag{5.5}$$

one has $W(\psi_{01},\psi_{02}) = 4/\pi i$. Next we suppose the general solution to (5.3a) to be of the form

$$\psi_1(x) = \alpha_1(x)\psi_{01}(x) + \alpha_2(x)\psi_{02}(x). \tag{5.4b}$$

The corresponding system of first-order equations for the functions α_k is easily solved, giving

$$\psi_1(x) = \frac{\pi i}{4}\psi_{01}(x)\int_0^x \chi(y)\psi_{02}(y)dy - \frac{\pi i}{4}\psi_{02}(x)\int_0^x \chi(y)\psi_{01}(y)dy + \alpha_1\psi_{01}(x) + \alpha_2\psi_{02}(x), \tag{5.6}$$

where the constants α_1, α_2 are arbitrary up to now. In order to fix them, let us look at the behavior of ψ_1 near the origin. It can be found from the formulas (4.9), (5.3), and (5.5). A short calculation shows that the leading-order terms (behaving as $x^{5/2-3\nu}$) in the first two expressions of (5.6) cancel mutually, and that the asymptotics are determined by the last two expressions, specifically

$$\psi_1(x) = (\alpha_1 - \alpha_2)AE^{-\nu/2}|x|^{1/2-\nu} - AB(\alpha_1\bar{\epsilon}^{4\nu} - \alpha_2\epsilon^{4\nu})E^{\nu/2}|x|^{1/2+\nu} + O(|x|^{5/2-\nu}). \quad (5.7)$$

Before proceeding further, let us look for the solution to (5.3a) for $x<0$. It can be found easily: $\psi_1(x) = \tilde{\psi}_1(-x)$, where $\tilde{\psi}_1$ fulfills the same equation as ψ_1 with replacement of $\chi(x)$ by $\chi(-x)$. Since $\varphi_\pm^{(2)}(x) = \varphi_\pm^{(1)}(-x)$, the function $\tilde{\psi}_1$ differs from ψ_1 just by the coefficients: in the first two terms of (5.6), $c_\pm^{(2)}$ stand in the place of $c_\pm^{(1)}$, and α_1, α_2 may assume other values. With this difference, the asymptotics of ψ_1 for $x \to 0^-$ are again given by (5.7).

Lemma 5.1: Let a function $\psi \in L^{(2)}_{\rm loc}(\mathbb{R})$ fulfill the following conditions: (a) ψ, ψ' are absolutely continuous in $\mathbb{R}\setminus\{0\}$; (b) there is a positive η such that $\psi'' \in L^2_{\rm loc}(\mathbb{R}\setminus[-\eta,\eta])$; and (c) it holds that $\psi(x) = [a_1x^{1/2-\nu} + a_2x^{1/2+\nu}]\theta(x) + [a_3(-x)^{1/2-\nu} + a_4(-x)^{1/2+\nu}]\theta(-x) + \xi(x)$, where ξ belongs locally to $D(H_1)$ [equivalently, ξ' is absolutely continuous in $\mathbb{R}$ and the functions ξ'', $x^{-2}\xi$ belong to $L^2(-\eta,\eta)$].

Then either $a_1 = \cdots = a_4 = 0$ or ψ does not belong to $D(\overline{H_1})$.

Proof: We write the function $\sigma \equiv \psi - \xi$ as $\varphi - \omega$, where φ is a suitable linear combination of the functions (4.3) and ω is regular at the origin. Using (4.9) and the fact that $\sin(\nu\pi/2)$ is nonzero for $\nu \in (\frac{1}{2}, 1)$, we see that φ is determined uniquely by the numbers a_j and that $\omega(x) = O(|x|^{5/2-\nu})$ near the origin. Then ω belongs locally to $D(H_1) \subset D(\overline{H_1})$ and the same is true for $\xi - \omega = \psi - \varphi$ in view of the assumption (c). Suppose that ψ belongs locally to $D(\overline{H_1})$; then the same should hold for φ. Moreover, φ is square integrable so $\varphi \in D(\overline{H_1})$. However, φ lies in the subspace $K_+ \oplus_{H_1} K_-$ of $D(H_1^*)$ that is H_1 orthogonal to $D(\overline{H_1})$ (Ref. 3, Sec. X.1). It is possible only if $\varphi = 0$, or equivalently, $a_1 = \cdots = a_4 = 0$. ■

The function ψ_1 is supposed to be locally of $D(\overline{H_1})$ so the above lemma implies easily that the last two terms on the rhs of (5.6) must vanish, and similarly for $\tilde{\psi}_1$. Now we express the obtained solution more explicitly, substituting for χ from (5.3b). We denote

$$J_{k\pm}(x) = \int_x^\infty \varphi_\pm^{(1)}(y)\psi_{0k}(y)dy, \quad x \geqslant 0, \quad (5.8a)$$

$$J_{k\pm} = J_{k\pm}(0). \quad (5.9)$$

The asymptotic behavior of $J_{k\pm}(\cdot)$ for $x \to \infty$ can be found using that of the cylindrical functions in (4.3) and (5.4a): there is a constant K (depending on E) such that

$$|J_{k\pm}(x)| \leqslant K\exp(-2^{-1/2}x), \quad x \geqslant 0. \quad (5.8b)$$

The function ψ_1 for $x>0$ can be now rewritten as

$$\psi_1(x) = (\pi i/4)[c_+^{(1)}(J_{2+} - J_{2+}(x)) + c_-^{(1)}(J_{2-} - J_{2-}(x))]\psi_{01}(x) - (\pi i/4)[c_+^{(1)}(J_{1+} - J_{1+}(x)) + c_-^{(1)}(J_{1-} - J_{1-}(x))]\psi_{02}(x), \quad (5.10a)$$

and due to (5.4a) and (5.8b), its asymptotic behavior for $x \to \infty$ is the following:

$$\psi_1(x) = \frac{i}{2}\left(\frac{\pi}{2\lambda}\right)^{1/2}\Big\{(c_+^{(1)}J_{2+} + c_-^{(1)}J_{2-}) \times\exp\Big[i\Big(\lambda x - \frac{\nu\pi}{2} - \frac{\pi}{4}\Big)\Big] - (c_+^{(1)}J_{1+} + c_-^{(1)}J_{1-})\exp\Big[-i\Big(\lambda x - \frac{\nu\pi}{2} - \frac{\pi}{4}\Big)\Big]\Big\} + O(x^{-1}), \quad (5.11a)$$

where again $\lambda = E^{1/2}$. Similarly, one has the expression

$$\psi_1(x) = (\pi i/4)[c_+^{(2)}(J_{2+} - J_{2+}(-x)) + c_-^{(2)}(J_{2-} - J_{2-}(-x))]\psi_{01}(-x) - (\pi i/4) \times [c_+^{(2)}(J_{1+} - J_{1+}(-x)) + c_-^{(2)}(J_{1-} - J_{1-}(-x))]\psi_{02}(-x), \quad (5.10b)$$

for $x<0$, which behaves for $x \to -\infty$ as

$$\psi_1(x) = \frac{i}{2}\left(\frac{\pi}{2\lambda}\right)^{1/2}\Big\{(c_+^{(2)}J_{2+} + c_-^{(2)}J_{2-}) \times\exp\Big[-i\Big(\lambda x + \frac{\nu\pi}{2} + \frac{\pi}{4}\Big)\Big] - (c_+^{(2)}J_{1+} + c_-^{(2)}J_{1-})\exp\Big[i\Big(\lambda x + \frac{\nu\pi}{2} + \frac{\pi}{4}\Big)\Big]\Big\} + O(|x|^{-1}). \quad (5.11b)$$

In order to make use of the relations (5.11), we should know the coefficients $J_{k\pm}$ explicitly. It can be easily achieved (cf. Ref. 14, 6.521.3 and 8.407.1; integrals $J_{k\pm}$ can be computed with the help of equations satisfied by functions $\varphi_\pm^{(1)}$ and ψ_{0k}). One has

$$J_{1+} = \frac{2}{\pi}\frac{\bar{\epsilon}^{4\nu}}{\sin\nu\pi}\frac{\bar{\epsilon}^{\nu}E^{\nu/2} - \epsilon^{\nu}E^{-\nu/2}}{E-i}, \quad (5.12a)$$

$$J_{1-} = \frac{2}{\pi}\frac{\epsilon^{3\nu}E^{-\nu/2} - \bar{\epsilon}^{3\nu}E^{\nu/2}}{\sin\nu\pi\,(E+i)}, \quad (5.12b)$$

$$J_{2+} = \overline{J_{1-}}, \quad J_{2-} = \overline{J_{1+}}. \quad (5.12c)$$

Now the crucial point is that the function $\varphi_U^{(k)}$ in (5.2) decays exponentially at infinity. Thus the asymptotic behavior of ψ coincides with that of ψ_1, and it is fully determined by the choice of the coefficients $c_\pm^{(k)}$. One has to know the correspondences $(c_+^{(1)}, c_-^{(1)}) \leftrightarrow (c_1, c_2) \leftrightarrow (c_+^{(2)}, c_-^{(2)})$. Two cases should be distinguished.

(i) *The matrix U is diagonal:* Then the pairs $c_\pm^{(1)}$ and $c_\pm^{(2)}$ are independent: the relations (5.3c) give

$$c_-^{(k)} = -u_{kk}\frac{E+i}{E-i}c_+^{(k)}, \quad k = 1,2. \quad (5.13)$$

Consider the situation when one of these pairs is zero, say, $c_\pm^{(2)} = 0$. Then the formulas (5.11a), (5.12c), and (5.13) yield the asymptotics

$$\psi(x) = \frac{i}{2}\left(\frac{\pi}{2\lambda}\right)^{1/2}c_+^{(1)}\Big\{\Big(J_{2+} - u_{11}\frac{E+i}{E-i}J_{2-}\Big) \times\exp\Big[i\Big(\lambda x - \frac{\nu\pi}{2} - \frac{\pi}{4}\Big)\Big] + u_{11}\frac{E+i}{E-i} \times \overline{\Big(J_{2+} - u_{11}\frac{E+i}{E-i}J_{2-}\Big)}\exp\Big[-i\Big(\lambda x$$

$$-\frac{\nu\pi}{2}-\frac{\pi}{4}\Big)\Big]\Big\}+O(x^{-1}), \tag{5.14a}$$

for $x\to\infty$. On the other hand, $c_{\pm}^{(2)}=0$ implies $c_2=0$ so the relations (5.2) and (4.5) together with $u_{21}=0$ give $\psi(x)=\psi_1(x)$ for $x<0$. In that case, however,

$$\psi(x)=0, \quad x<0 \tag{5.14b}$$

holds due to (5.10b). Hence we have total reflection in this case; the phase shift of the reflected wave can be easily derived from (5.14a).

(ii) *The matrix U is nondiagonal:* Now the appropriate determinants are nonzero so the correspondence $(c_+^{(1)},c_-^{(1)})\leftrightarrow(c_+^{(2)},c_-^{(2)})$ is bijective. We choose the initial conditions in such a way that we have the transmitted wave on the positive semiaxis only, i.e.,

$$c_+^{(1)}J_{1+}+c_-^{(1)}J_{1-}=0. \tag{5.15a}$$

Then the reflection and transmission coefficients are defined by the expressions

$$R=R_\nu(E;U)=\left|\frac{c_+^{(2)}J_{2+}+c_-^{(2)}J_{2-}}{c_+^{(2)}J_{1+}+c_-^{(2)}J_{1-}}\right|^2 \tag{5.15b}$$

and

$$T=T_\nu(E;U)=\left|\frac{c_+^{(1)}J_{2+}+c_-^{(1)}J_{2-}}{c_+^{(2)}J_{1+}+c_-^{(2)}J_{1-}}\right|^2, \tag{5.15c}$$

respectively. Using (5.12c) together with the unitarity condition (4.4), one can check that the equality

$$|c_+^{(2)}J_{2+}+c_-^{(2)}J_{2-}|^2+|c_+^{(1)}J_{2+}+c_-^{(1)}J_{2-}|^2$$
$$=|c_+^{(2)}J_{1+}+c_-^{(2)}J_{1-}|^2$$

holds if (5.15a) is valid, i.e.,

$$R+T=1. \tag{5.16}$$

Let us now express the transmission coefficient more explicitly. The system of equations for $c_{\pm}^{(2)}$ that follows from (5.3c) is easily solved. Further, $c_+^{(1)}$ and $c_-^{(1)}$ are related by (5.15a), so we obtain

$$c_+^{(2)}=-\frac{1}{u_{12}}\frac{E-i}{E+i}(1+u_{11}\beta)c_-^{(1)},$$
$$c_-^{(2)}=[-u_{21}+(u_{22}/u_{12})(1+u_{11}\beta)]c_-^{(1)},$$

where

$$\beta=-\frac{E+i}{E-i}\frac{J_{1-}}{J_{1+}}=\epsilon^{6\nu}\frac{1-\bar{\epsilon}^{6\nu}E^{\nu}}{1-\bar{\epsilon}^{2\nu}E^{\nu}}. \tag{5.17a}$$

Substituting now to (5.15c) and denoting

$$\gamma=J_{2+}/J_{1+}=-(1-\epsilon^{6\nu}E^{\nu})/(1-\bar{\epsilon}^{2\nu}E^{\nu}), \tag{5.17b}$$

we arrive after a short calculation at the expression

$$T=\left|\frac{\gamma(\beta-\bar{\beta}^{-1})u_{12}}{1+\beta\,\mathrm{Tr}\,U+\beta^2\det U}\right|^2. \tag{5.17c}$$

After some more simple manipulations, we can rewrite it in the following final form:

$$T_\nu(E;U)=16|u_{12}|^2\sin^2\nu\pi$$
$$\times\sin^2\frac{\nu\pi}{2}|aE^{\nu}+b+cE^{-\nu}|^{-2}, \tag{5.18a}$$

where

$$a=a_\nu(U)=\bar{\epsilon}^{4\nu}[1+\epsilon^{2\nu}\,\mathrm{Tr}\,U+\epsilon^{4\nu}\det U], \tag{5.18b}$$

$$b=b_\nu(U)=-2\bar{\epsilon}^{2\nu}\Big[1+\epsilon^{4\nu}\cos\frac{\nu\pi}{2}\,\mathrm{Tr}\,U+\epsilon^{8\nu}\det U\Big], \tag{5.18c}$$

$$c=c_\nu(U)=1+\epsilon^{6\nu}\,\mathrm{Tr}\,U+\epsilon^{12\nu}\det U. \tag{5.18d}$$

Let us collect some simple properties of T.

Theorem 5.2: Let $E>0$ and $\nu=(g+\frac{1}{4})^{1/2}\in(\frac{1}{2},1)$. The transmission coefficient $T_\nu(E;U)$ referring to a self-adjoint extension H_U of H_1 with the potential (4.1) is then given by the formulas (5.18), where the coefficients (5.18b)–(5.18d) cannot be simultaneously zero. It assumes values from [0,1] and depends continuously on ν, E, U. In particular, for a diagonal matrix U we have total reflection, $T_\nu(E;U)=0$. For a nondiagonal U, the following alternative is valid: *either* U is unitarily equivalent to the matrix

$$-\bar{\epsilon}^{4\nu}\begin{pmatrix}\epsilon^{2\nu} & 0\\ 0 & \bar{\epsilon}^{2\nu}\end{pmatrix} \tag{5.19a}$$

and

$$T=4|u_{12}|^2\cos^2(\nu\pi/2) \tag{5.19b}$$

is energy independent, *or* at most one of the coefficients (5.18b)–(5.18d) can be zero. In that case, T has the following asymptotic behavior:

$$T_\nu(E;U)=16|u_{12}|^2\sin^2\nu\pi\sin^2(\nu\pi/2)f(E), \tag{5.20a}$$

where

$$f(E)=|c|^{-2}E^{2\nu}-2|c|^{-4}\,\mathrm{Re}\,\bar{b}cE^{3\nu}+O(E^{4\nu}),\quad c\neq0, \tag{5.20b}$$

$$f(E)=|b|^{-2}-2|b|^{-4}\,\mathrm{Re}\,\bar{a}bE^{\nu}+O(E^{2\nu}),\quad c=0 \tag{5.20c}$$

holds for $E\to0^+$, and similarly

$$f(E)=|a|^{-2}E^{-2\nu}-2|a|^{-4}\,\mathrm{Re}\,\bar{a}bE^{-3\nu}+O(E^{-4\nu}),\quad a\neq0, \tag{5.20d}$$

$$f(E)=|b|^{-2}-2|b|^{-4}\,\mathrm{Re}\,\bar{b}cE^{-\nu}+O(E^{-2\nu}),\quad a=0, \tag{5.20e}$$

for $E\to\infty$.

For fixed E and U, the transmission coefficient tends not necessarily to 0 as $g\to\frac{3}{4}^-$ and to 1 as $g\to0^+$. In fact, we have

$$\lim_{\nu\to1/2^+}T_\nu(E;U)=1, \tag{5.21a}$$

for all E iff the matrix U is of the form

$$2^{-1/2}\begin{pmatrix}i & e^{i\omega}\\ e^{-i\omega} & i\end{pmatrix}, \tag{5.21b}$$

with $\omega\in\mathbb{R}$. For

$$U=e^{i\xi}\begin{pmatrix}\cos\varphi & -e^{i\omega}\sin\varphi\\ e^{-i\omega}\sin\varphi & \cos\varphi\end{pmatrix}, \tag{5.21c}$$

with $\xi=(\alpha_1+\alpha_2)/2$, $\varphi=(\alpha_1-\alpha_2)/2$, $\alpha_1\in(\pi/4,3\pi/4)$, $\alpha_2\in(3\pi/4,9\pi/4)$, $\omega\in\mathbb{R}$, Eq. (5.21a) holds iff

$$E=-\frac{\sin(\alpha_1/2-\pi/8)\sin(\alpha_2/2-\pi/8)}{\cos(\alpha_1/2+\pi/8)\cos(\alpha_2/2+\pi/8)}. \tag{5.21d}$$

Here $\exp(i\alpha_1)$ and $\exp(i\alpha_2)$ are eigenvalues of the matrix

(5.21c). For other matrices U, Eq. (5.21a) does not hold with any $E\in(0,\infty)$.

On the other hand,

$$\lim_{\nu\to 1^-} T_\nu(E;U)=0 \qquad (5.22a)$$

holds almost everywhere; there are at most two values of E where the limit is nonzero for a nondiagonal U, namely

$$E=\tan(\alpha/2+\pi/4), \qquad (5.22b)$$

where $e^{i\alpha}$ is an eigenvalue of U such that $\alpha\in(-\pi/2,\pi/2)$.

Proof: The inequality $0\leqslant T\leqslant 1$ follows from (5.15) and (5.16). Each of the coefficients (5.18b)–(5.18d) may be zero (examples of such matrices U can be easily found), but they cannot vanish simultaneously. Suppose, e.g., that $a=b=0$. The relations (5.18b) and (5.18c) then yield

$$\det U = -\bar{\epsilon}^{4\nu}-\bar{\epsilon}^{2\nu}\,\mathrm{Tr}\,U$$
$$= -\bar{\epsilon}^{8\nu}-\bar{\epsilon}^{4\nu}\cos(\nu\pi/2)\mathrm{Tr}\,U \qquad (5.23a)$$

and the last equality further implies

$$\mathrm{Tr}\,U=-2\bar{\epsilon}^{2\nu}. \qquad (5.23b)$$

The relations (5.23) are fulfilled iff $U=-\bar{\epsilon}^{2\nu}I$; then $a=b=0$, while $c=(1-\epsilon^{4\nu})^2$ is nonzero. Similarly, $b=c=0$ is possible for $U=-\bar{\epsilon}^{6\nu}I$ only, in which case $a=\bar{\epsilon}^{4\nu}(1-\bar{\epsilon}^{4\nu})^2\neq 0$. Both these U are multiples of the unit matrix so the transmission coefficient is zero for them. For a nondiagonal matrix U, therefore, either $a=c=0$ or at least two of the coefficients (5.18b)–(5.18d) are nonzero. In the first case, relations (5.18b) and (5.18d) determine uniquely the values

$$\mathrm{Tr}\,U=-2\bar{\epsilon}^{4\nu}\cos(\nu\pi/2),\quad \det U=\bar{\epsilon}^{8\nu},$$

which can be achieved just for the matrices U that are unitarily equivalent to matrix (5.19a). The relation (5.19b) then follows readily from (5.18). Check of the relations (5.20) is elementary.

In order to prove the continuity of $T_\nu(E;U)$, one has to verify that the denominator in (5.18a) does not vanish in the allowed region of parameters. In view of (5.17), it equals $E^{-\nu}(1+\beta\,\mathrm{Tr}\,U+\beta^2\det U)(1-\bar{\epsilon}^{2\nu}E^{\nu})^2$ and since the last factor is nonzero, we must solve the corresponding quadratic equation for β. Its roots are easily seen to be $\beta_k=-\exp(-i\alpha_k)$, $k=1,2$, where $\exp(i\alpha_k)$ are eigenvalues of U. Hence we have to solve the equation

$$\beta=\epsilon^{6\nu}\frac{1-\bar{\epsilon}^{6\nu}E^{\nu}}{1-\bar{\epsilon}^{2\nu}E^{\nu}}=-e^{-i\alpha},\quad \alpha=\alpha_1,\alpha_2. \qquad (5.24)$$

Obviously $|\beta|=1$ in (5.24) and it yields the condition $\sin\nu\pi\times\sin(\nu\pi/2)=0$, which cannot be fulfilled for $\nu\in(\frac{1}{2},1)$.

Let us pass to the limits in ν. Equation (5.24) has no solution for $\nu=\frac{1}{2}$ as well, so we have

$$\lim_{\nu\to 1/2^+} T_\nu(E;U)=8|u_{12}|^2|(-i+\bar{\epsilon}\,\mathrm{Tr}\,U+\det U)E^{1/2}$$
$$-2\bar{\epsilon}(1+2^{-1/2}i\,\mathrm{Tr}\,U-\det U)$$
$$-i(i-\epsilon\,\mathrm{Tr}\,U+\det U)E^{-1/2}|^{-2}, \qquad (5.25)$$

with the rhs properly defined. It yields particularly (5.21a) if $a=c=0$ and $|u_{12}|=2^{-1/2}$, what is possible just for the matrices (5.21b). If these requirements are not fulfilled, then (5.21a) can hold only for the values of E where the denominator in the rhs of (5.25) reaches its minimum. Expressing Tr U and det U in terms of eigenvalues $\exp(i\alpha_1)$ and $\exp(i\alpha_2)$, we see that the denominator has a form

$$|a|^2E+|b|^2+2\,\mathrm{Re}\,\bar{a}c+|c|^2E^{-1},$$

since $\mathrm{Re}\,\bar{a}b=\mathrm{Re}\,\bar{b}c=0$. It has a minimum at energy $E\in(0,\infty)$ given by Eq. (5.21d) for $\alpha_1\in(\pi/4,3\pi/4)$, $\alpha_2\in(3\pi/4,9\pi/4)$; possible minima for other values of α_k do not lead to (5.21a). Further, $|u_{12}|$ must have its maximal value $|\sin[(\alpha_1-\alpha_2)/2]|$ at given α_1,α_2 to obtain the maximum in expression (5.25). Then the matrix U has the form (5.21c).

On the other hand, Eq. (5.24) has two solutions $E_k, k=1,2$, if $\nu=1$; they are possibly equal to each other (if $\alpha_1=\alpha_2$) or to ∞ (if $\alpha_k=\pi/2$). It is easy to find that they are given by (5.22b), where only the case $\alpha\in(-\pi/2,\pi/2)$ is interesting, giving a positive solution. With the exception of $E=E_k$, the relation (5.22a) holds. If E is equal to $E_k>0$, one obtains

$$\lim_{\nu\to 1^-} T_\nu(E_k;U)=16\pi^2|u_{12}|^2|d(E_k;U)|^{-2}, \qquad (5.26a)$$

where

$$d(E;U)=-\frac{d}{d\nu}[a_\nu(U)E^{\nu}+b_\nu(U)$$
$$+c_\nu(U)E^{-\nu}]_{\nu=1}; \qquad (5.26b)$$

the last expression cannot be zero at $E=E_k$ since T is bounded by 1. ■

Hence we have confirmed the conjecture formulated in the preceding section: the tunneling actually occurs for the barrier (4.1) if $g\in(0,\frac{3}{4})$, unless the matrix U specifying the Hamiltonian is diagonal. The transmission coefficient is "almost continuous" when the coupling constant g reaches the critical value $g=\frac{3}{4}$, above which H_1 is self-adjoint (cf. Remark 4.1) and the tunneling is forbidden due to Theorem 3.3. More explicitly, the relation (5.22a) holds with a possible exception of the "resonant" energies (5.22b) for which the tunneling vanishes discontinuously at $g=\frac{3}{4}$. On the other hand, the barrier does not generally become fully transparent as $g\to 0^+$, except for the special class of the matrices U specified above. This is not strange, however. Even for a very small g, the barrier still has the singularity at $x=0$, which makes the motion over it different from the free-particle case.

ACKNOWLEDGMENTS

The authors are indebted for discussions to Dr. V. I. Inozemtsev and to Dr. V. A. Rubakov, who suggested the quantum-mechanical tunneling as a model to the problem mentioned in the Introduction. The hospitality extended to the authors at the Laboratory of Theoretical Physics of the Joint Institute for Nuclear Research in Dubna, where the work was done, is also gratefully acknowledged.

[1]S. Coleman, "Classical lumps and their quantum descendants," in *New Phenomena in Subnuclear Physics*, edited by A. Zichichi (Plenum, New York, 1977).

[2]J. Dittrich, Czech. J. Phys. B **34**, 832 (1984).
[3]M. Reed and B. Simon, *Methods of Modern Mathematical Physics, II. Fourier Analysis. Self-adjointness, III. Scattering Theory* (Academic, New York, 1975).
[4]J. Blank and P. Exner, *Selected Topics of Mathematical Physics: Linear Operators in Hilbert Space II* (SPN, Prague, 1978) (in Czech).
[5]K. M. Case, Phys. Rev. **80**, 797 (1950).
[6]E. Nelson, J. Math. Phys. **5**, 332 (1964).
[7]K. Meetz, Nuovo Cimento **34**, 690 (1964).
[8]A. M. Perelomov and V. S. Popov, Teor. Mat. Fiz. **4**, 48 (1970).
[9]W. M. Frank, D. J. Land, and R. M. Spector, Rev. Mod. Phys. **43**, 36 (1971).
[10]Ch. Radin, J. Math. Phys. **16**, 544 (1975).
[11]M. Schechter, *Operator Methods in Quantum Mechanics* (North-Holland, New York, 1981).
[12]W. O. Amrein, J. M. Jauch, and K. B. Sinha, *Scattering Theory in Quantum Mechanics* (Benjamin, Reading, MA, 1977).
[13]P. Exner, J. Math. Phys. **24**, 1129 (1983).
[14]I. S. Gradshtein and I. M. Ryzhik, *Tables of Integrals, Sums, Series and Products* (Nauka, Moscow, 1971) (in Russian).

Global aspects of shear-free perfect fluids in general relativity

C. B. Collins
Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

(Received 29 November 1984; accepted for publication 21 March 1985)

In an earlier investigation of the class of shear-free expanding (or contracting) irrotational perfect fluids obeying a barotropic equation of state $p = p(\mu)$, and satisfying the field equations of general relativity, it was shown that the space-times form three distinct classes. In one class, the fluid acceleration is zero (i.e., the flow is geodesic), and the space-times are the well-known spatially homogeneous and isotropic Friedmann–Robertson–Walker (FRW) models. In the other two classes, the acceleration is nonzero, and the space-times are spatially anisotropic and are less familiar. One of these classes consists of the spherically symmetric Wyman solutions, whereas models that are plane symmetric, and either spatially or temporally homogeneous, constitute the final class. Analytic forms for these anisotropic space-time metrics were given, although in each case their exact determination would depend upon the solution of a single nonlinear ordinary differential equation, which has not been achieved. The purpose of the present article is to provide a pictorial description of the solutions of these equations, to depict qualitatively similar and distinct subclasses of solution, and hence to discover in some detail all possible global behaviors of the associated space-times. In all cases, it is found that, when the space-time is sufficiently extended, the fluid exhibits unphysical properties. The conclusion is that shear-free expanding (or contracting) relativistic perfect fluids that obey an equation of state $p = p(\mu)$ must be FRW, or else must be restricted to "local" regions, by means of a suitable extension in which at least one of the conditions defining the entire family is relaxed.

I. INTRODUCTION

For some time, the class of shear-free perfect fluids in general relativity has received attention (see, e.g., References 1–3, together with works cited therein). Perhaps the most challenging problem to be addressed is the question of whether or not it is true that if such a fluid obeys a barotropic equation of state $p = p(\mu)$, with $\mu + p \neq 0$, where μ is the energy density of the fluid and p is its isotropic pressure, then either the fluid is irrotational, or the fluid's overall volume expansion rate is zero. This is indeed the case under special conditions, such as for dust[1] ($p = 0$) and radiation[4] ($p = \frac{1}{3}\mu$), although to my knowledge the answer to this question has not been established in the general case. Other special situations in which the result is valid are given in Refs. 2 and 5, and in cited articles. Furthermore, at least as far as I am aware, there are no known examples where the result does not hold, a state of affairs that contrasts sharply with the corresponding situation in Newtonian gravity.[6] These facts lead one to conjecture that the result enjoys general validity.

With the evidence mounting that the result is universally valid in general relativity, it is tempting to investigate its consequences. Thus, assuming that the result is true, there naturally arise three mutually exclusive classes of shear-free fluids. Denoting the fluid's vorticity by ω and the volume expansion rate by θ, in the usual way,[6] we have the following alternatives:

$$\omega = 0 \neq \theta$$

or

$$\omega = \theta = 0$$

or

$$\theta = 0 \neq \omega.$$

Of course, if the conjectured result should turn out to be false, there would be a fourth possibility, viz., $\theta\omega \neq 0$, but this seems unlikely, and it is therefore omitted. From a physical point of view, the most interesting of the three alternatives is the first, since both stellar and cosmological investigations frequently center on fluids in a state of overall volume expansion or contraction, in which $\theta \neq 0$. This particular case was the subject of an earlier article,[3] in which it was shown that there were three distinct subclasses. In one subclass, labeled (i) in Ref. 3, the space-times involved are the well-known spatially homogeneous and isotropic Friedmann–Robertson–Walker (FRW) models; these can be characterized by the requirements that the fluid satisfy an equation of state $p = p(\mu)$ and that its motion not only be shear-free and irrotational, but also be geodesic.[6] In the other two subclasses, the fluid flow has nonzero acceleration, and is consequently anisotropic. In one of these subclasses, labeled (iii) in Ref. 3, the geometry is spherically symmetric, and the solutions are due to Wyman,[7] although they were not originally characterized as in the present manner; the global properties of these solutions are discussed in Sec. IV of the present article. In the other subclass, labeled (ii) in Ref. 3, the geometry is plane symmetric, and either spatially or temporally homogeneous. This subclass of solutions appeared to be new, and was unexpected, particularly in view of an earlier result[8] which contradicted its existence. As a consequence of this, rather little is understood of the nature of the solutions in this subclass; the global properties of these solutions are discussed in Sec. III of the present article. In each of the two subclasses of anisotropic solutions, Collins and Wainwright[3] showed that the space-time metric could be determined once the solution of a certain nonlinear ordinary differential equation was found. Neither differential equation has been solved

 0022-2488/85/082009-09$02.50

exactly, although the solutions in the spherically symmetric case can be expressed in terms of the Weierstrass elliptic function.[7] Since solutions have not been found exactly in terms of elementary functions, it becomes necessary to investigate their behavior qualitatively in some manner, in order to obtain a better appreciation of the nature of the space-times that they govern. The first steps in this direction have already been taken by Mashhoon and Partovi,[9] who in essence use inequalities to deduce that the fluid's behavior becomes unphysical if the space-time is sufficiently far extended, thus strengthening suspicions that were alluded to in the work of Collins and Wainwright.[3] However, the approach of Mashhoon and Partovi[9] is somewhat unsatisfactory, owing to an apparent reliance on the assumption of power series expansions, and to the fact that an incomplete set of conclusions is drawn. Further, no clear picture emerges of the full range of possibilities that can arise. This situation is rectified in the present article, by writing each governing differential equation in the form of a plane autonomous system, and then employing standard phase-plane analysis (Poincaré–Bendixson theory) to determine a vivid pictorial description for each case. From these diagrams, the various qualitatively similar families of solutions are clearly discernible, and the asymptotic behaviors are readily obtained, thus leading to a fairly complete understanding of the global features of the associated space-times.

The plan of this article is as follows. In Sec. II, the fundamental differential equations that govern the space-time metrics are provided and expressed in the form of plane autonomous systems. In Secs. III and IV, these systems are examined qualitatively, and various deductions are made about their asymptotic features and about the global behaviors of the associated space-times. The results are discussed and summarized in Sec. V, and, in particular, a comparison is made with the claims of Mashhoon and Partovi.[9] For the most part, I shall assume that Einstein's field equations of general relativity hold, with zero cosmological constant Λ, although in Sec. V I shall also briefly consider the effect of allowing Λ to be nonzero.

II. THE BASIC EQUATIONS

This section deals with the basic differential equations that determine the space-time metrics of anisotropic shear-free expanding (or contracting) irrotational perfect fluids which satisfy an equation of state $p = p(\mu)$ with $\mu + p \not\equiv 0$. From Theorem 1 of Collins and Wainwright,[3] there are two cases; these were labeled (ii) and (iii), although they will be denoted by (a) and (b), respectively, in the present article [case (i) of Ref. 3 consists of the Friedmann–Robertson–Walker models, which are not considered here, in view of the degree of their familiarity].

In case (a), comoving coordinates can be found in which the space-time metric is

$$ds^2 = \frac{C^2}{U^2}\left[-\frac{U'^2}{m^2}\,dt^2 + dx^2 + e^{-2x}(dy^2 + dz^2)\right], \qquad (2.1)$$

where $U = U(v) \not\equiv 0$, $v = t + x$, and

$$U'' + U' = -U^2, \qquad (2.2)$$

a prime denoting differentiation with respect to v. The fluid's energy density μ and pressure p are given by

$$\mu = (1/C^2)[3m^2 - 2U^3 - 3(U' + U)^2] \qquad (2.3)$$

and

$$\mu + p = 2U^4/(C^2U'). \qquad (2.4)$$

The quantitites C and m are nonzero constants, which, without loss of generality, are positive; the fluid's four-velocity is then $\mathbf{u} = (m|U|/C|U'|)(\partial/\partial t)$. The ratio of the derivatives of Eqs. (2.3) and (2.4) reveals that

$$1 + \frac{dp}{d\mu} = \frac{1}{3U'^2}[4U'^2 + UU' + U^3], \qquad (2.5)$$

where use has been made of (2.2). The space-times described by Eq. (2.1) are locally either spatially or temporally homogeneous according to whether $0 < U'^2 < m^2$ or $m^2 < U'^2$. They are locally rotationally symmetric, and admit a four parameter isometry group acting multiply transitively on the hypersurfaces $\{v = \text{const}\}$. This group contains a three-parameter subgroup of Bianchi type V that acts simply transitively on the hypersurfaces $\{v = \text{const}\}$. Note by differentiation of Eq. (2.2) that $U'' \equiv 0$ requires $UU' \equiv 0$, and so, using (2.2) again, $U \equiv 0$, which is inadmissible; hence the cases excluded by these inequalities, viz., $U'^2 = 0$ and $U'^2 = m^2$, can occur only at isolated values of v, and not in an open interval. This rules out the possibility of there being space-times in our discussion that are locally null homogeneous, while still allowing the occurrence of space-times that are spatially homogeneous in one region, and yet temporally homogeneous in another, with a common boundary being a single null homogeneous hypersurface (cf. Ref. 10). This type of possibility is considered in more detail in Sec. III. The "tilt" angle that the fluid flow vector makes with the normal to the hypersurfaces of homogeneity may be defined, in the case of spatial homogeneity, as β, where $-\cosh\beta$ is equal to the scalar product of the unit future-pointing hypersurface normal and the unit future-pointing fluid-flow vector.[8] In the present case, this means that

$$\cosh\beta = (1 - U'^2/m^2)^{-1/2}.$$

A similar definition can be given for temporal homogeneity, wherein the hyperbolic cosine is changed to a trigonometric cosine (cf. Ref. 11).

In case (b), comoving coordinates can be found in which the space-metric is

$$ds^2 = \frac{1}{U^2}\left[-\frac{U'^2}{At+B}dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right], \qquad (2.6)$$

where $U = U(v) \not\equiv 0$, $v = t + r^2$, and

$$U'' = U^2 \qquad (2.7)$$

$$\Leftrightarrow U'^2 = \tfrac{2}{3}U^3 - \tfrac{1}{3}A, \qquad (2.8)$$

a prime denoting differentiation with respect to v. The fluid's energy density μ and pressure p are given by

$$\mu = 3(Av + B) + 12UU' \qquad (2.9)$$

and

$$\mu + p = 20U^4/3U'. \qquad (2.10)$$

The quantities A and B are constants, and, without loss of generality, either $A = 0 < B$ or $B = 0 \neq A$. In the latter case, a coordinate singularity may arise at $t = 0$, details of which are provided in Sec. IV. The fluid's four-velocity is $\mathbf{u} = (At + B)^{1/2}(|U|/|U'|)(\partial/\partial t)$. The ratio of derivatives of Eqs. (2.9) and (2.10) reveals that

$$1 + \frac{dp}{d\mu} = \frac{4(5U^3 - 6A)}{3(8U^3 - 3A)}, \tag{2.11}$$

where use has been made of (2.7) and (2.8). The space-times described by Eq. (2.6) possess spherical symmetry (a particular case of local rotational symmetry) and are due to Wyman.[7] They admit a three-parameter isometry group acting multiply transitively on the two-surfaces $\{t = \text{const}, r = \text{const}\}$.

The differential equations that are to be examined qualitatively in Secs. III and IV are (2.2) and (2.7). Equation (2.7) differs somewhat from (2.2), in that a first integral, given by (2.8), is readily available. This means that the procedure for sketching the solutions in the U-U' phase plane is particularly straightforward. In each case, we shall denote the variable U' by Y. Thus Eq. (2.2) is written as the plane autonomous system

$$Y' = -Y - U^2, \quad U' = Y, \tag{2.12}$$

whereas Eq. (2.7) becomes

$$Y' = U^2, \quad U' = Y, \tag{2.13}$$

where, by (2.8)

$$Y^2 = \tfrac{2}{3}U^3 - \tfrac{1}{4}A. \tag{2.14}$$

In the next two sections, the results of a qualitative analysis of these systems of equations is provided. This leads to a pictorial representation in the U-Y phase planes of the solutions of these equations. The techniques that are employed are fairly standard methods of analyzing solutions of plane autonomous systems, details of which can be found in, e.g., Refs. 12–14.

III. THE PLANE-SYMMETRIC CASE (a): QUALITATIVE ANALYSIS OF THE SYSTEM (2.12)

In this case, the differential equations are

$$Y' = -Y - U^2, \quad U' = Y. \tag{2.12}$$

The associated space-times are plane symmetric and either spatially or temporally homogeneous. The solution curves are drawn in Fig. 1, with the arrows indicating the direction of increasing $v := t + x$. Each curve represents the evolution of a two-parameter family of space-times, as determined by the constants C and m in the metric.[3] There are two exceptional curves which either begin or end at the origin. These curves divide the remaining ones into two broad classes. In one of these, the curves start out in the top left of the figure, pass through the positive Y axis, move down to cross the positive U axis, then turn to traverse the negative Y axis, and finally turn again to end up in the bottom left of the figure. The behavior of the curves in the other class is less complicated. They simply start at the top left and curve around to cross the negative U axis, finishing up in the bottom left of the figure, with U remaining negative throughout. Thus, in-

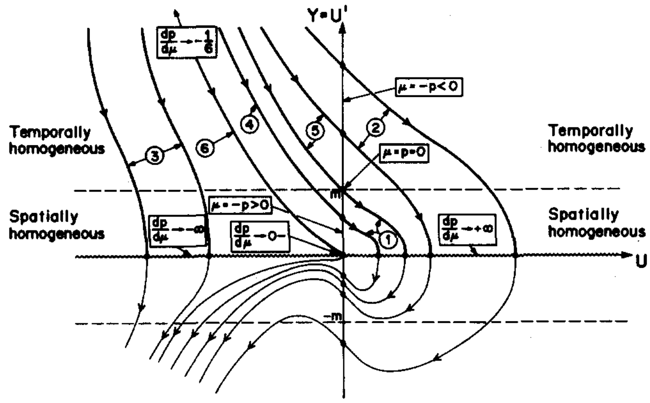


FIG. 1. Plane-symmetric, spatially or temporally homogeneous models. Solution curves describe the evolution of the variables U and U' in the space-time metric (2.1), as determined by the system (2.12). Each curve depicts the evolution of a two-parameter family of space-times. Arrows are drawn in the direction of increasing $v := t + x$. Solutions in the top half satisfy the physically relevant condition that $\mu + p > 0$; the space-times are locally spatially homogeneous if $|U'| < m$ and temporally homogeneous if $|U'| > m$. The curves are classified into six types, labeled 1–6, which correspond respectively to cases (1)–(6) in Sec. III. A solid dot (●) denotes the end of the evolution of the associated space-times. Excluding the origin, the U axis represents a space-time singularity at which $\mu + p$ and $dp/d\mu$ become infinite, and the U' axis represents a limit in which $\mu + p$ approaches zero, with μ tending to zero only if U' approaches the value m. Boxes provide information on some limits that indicate unphysical asymptotic behavior. The limit associated with the point $(0,m)$ is physically valid. For further details, see Sec. III.

cluding the solution represented by the point at the origin, and temporarily neglecting reference to the physical interpretation of the solutions, there are five qualitatively distinct types of behavior. However, the zero solution corresponds to $U \equiv 0$, which is inadmissible (see Sec. II). Moreover, by Eq. (2.4), $\mu + p$ becomes infinite whenever curves cross the U axis at nonzero U values, and, further, $\mu + p$ is negative if Y is negative and U is nonzero. From (2.3), it follows that μ becomes negative on the Y axis at values of Y that exceed m in magnitude. Thus it is clear that many of the solution curves will, at least for some interval in their evolution, represent space-times that would be regarded as unphysical. The question of the sign of quantities such as μ, p, or $\mu + p$ is of great relevance, but it is not as serious an issue as that of the infinite value of $\mu + p$ that occurs when the curves cross the U axis, since its occurrence within a finite proper distance or proper time would correspond to a space-time singularity, whereas if it were to arise within an infinite proper distance and proper time, the parts of the curve to either side of the U axis would be depicting the asymptotic behavior of two distinct space-times. Thus, in any case, a curve that crosses the U axis in Fig. 1 actually represents two separate space-times. Two other physically important aspects must also be addressed. One of these is the question of whether the associated space-time is spatially or temporally homogeneous, i.e., of whether or not $Y^2 < m^2$. The other question is whether there are solution curves in the U-Y plane that pass through particular points, which in the associated space-times are an infinite proper distance or proper time away, and whether there are points at infinity in the U-Y plane which refer to finite proper distances or times in the space-time. For example, since the coordinates are comoving, Eq. (2.1) implies that the proper time of a fluid particle is measured by integrating CU'/mU with respect to t, at constant x, and so the proper time behaves as $\ln|U|$, from which it follows that points on the Y axis should be regarded as being at infinity as measured by the fluid particles. By referring to the metric (2.1), it is apparent that special consideration of this sort to distances and times must be made only in the cases where U or Y approach either zero of infinity; a detailed discussion of these various possibilities is provided below, when the physical properties of the space-times are examined in detail.

The qualitatively distinct types of solution may now be examined, subject to the requirements that $U\not\equiv 0 \Leftrightarrow Y^2 + U^2 > 0$ and $\mu + p > 0 \Leftrightarrow Y > 0$. This consideration is also made by reference to the behavior of the remaining curves, not only with regard to the region defined by the restriction $0 < Y < m$, since this is of relevance to the character of the homogeneity of the associated space-times, but also with regard to their intersection with the (positive) Y axis. It is found that there are now six distinct classes of solutions, and hence of space-times, that are of interest. These are given by (1) curves which start on the Y axis with $0 < Y \leqslant m$, and end on the U-axis at some positive value of U; (2) curves which start on the Y axis with $Y > m$, and end on the U-axis at some positive value of U; (3) curves which begin in the top left of the figure and which end on the negative U axis; (4) curves which begin in the top left of the figure and terminate on the positive Y axis with $Y < m$; (5) curves which begin in the top left of the figure and terminate on the positive Y axis with $Y > m$; and finally (6) the curve that starts in the top left of the figure and runs into the origin.

The cases (1)–(6) will eventually be examined in turn, with conclusions drawn about the nature of the space-times involved. However, before embarking on a case-by-case study, some general observations are worthwhile. First, the space-times in cases (2), (3), (4), and (6) are initially temporally homogeneous, and the fluid evolves into a spatially homogeneous region. Associated with this transition is a "whimper" singularity,[10] which the fluid avoids. Next, by considering the various possibilities at large negative U, it is found that $Y^2 \sim -\frac{2}{3}U^3$, from which, using (2.12), it follows that $U \to -\infty$ as v approaches some finite value, v_*, say. By Eqs. (2.4) and (2.5), this means that $\mu + p \to +\infty$ and that $dp/d\mu \to -\frac{1}{6}$ (and hence $\mu \to +\infty$ and $p \to -\infty$) as $v \to v_*$. From the metric (2.1), it is seen that this state of affairs is encountered by the fluid in infinite proper time, although in each hypersurface $\{t = \text{const}\}$ orthogonal to the fluid flow, there is a singularity located a finite proper distance from any fluid particle. The approach of curves to points on the positive Y axis has already been mentioned. If $Y \to Y_0 > 0$ and $U \to 0$, then by Eq. (2.12) this occurs within a finite value of v, but, as previously mentioned, within an infinite proper time for the fluid particles. By Eq. (2.3), the energy density μ approaches $(3/C^2)(m^2 - Y_0^2)$, which shows that in the limit, μ is positive, negative, or zero, according as Y_0 is less than, exceeds, or equals m. Equation (2.4) shows that $\mu + p$ tends to zero, while it follows that $dp/d\mu \to \frac{1}{3}$ from (2.5). By considering the metric (2.1), it is found that this limit corresponds to space-time regions that are an infinite proper distance away from the fluid particles, as measured in any hypersurface $\{t = \text{const}\}$ orthogonal to the flow. Reference has also been made to curves that approach the positive or negative U axis, at $U = U_0$, say. By Eqs. (2.4) and (2.5), it follows that $\mu + p \to +\infty$ with $dp/d\mu \to \pm\infty$. According to Eq. (2.3), the limiting value of μ is finite, and its sign depends on the sign of the quantity $3m^2 - 2U^3 - 3U^2$. If $3m^2 > 1$, this cubic has only one real zero, U_*, say, and the limiting sign of μ is the same as that of $U_* - U_0$. If $3m^2 = 1$, the cubic has a single zero at $\frac{1}{2}$ and a double zero at -1, from which the limiting sign of μ is negative if U_0 is greater than $\frac{1}{2}$, zero if U_0 is equal to $\frac{1}{2}$ or -1, and negative otherwise. If $3m^2 < 1$, the cubic possesses three distinct real zeros, U_1, U_2, and U_3, where $U_1 < U_2 < 0 < U_3$, and the limiting value of μ is positive if $U_2 < U_0 < U_3$ or if $U_0 < U_1$, zero if $U_0 = U_1$, U_2 or U_3, and negative otherwise. From Eq. (2.12) it can be seen that this limit is approached as v tends to a finite value, and from the metric (2.1) it is deduced that this occurs within a finite proper time for the fluid, and within a finite proper distance on the $\{t = \text{const}\}$ sections. Finally, consideration must be given to the curve that runs into the origin. A study of the various possibilities that are allowed by Eq. (2.12) shows that $Y \sim -U$, and hence that this limit occurs as $v \to +\infty$. Moreover, by Eqs. (2.3)–(2.5), it follows that $\mu \to 3m^2/C$, $\mu + p \to 0$, and $dp/d\mu \to 0$ from below. The metric (2.1) shows that this limit occurs within an infinite proper time for the fluid, and corresponds to an infinite proper distance in the slices $\{t = \text{const}\}$. The detailed nature of the six types of space-times can now be discussed.

In case (1), the space-times are globally spatially homogeneous. The fluid starts its history in the infinite past, with $\mu \geqslant 0$ and $p \leqslant 0$, in such a way that $\mu + p = 0$ initially, and $dp/d\mu = \frac{1}{3}$. The tilt angle is nonzero, and infinite if and only if the initial value of μ is zero [corresponding to the particular curve 1 issuing from the point $(0,m)$ in Fig. 1]. The quantity $dp/d\mu$ has positive sign throughout the evolution of each solution. To the future, the fluid runs into a singularity (within a finite proper time), at which the limiting tilt angle is zero, and the energy density is finite (positive, negative, or zero), but at which the pressure becomes infinite, and associated with this is the fact that $dp/d\mu$ is infinite. Both asymptotic behaviors can be regarded as generally unphysical, the former on the grounds of negative pressures, and the latter by the fact that the speed of sound, being governed by $dp/d\mu$, is infinite. However, there is one interesting allowable limit, this occurring in the remote past in the case when the limiting value of μ is zero. This corresponds to the particular curve 1 in Fig. 1 which starts at $(0,m)$. A detailed calculation using the derivative of (2.3) with respect to v, and employing (2.2), shows that $d\mu/dv = 6U^3/C^2$, so it follows that, for this particular case, the energy density increases to positive values, thus confirming the physical validity of the limit.

In case (2), the fluid starts in the infinite past in a temporally homogeneous region, and evolves into a spatially homogeneous region, with the onset of a "whimper" singularity.[10] The asymptotic behaviors in this case are very similar to those in case (1), the main difference being that the early values of the energy density are negative, and that the initial pressure is positive. Here again, $dp/d\mu$ is positive during the entire history of the fluid.

In case (3), there is also an initial temporally homogeneous region, with evolution to a spatially homogeneous space-time, and the future behavior is very similar to that of cases (1) and (2), except that the final limit of $dp/d\mu$ is *negative* and infinite. However, there is a significant difference in the behavior at early times. This is in the infinite past of the fluid, but the initial energy density of the fluid is infinite, and the limit is unphysical, since the equation of state is such that $dp/d\mu$ is negative. In fact, $dp/d\mu$ stays negative for all time.

In case (4), the early stages are much the same as in case (3), and while the fluid evolves into a spatially homogeneous region, it does so for an infinite proper time, and the pressure has a finite negative limit, thus yielding an unphysical regime. The tilt angle never approaches zero. The quantity $dp/d\mu$ changes sign once as the fluid evolves.

In case (5), the initial behavior is similar to that of case (4), but the fluid never evolves into a spatially homogeneous region. Instead, the fluid exhibits, in its infinite future, a behavior reminiscent of the *early* stages of the case (2) solutions, i.e., the energy density and the pressure have finite limits, such that their sum is zero. In this case, however, the limiting value of the energy density is negative or zero; a calculation along the lines of that employed in case (1) shows that the particular case of the zero limit is of physical validity, since this limit is approached through positive values. As in case (4), $dp/d\mu$ changes sign once as the fluid evolves.

Finally, there is case (6). This is qualitatively similar to case (3), except that the fluid evolves indefinitely, and its energy density μ and pressure p approach finite limits, with $\mu > 0$ and $p > 0$, in such a way that $\mu + p$ tends to zero, and such that $dp/d\mu$ approaches zero through negative values, this behavior being very unphysical. In fact, $dp/d\mu$ stays negative during the entire evolution.

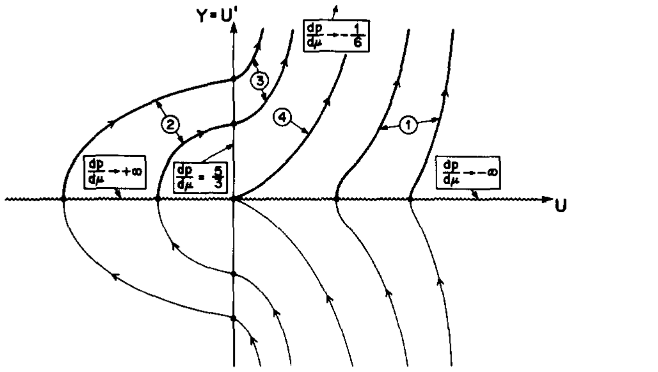


FIG. 2. Spherically symmetric Wyman models. Solution curves describe the evolution of the variables U and U' in the space-time metric (2.6), as determined by the system (2.13). Each curve depicts the evolution of a one-parameter family of space-times. Arrows are drawn in the direction of increasing $v: = t + r^2$, and for clarity the possibility of continuation through a coordinate singularity is ignored. The constant A in the first integral equation (2.14) is positive, negative, or zero, according as curves intersect the U axis at a value of U that is positive, negative, or zero. Solutions in the top half satisfy the physically relevant condition that $\mu + p > 0$. The curves are classified into four types, labeled 1–4, which correspond, respectively, to cases (1)–(4) in Sec. IV. A solid dot (●) denotes the end of the evolution of the associated space-times. Excluding the origin, the U axis represents a space-time singularity at which $\mu + p$ and $dp/d\mu$ become infinite, and the U' axis represents a limit in which $\mu + p$ approaches zero. Boxes provide information on some limits that indicate unphysical asymptotic behavior. A more complicated pattern of evolution, involving reversal of the direction of the arrows, is also possible. For further details, see Sec. IV.

IV. THE SPHERICALLY SYMMETRIC "WYMAN" CASE (b): QUALITATIVE ANALYSIS OF THE SYSTEM (2.13)

In this case, the differential equations are

$$Y' = U^2, \quad U' = Y, \tag{2.13}$$

which has the first integral

$$Y^2 = \tfrac{2}{3}U^3 - \tfrac{1}{4}A. \tag{2.14}$$

The associated space-times are spherically symmetric. The solution curves are drawn in Fig. 2, with the arrows indicating the direction of increasing $v: = t + r^2$. Each curve represents the evolution of a one-parameter family of space-times, as determined by a constant of integration.[3] For simplicity, in the case $A \neq 0 = B$, discussion will be deferred of the issues associated with a coordinate singularity at $t = 0$. As in the previous section, a preliminary study of the qualitatively distinct types of solutions reveals some valuable information, but this in itself is incomplete, and a further examination involving physical considerations becomes necessary. In preparation for this, it should first be noted that solutions of the system of equations (2.13) have one of five types of behavior. These are depicted in Fig. 2 by the solution at the origin, by the two curves that either leave or run into the origin and that extend to infinite values of U and Y, and by the remaining curves, which divide into two types of behavior, depending on whether intersection with the U axis occurs at positive or negative values of U. Excluding the degenerate case $U \equiv Y \equiv 0$, all curves have the same qualitative character at large distances from the origin; they behave approximately as $Y^2 = \frac{2}{3}U^3$, which can be seen either by a complete examination of all the possibilities as U and Y become infinite, in a manner similar to that of Sec. III, or more simply by direct referral to the first integral equation (2.14). It is also seen by Eq. (2.14) that curves meet the U axis at positive, negative, or zero values of U, according as A is positive, negative, or zero. Again, the zero solution corresponds to $U \equiv 0$, which is inadmissible (see Sec. II). Also, by Eq. (2.10), $\mu + p$ becomes infinite whenever curves cross the U axis at nonzero U values, and $\mu + p$ is negative if $Y < 0$. Moreover, reference to the metric (2.6) shows that special consideration must be given to the cases when U or Y approach either zero or infinity. In fact, unlike the situation in the previous section, care must also be taken to allow for the possibilities of $At + B$ becoming either zero or infinite. This particular point will be addressed later, specifically at the end of this section.

Taking into account these physical aspects, and in particular rejecting the zero solution and those solutions in which Y is always negative, along much the same lines as in Sec. III, it is seen that there are now basically four qualitatively distinct classes of solutions, and hence of space-times, that are of interest. These are given by (1) curves which start on the U axis with $U > 0$ and end in the top right of the figure; (2) curves which start on the U axis with $U < 0$ and terminate on the positive Y axis; (3) curves which begin on the positive Y

axis and end in the top right of the figure; and finally (4) the curve that starts at the origin and ends in the top right of the figure.

The cases (1)–(4) will soon be examined in turn, with conclusions drawn about the natures of the associated space-times. Prior to this examination, some general observations will first be made. It follows from earlier remarks on the relationship between the sign of A and the intersection of the curves with the U axis that in case (1), where $A>0$ (and $B=0$), while in cases (2) and (3), $A<0$ (and $B=0$), and in case (4), $A=0$ (and $B>0$). As has been observed already, for large U, $Y^2 \sim \frac{2}{3}U^3$, and so it follows from (2.10) that $\mu + p \rightarrow +\infty$. By (2.13), this occurs as v tends to a finite limit, v_0, say, from below. In case (1), where $A>0$, the requirement that the metric (2.6) be Lorentzian imposes the restriction that $t>0$, and similarly in cases (2) and (3), where $A<0$, it follows that $t<0$. Thus in case (1), indefinitely large values of U are reached along the fluid flow lines (on which r is constant) as t approaches a limit t_0 from below, and since t is constrained to be positive, t_0 must be positive. On the other hand, in case (3), the limiting value of t along a fluid flow line could be either negative or zero. The possibility of having a zero limiting value of t does not in fact affect any of the qualitative conclusions that are drawn below, although *a priori* this is conceivable, as can be seen from a discussion of whether or not this limit occurs within a finite proper time for the fluid. Using the metric (2.6), it is found that in both cases (1) and (3) the limit occurs within an infinite proper time, since the integral of $U'/Ut^{1/2}$ with respect to t diverges as t tends to t_0, regardless of whether or not t_0 is zero. In case (4), the situation is easier to examine, since $A=0$, and so the proper time behaves as $\ln U$, which becomes infinite. Similarly, it may be shown that U and Y become infinite within a finite proper distance of any one fluid particle, on each hypersurface $\{t = \text{const}\}$ orthogonal to the flow. By Eq. (2.11), it follows that $dp/d\mu$ approaches $-\frac{1}{6}$ for large U, while from (2.9), $\mu \rightarrow +\infty$. The limiting situation as curves approach the positive or negative U axis will next be considered. From (2.10), $\mu + p \rightarrow +\infty$, and by (2.13), this occurs as v tends to a finite value, from above. From the metric (2.6), it is found that this corresponds to a finite proper fluid time; again, a special calculation must be performed in the case when the coordinate time t tends to zero, which is now possible in case (1), but there is no qualitative difference in the conclusion. This limit, in which Y tends to zero, and U approaches a nonzero finite value, also corresponds to a singularity on each hypersurface $\{t = \text{const}\}$ orthogonal to the fluid flow, since the proper distance according to (2.6) is finite. By Eq. (2.14), the limiting value of U is $(\frac{3}{8}A)^{1/3}$, and so, by (2.11), $dp/d\mu \rightarrow +\infty$ when $A>0$, and $dp/d\mu \rightarrow -\infty$ when $A<0$. The limiting value of μ itself is positive, negative, or zero. If U tends to zero for finite (positive) Y, then A and t must both be negative. From Eq. (2.13), it is seen that this state of affairs is reached as v tends to a finite limit. If this limit is denoted by v_1, it is found that for $U>0$, $v \rightarrow v_1$ from above, whereas if $U<0$, $v \rightarrow v_1$ from below. Thus in this latter case the question arises of whether or not a limiting coordinate time value of zero will make any qualitative difference in a computation of the proper time along the fluid flow. It is found that it has no such effect: the proper time is always infinite. By Eqs. (2.10) and (2.11), $\mu + p$ tends to zero (from above), and $dp/d\mu$ approaches $\frac{5}{3}$; by (2.9), the quantity μ approaches a finite limit. This situation is reached at infinity in the $\{t = \text{const}\}$ slices. Finally, with regard to case (4), a study is made of the solution for which both U and Y tend to zero. Here, $A=0$, so it is clear from (2.6) that this takes an infinite proper fluid time in which to occur. Moreover, by Eqs. (2.13) and (2.14), it takes place as $v \rightarrow -\infty$. By (2.9)–(2.11), it is found that $\mu \rightarrow 3B$, $\mu + p \rightarrow 0$, and $dp/d\mu \equiv -\frac{1}{6}$. The fact that $v \rightarrow -\infty$ means that, unlike all other cases so far encountered, either in this section or in Sec. III, this limit does not pertain to a situation in *any* slice $\{t = t_*\}$, where t_* is constant, since then v is bounded below by t_*, and so cannot approach indefinitely large negative values. The detailed nature of the four types of space-times can be now be discussed.

In case (1), the fluid starts its evolution at a finite time in the past, at a singularity at which the fluid pressure p is infinite, and yet the energy density μ is finite (positive, negative, or zero). Associated with this is the fact that $dp/d\mu$ if infinite (and negative). The fluid evolves for an infinite time, and asymptotically μ and p become infinite, with $\mu>0$ and $p<0$, and with $dp/d\mu$ approaching $-\frac{1}{6}$. Both extremes are quite unphysical, since $dp/d\mu$ is asymptotically negative in both directions. In fact, $dp/d\mu$ is negative throughout the fluid's evolution.

In case (2), the early stages are much as in case (1), although $dp/d\mu$ is initially infinite and *positive*. To the future, the fluid still evolves for an infinite time, but now $\mu + p$ tends to zero, with μ having a finite limit (positive, negative, or zero), and with $dp/d\mu$ approaching $\frac{5}{3}$. This asymptotic state is also unphysical, since the speed of sound in the fluid would exceed that of light. The quantity $dp/d\mu$ is positive for all times.

In case (3), the initial stages occur in the infinite past, and are similar to the *final* stages in case (2). The fluid evolves into the same sort of future as in case (1). As it does so, the quantity $dp/d\mu$ changes sign once.

Finally, there is case (4), in which the equation of state is such that $dp/d\mu$ is always equal to $-\frac{1}{6}$. The early stages occur in the infinite past, with the initial values of μ and $\mu + p$ being $3B$ and 0, respectively. The final stages are similar to those in case (1).

The above considerations have been made largely without reference to the term $At + B$ in the metric (2.6). This is of particular importance only when $A \neq 0$ (and $B=0$, without loss of generality). It is possible for t to become zero in the course of evolution, so that the term then vanishes, although, as shown by Mashhoon and Partovi,[9] this is merely a coordinate singularity for the metric (2.6), and is removed by the transformation $t = A\tau^2$, thus yielding an analytic extension. If $A>0$, it may be assumed that locally $\tau > 0$, and that τ increases to the fluid's future. Hence, into the past (i.e., following curve 1 in Fig. 2 against the direction of the arrows), the fact that $v = A\tau^2 + r^2$ tends to a finite limit as U' approaches zero means that, at least for sufficiently large fixed r, τ must reach the value zero *before* encountering the singularity on the U axis. Retaining the condition that τ decreases to the past along the fluid flow, it follows that v must *increase* further into the past. Such evolution would be depicted in Fig. 2 by a point that comes in from infinity along curve 1, and which retraces its path before it encounters the U axis. Similarly, if $A < 0$, it may be assumed that locally $\tau < 0$, and that τ increases to the fluid's future, and so the evolution would be depicted in Fig. 2 by a point that moves up along curve 2, and retraces its path before it meets with the Y axis, or by a point that moves up curve 3, and retraces its path to return to the Y axis, instead of approaching infinity. Now, however, the fact that $A < 0$ means that this only takes place for certain values of the integration constant, and then only for sufficiently small fixed values of r. There are no similar effects for which account must be taken in the event that t become infinite, since for $A \neq 0$ and fixed r, the range of values of v, and hence of t, is bounded. Allowance for this type of behavior does not affect earlier conclusions concerning the unphysical behavior of the fluid on a global scale.

V. DISCUSSION AND CONCLUSION

It will be seen from the foregoing examinations of the asymptotic behaviors of the shear-free plane-symmetric solutions (Sec. III) and of the spherically symmetric Wyman models (Sec. IV) that the fluid typically exhibits unphysical characteristics. The question of what actually constitutes a "reasonable" equation of state is debatable, but one frequently accepted criterion is the dominant energy condition,[15] which requires that the inequality $-\mu \leqslant p \leqslant \mu$ hold for a perfect fluid. However, this condition alone is somewhat insufficient. It is also reasonable to suppose that $dp/d\mu$ satisfies the inequality $0 \leqslant dp/d\mu \leqslant 1$, since the lower limit is required for local mechanical stability, while the upper limit expresses the condition that the speed of sound in the fluid should not exceed that of light.[6] In addition, kinetic theory considerations lead some authors to invoke a "positive pressure" criterion, according to which $p \geqslant 0$ (see, e.g., Ref. 16). In the global extensions of all of the models considered herein, $dp/d\mu$ violates the inequality $0 \leqslant dp/d\mu \leqslant 1$, by becoming either negative or infinite, or both. In some solutions, $dp/d\mu$ is *always* negative, as in cases (3) and (6) of Sec. III and in cases (1) and (4) of Sec. IV, and these solutions are totally unrealistic. On the other hand, there are solutions in which $dp/d\mu$ is positive throughout the entire history of the fluid, such as in cases (1) and (2) of Sec. III and in case (2) of Sec. IV, and there are also solutions in which $dp/d\mu$ changes sign, as in cases (4) and (5) in Sec. III and in case (3) in Sec. IV. As already observed, these models are not realistic throughout their entire history, but it is important to recognize that, by appropriate choice of initial conditions, it might very well be possible to arrange for every preconceived notion of reasonable physical behavior to be satisfied within a sufficiently local region of space-time.

A comparison will now be made with the claims of Mashhoon and Partovi,[9] who base their judgement of the reasonability of the equation of state on the signs of μ, of p, and of either $\mu - 3p$ or $\mu - p$; that all these should be nonnegative is regarded as a necessary condition for the equation of state to be physically reasonable. Their discussion of the plane-symmetric case may be followed in the notation of the present article by writing $Z = -U$ and $\alpha = 1$. First, they provide an argument which purports to prove that, in all cases, $U \rightarrow -\infty$ within a finite value of v to the past. This is done by comparing Eq. (2.2) with the differential equation $U'' = -U^2$, noting from (2.4) that the physical requirement that $\mu + p > 0$ implies that $Y = U' > 0$. While it is true, from the point of view of differential inequalities alone, that this means that $U \rightarrow -\infty$ to the past, it must also be borne in mind, as was done in Sec. III of the present article, that the initial value of U could be positive, in which case, as U decreased, there would first be a value of v at which U would be zero, and that this would signify, through the metric (2.1), an end of the fluid's evolution. The curves in Fig. 1 clearly attest to the correct predictions of the differential inequality, and to the caution that must be exercised in a consideration of initial values of U (in particular, compare the behavior of curves 3–6 with curves 1 and 2). Next, it is deduced by Mashhoon and Partovi[9] that *if* $U \rightarrow -\infty$, and *if* Y^{-1} is analytic in $(-U)^{-1/2}$, then the fluid is unphysical, because $p/\mu \rightarrow -\frac{1}{6}$. As far as I am aware, no justification was made for the analyticity assumption, and the present considerations serve to show its superfluousness: the conclusion relating to the behavior of p/μ is in agreement with those results of Sec. III that pertain to the early stages of the solutions in cases (3)-(6), when $U \rightarrow -\infty$. Furthermore, attention is focused by Mashhoon and Partovi[9] on the finiteness of the *coordinate* values in this limit, whereas the issue of the finiteness or otherwise of the fluid proper time, or of the associated proper distances in the hypersurfaces $\{t = \text{const}\}$, is the physically relevant issue. In particular, the hypersurfaces $\{v = \text{const}\}$ are *timelike* for large negative U, and the limiting case is not correctly described as a hypersurface, since it does not relate to a feature actually in the space-time. These considerations of proper time and distance were addressed in Sec. III. Mashhoon and Partovi[9] also discuss the nature of those solutions in which $U \rightarrow 0$. It is correctly claimed that there are two allowable possibilities. In one, $U \rightarrow 0$ as $v \rightarrow +\infty$, in such a way that $Y = U' \rightarrow 0$ also. This corresponds to the final stages of case (6) in the present article. It can be seen from Fig. 1 that there is indeed only one solution with this behavior, as asserted by Mashhoon and Partovi.[9] In the other possibility, $U \rightarrow 0$ within a finite value of v, in such a way that $Y = U'$ approaches a finite nonzero value, with $dp/d\mu \rightarrow \frac{1}{3}$ and $\mu + p \rightarrow 0$. This corresponds to the early stages of solutions in cases (1) and (2), and to the late stages in cases (5) and (6). As discussed by Mashhoon and Partovi,[9] the fluid energy density and pressure will in general tend to nonzero limits whose sum is zero, which is regarded as unphysical by the requirement that both μ and p should be non-negative. However, it is also pointed out that there is a special case in which both μ and p tend to zero, thus allowing the possibility that the limit could be physical. Such a limit corresponds to both the early stages of the particular solution in case (1), which is represented in Fig. 1 by the curve 1 that issues from the point $(0,m)$, and the late stages of the particular solution in case (5), depicted in Fig. 1 by the particular curve 5 that runs into the point $(0,m)$. Such allowable limits were treated in Sec. III. These claims of Mashhoon and Partovi[9] appear to rely on the assumption of the validity of power series expansions of a certain type, and this assumption is somewhat unsubstantiated (cf. footnote 15 in the bibliography of Ref. 9). It is clear that this approach fails to yield *every* allowable limit, since the late stages of solutions in cases (1), (2), and (3), in which $Y = U'$ tends to zero and U approaches a nonzero finite value, are not discovered, and it appears to be entirely fortuitous that Mashhoon and Partovi[9] succeeded in obtaining what corresponds to either the early or the late stages in all six cases. It is quite conceivable that not every solution would have an asymptotic form which happens to coincide with the particular ones that were treated using power series.

The treatment by Mashhoon and Partovi[9] of the spherically symmetric Wyman solutions is discussed in the present notation by writing $\rho_{MP} = Kr$, $Y_{MP} = KU$, $t_{MP} = K^2 t$, $u_{MP} = K^2 v$, $C_{MP} = B$, and $\xi_{MP} = -A/4K^2$, where $K = (\frac{2}{3})^{1/5}$ and the subscript "MP" denotes the notation of Mashhoon and Partovi.[9] First an argument is provided which purports to show that (in the case when $A \neq 0$) $U \to +\infty$ within a finite value of v to the future. This is done in a manner analogous to that in the plane-symmetric case. It is deduced from Eq. (2.14) that this limit is attained by all solutions of the differential equation, but again it is not recognized that the sign of the initial value of U is important. That this is so is made very clear by reference to Fig. 2. If the initial value of U is negative, then as v increases so also does U, and eventually U tends to zero within a finite value of v. By the metric (2.6), this signifies an end to the fluid's evolution. Thus Mashhoon and Partovi[9] completely miss the case (2) solutions. They do however correctly conclude that in the other solutions, viz., cases (1), (3), and (4), the pressure becomes negative as $U \to +\infty$, and that this occurs within a finite value of v; although mention is made of the relationship to a physical radius, it is not pointed out that this limit corresponds to a time in the infinite future for the fluid (in addition, as was the case in the plane-symmetric models, the limiting value of v cannot be said to determine a hypersurface). No attempt was made to discuss the physics of the *early* stages of these models, and it is worth noting that there are solutions in case (1) in which the initial values of μ and p *would* satisfy the algebraic criteria of Mashhoon and Partovi,[9] but which are nevertheless unreasonable on account of the fact that $dp/d\mu$ is negative throughout their history. Furthermore, no investigation was made of the details of the continuation through the coordinate singularity at $t = 0$ in solutions of cases (1)–(3). In case (1), for example, this offers the possibility of avoiding the unphysical asymptotic region in which $U \to +\infty$ and $dp/d\mu \to -\frac{1}{6}$, which is regarded as inevitable in Ref. 9. It should be noted that Srivastava and Prasad[17] began an investigation of the behavior of the fluid in the Wyman solutions. However, they appear to assume (in the present notation) that $d\mu/dr \leqslant 0$ and that in the course of evolution the function U always achieves values in a specified range.

It can be observed by referral to Sec. III that, for the plane-symmetric models, the evolution is governed by the behavior of the quantity U in the metric (2.1), as a function of the variable $v := t + x$. Since the coordinates are comoving, this permits a fairly direct assessment of the evolution along the fluid flow lines, by fixing the value of x and regarding t as the essential variable. On the other hand, the value of t may be fixed, and x regarded as the fundamental variable, so that the behavior of U as a function of v is interpreted in terms of features in the hypersurfaces $\{t = \text{const}\}$ that are orthogonal to the flow. Frequently, infinite proper time along the flow lines corresponds to infinite proper distance in these hypersurfaces, although this is not always the case. When U becomes large and negative, as happens in the early stages of solution in cases (3)–(6), the limiting behavior occurs within a *finite* proper distance in the hypersurfaces, but within an *infinite* proper fluid time. While the specific nature of this limit means that, sufficiently early on in the evolution, in each hypersurface $\{t = \text{const}\}$ the quantity $dp/d\mu$ becomes negative, and that consequently these regions of space-time are unphysical, it raises the question of principle of whether or not it is compelling in a cosmological setting to require that hypersurfaces transverse to the fluid flow should exhibit regular behavior (cf., for example, Refs. 1, 18, and 19). Similar remarks pertain to the spherically symmetric models of Sec. IV, although now an additional feature is present. Here, when U becomes large and positive, as occurs in cases (1), (3), and (5), infinite proper time for the fluid corresponds to a finite proper distance in the hypersurfaces $\{t = \text{const}\}$. Moreover, as mentioned in Sec. IV, in the early stages of case (4) solutions, $U \to 0$ as $v \to -\infty$, and while this corresponds to a limit in infinite proper time to the past for the fluid, it does not represent any limiting situation on a hypersurface $\{t = \text{const}\}$, on which v is always bounded below.

The conclusions of the present work are largely unaffected by the inclusion of a cosmological constant Λ in Einstein's field equations. The local effects of such a modification were addressed by Collins and Wainwright.[3] The quantities μ and p are replaced bty $\mu + \Lambda$ and $p - \Lambda$, respectively, and so, depending on the size of Λ, this might cause a change of sign in the limiting values of μ and p, but not, of course in $\mu + p$. However, the value of $dp/d\mu$ is unaffected by this transformation, and so it may be deduced that all of the solutions considered herein, when generalized to include a cosmological term Λ fail to be physically valid on a global scale, because such is the case when $\Lambda = 0$.

In conclusion, the following theorem, suggested in previous works,[3,9] is now hopefully unequivocally established.

Theorem: The spatially homogeneous and isotropic Friedmann–Robertson–Walker models are characterized globally as the only shear-free irrotational expanding or contracting perfect fluid solutions of Einstein's field equations of general relativity, with or without cosmological constant, in which the fluid satisfies a barotropic equation of state, $p = p(\mu)$, that is physically realistic everywhere, in the sense that $dp/d\mu$ satisfies the inequality $0 \leqslant dp/d\mu \leqslant 1$ throughout the space-time.

ACKNOWLEDGMENTS

I thank Helen Warren for her careful typing of the manuscript.

This work was partially supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada under Grant No. A3978.

[1] G. F. R. Ellis, J. Math. Phys. **8**, 1171 (1967).
[2] A. J. White and C. B. Collins, J. Math. Phys. **25**, 332 (1984).

[3]C. B. Collins and J. Wainwright, Phys. Rev. D **27**, 1209 (1983).
[4]R. Treciokas and G. F. R. Ellis, Commun. Math. Phys. **23**, 1 (1971).
[5]C. B. Collins, J. Math. Phys. **25**, 995 (1984).
[6]G. F. R. Ellis, "Relativistic Cosmology," in *General Relativity and Cosmology*, Course XLVII, Varenna, Italy, 1969, edited by R. K. Sachs (Academic, New York, 1971).
[7]M. Wyman, Phys. Rev. **70**, 396 (1946).
[8]A. R. King and G. F. R. Ellis, Commun. Math. Phys. **31**, 209 (1973).
[9]B. Mashhoon and M. H. Partovi, Phys. Rev. D **30**, 1839 (1984).
[10]G. F. R. Ellis and A. R. King, Commun. Math. Phys. **38**, 119 (1974).
[11]C. B. Collins, Commun. Math. Phys. **39**, 131 (1974).
[12]V. V. Nemytskii and V. V. Stepanov, *Qualitative Theory of Differential Equations* (Princeton U. P., Princeton, NJ, 1960).
[13]G. Sansone and R. Conti, *Non-Linear Differential Equations* (New York, Macmillan, 1964).
[14]S. Vojtášek and K. Janáč, *Solution of Non-Linear Systems* (Iliffe, London, 1969).
[15]S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge U. P., Cambridge, 1973).
[16]J. M. Stewart, Mon. Not. R. Astron. Soc. **145**, 347 (1969).
[17]D. C. Srivastava and S. R. Prasad, Gen. Relativ. Gravit. **15**, 65 (1983).
[18]G. F. R. Ellis, R. Maartens, and S. D. Nel, Mon. Not. R. Astron. Soc. **184**, 439 (1978).
[19]S. W. Goode and J. Wainwright, Mon. Not. R. Astron. Soc. **198**, 83 (1982).

Anisotropic spheres admitting a one-parameter group of conformal motions

L. Herrera[a)]
Departamento de Física, Facultad de Ciencias, Universidad Central de Venezuela, Caracas 1051, Venezuela

J. Ponce de León
Departamento de Física, Facultad de Ciencias, Universidad Central de Venezuela, Caracas 1051, Venezuela and Centro de Física, Instituto Venezolano de Investigaciones Científicas, Apdo. 1827, Caracas 1010-A, Venezuela

(Received 5 November 1984; accepted for publication 18 January 1985)

The Einstein equations for spherically symmetric distributions of anisotropic matter (principal stresses unequal), are solved, assuming the existence of a one-parameter group of conformal motions. All solutions can be matched with the Schwarzschild exterior metric on the boundary of matter. Two families of solutions represent, respectively, expanding and contracting spheres which asymptotically tend to a static sphere with a surface potential equal to $\frac{1}{3}$. A third family of solutions describes "oscillating black holes." All solutions possess a positive energy density larger than the stresses everywhere.

I. INTRODUCTION

In a recent paper[1] we have integrated the Einstein equations for perfect fluids under the assumption that the space-time admits, besides the spherical symmetry, a one-parameter group of conformal motions, i.e.,

$$\underset{\xi}{L}\, g_{\alpha\beta} = \psi g_{\alpha\beta}\,, \tag{1}$$

where the left-hand side is the Lie derivative of the metric tensor and ψ is an arbitrary function of the coordinates. As a result of that work it appears that the Einstein equations reduce to a system of ordinary differential equations for three unknown functions, which may be solved analytically in many cases.

In this paper we propose to carry out the same program for anisotropic matter. The motivation to undertake such a task is twofold.

(a) It will be shown below that, unlike the perfect fluid case, the anisotropic matter solutions may be matched with the Schwarzschild vacuum metric on the boundary of the matter.

(b) The introduction of anisotropic matter is suggested by some theoretical works on more realistic equation of state and stellar models[2,3] which indicate that compact objects could have anisotropic pressures. The origin of the anisotropy could be found in the existence of a solid core, in the presence of type P superfluid, or in the existence of an external field. Also, if the fluid is composed of two perfect fluids with different four-velocities, then the energy momentum tensor can be cast into the standard form for anisotropic fluids[4] (see, for example, the Landau model for He[4]). Finally it is worth noticing that the properties of anisotropic spheres may differ drastically from the properties of the isotropic.[5-8] We shall exhibit explicitly three families of solutions describing different self-similar evolution scenarios. Each family depends upon a function of the radial coordinate which has to satisfy certain physical requirements.

Two families of solutions represent expanding and contracting spheres, respectively. All solutions of these families tend asymptotically to a static sphere with a surface gravitational potential equal to $\frac{1}{3}$.

A third family of solutions represents spheres whose boundaries oscillate between the center and the horizon.

The paper is organized as follows. The field equations as well as the conventions used are included in Sec. II. In Sec. III we derive, from the fulfillment of the junction conditions on the boundary, a differential equation which describes the evolution of the boundary of the source (hereafter referred to as the surface equation). Three different solutions of this late equation are explicitly displayed. In Sec. IV we work out different models from the solutions of the surface equation and in Sec. V the results are discussed. Finally, some details of intermediate calculations are included in the Appendix.

II. THE FIELD EQUATIONS AND CONVENTIONS

Let us consider a nonstatic distribution of matter represented by an anisotropic fluid and which is spherically symmetric.

In comoving coordinates the line element may be written as[9]

$$ds^2 = e^{\nu}\, dt^2 - e^{\lambda}\, dr^2 - e^{\mu}\, d\Omega^2\,, \tag{2}$$

with

$$d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2\,, \quad x^{0,1,2,3} \equiv t,r,\theta,\phi\,,$$

where λ, ν, and μ are functions of r and t. For the energy momentum tensor we have the usual expression

$$T^{\mu}_{\nu} = (\rho + p_\perp) U^{\mu} U_{\nu} - p_\perp \delta^{\mu}_{\nu} + (p_r - p_\perp) X^{\mu} X_{\nu}\,, \tag{3}$$

with ρ, p_r, and U^{μ} denoting the energy density, the pressure in the direction of X_μ, and the four-velocity of the fluid, respectively, and X_μ and $p_\perp$ denoting a unit spacelike vector (in the radial direction) orthogonal to U^{μ} and the pressure on the two-space orthogonal to X_μ. Also, since we are in a comoving frame,

$$U^{\mu} = \delta^{\mu}_{0} e^{-\nu/2}\,. \tag{4}$$

Thus, the Einstein equations

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = 8\pi T_{\mu\nu} \tag{5}$$

a) Postal address: Apartado 80793, Caracas 1080-A, Venezuela.

 0022-2488/85/082018-06$02.50

read

$$-8\pi T^1_1 = 8\pi p_r = \tfrac{1}{2} e^{-\lambda}(\mu'^2/2 + \mu'\nu') - e^{-\nu}(\ddot{\mu} - \tfrac{1}{2}\dot{\mu}\dot{\nu} + \tfrac{3}{4}\dot{\mu}^2) - e^{-\mu}, \quad (6)$$

$$-8\pi T^2_2 = 8\pi p_\perp = \tfrac{1}{4} e^{-\lambda}(2\nu'' + \nu'^2 + 2\mu'' + \mu'^2 - \mu'\lambda' - \nu'\lambda' + \mu'\nu') + \tfrac{1}{4} e^{-\nu}(\dot{\lambda}\dot{\nu} + \dot{\mu}\dot{\nu} - \dot{\lambda}\dot{\mu} - 2\ddot{\lambda} - \dot{\lambda}^2 - 2\ddot{\mu} - \dot{\mu}^2), \quad (7)$$

$$8\pi T^0_0 = 8\pi\rho = e^{-\lambda}(\mu'' + \tfrac{3}{4}\mu'^2 - \mu'\lambda'/2) + \tfrac{1}{2} e^{-\nu}(\dot{\lambda}\dot{\mu} + \dot{\mu}^2/2) + e^{-\mu}, \quad (8)$$

$$8\pi T^1_0 = 0 = \tfrac{1}{2} e^{-\lambda}(2\dot{\mu}' + \dot{\mu}\mu' - \dot{\lambda}\mu' - \nu'\dot{\mu}) \quad (9)$$

(dots and primes denote differentiation with respect to t and r, respectively).

Next, we shall assume that the space-time admits a one-parameter group of conformal motions, i.e.,

$$\underset{\xi}{L}\, g_{\mu\nu} \equiv \xi_{\mu;\nu} + \xi_{\nu;\mu} = \psi g_{\mu\nu}, \quad (10)$$

where ψ is an arbitrary function of t and r. We shall further restrict the vector field ξ^α by demanding

$$\xi^\alpha U_\alpha = 0. \quad (11)$$

Then as a consequence of the spherical symmetry and from Eq. (11) we have

$$\xi^0 = \xi^2 = \xi^3 = 0. \quad (12)$$

Thus, using (2) and (12) we get from Eq. (10)

$$\nu'\xi^1 = \psi, \quad (13)$$

$$\xi^1{}_{,0} = 0, \quad (14)$$

$$\lambda'\xi^1 + 2\xi^1{}_{,1} = \psi, \quad (15)$$

$$\mu'\xi^1 = \psi \quad (16)$$

(comma denotes partial derivatives). It can be seen at once from (13) and (16) that

$$\nu - \mu = f_1(t), \quad (17)$$

where $f_1(t)$ is an arbitrary function of t.

Next, taking derivatives of (15) and (16) with respect to t, and using (14), we obtain the equation

$$\dot{\lambda}' = \dot{\mu}', \quad (18)$$

from which it follows that

$$\lambda - \mu = f_2(t) + g_1(r), \quad (19)$$

f_2 and g_1 being arbitrary functions of their arguments.

We still have the freedom to perform a coordinate transformation of the form[9]

$$t = t(\bar{t}), \quad r = r(\bar{r}).$$

Thus, without loss of generality we may choose $f_1(t) = g_1(r) = 0$, and then

$$\nu - \mu = 0, \quad \lambda - \mu = f(t). \quad (20)$$

Feeding (20) back into (15) and using (13) we obtain

$$\xi^1 = A = \text{const}, \quad \psi = A\nu'. \quad (21)$$

Expressions (20) and (21) contain all the implications derived from the existence of the conformal motion with ξ^μ orthogonal to U^μ.

Let us now turn to the field equations (6)–(9). Using (20) we get from Eq. (9)

$$2\dot{\lambda}' - \dot{\lambda}\lambda' = 0. \quad (22)$$

Introducing the new variable

$$Z \equiv e^{-\lambda/2},$$

Eq. (22) becomes

$$\dot{Z}' = 0,$$

whose solution has the form

$$Z \equiv e^{-\lambda/2} = h_1(r) + h_2(t), \quad (23)$$

where h_1 and h_2 are two unknown functions of their arguments. Using again Eq. (20) we obtain

$$e^{-\nu/2} = e^{-\mu/2} = e^{f(t)/2}[h_1(r) + h_2(t)]. \quad (24)$$

We can now write down the field equations (6)–(9) in terms of the functions $f(t)$, $h_1(r)$, and $h_2(t)$. We get

$$-8\pi T^1_1 = 8\pi p_r = [3h'^2_1(r) - 3\dot{h}^2_2(t)e^{f(t)}] + e^{-\lambda/2}e^{f(t)}[2\ddot{h}_2(t) - \dot{h}_2(t)\dot{f}(t)] + e^{-\lambda}e^{f(t)}[\ddot{f}(t) - \dot{f}^2(t)/4 - 1], \quad (25)$$

$$-8\pi T^2_2 = 8\pi p_\perp = [3h'^2_1(r) - 3\dot{h}^2_2(t)e^{f(t)}] - 2e^{-\lambda/2}[h''_1(r) - \ddot{h}_2(t)e^{f(t)}] + \tfrac{1}{2}\ddot{f}(t)e^{-\lambda}e^{f(t)}, \quad (26)$$

$$8\pi T^0_0 = 8\pi\rho = -[3h'^2_1(r) - 3\dot{h}^2_2(t)e^{f(t)}] + 2e^{-\lambda/2}[h''_1(r) + \dot{h}_2(t)\dot{f}(t)e^{f(t)}] + e^{-\lambda}e^{f(t)}[\dot{f}^2(t)/4 + 1], \quad (27)$$

and for the line element we have

$$ds^2 = \frac{e^{-f(t)}}{[h_1(r) + h_2(t)]^2}[dt^2 - e^{f(t)}dr^2 - d\Omega^2] \quad (28)$$

or

$$ds^2 = R^2(r,t)[dt^2 - e^{f(t)}dr^2 - d\Omega^2],$$

with

$$R(r,t) \equiv e^{-f(t)/2}/[h_1(r) + h_2(t)].$$

III. THE SURFACE EQUATION AND THE JUNCTION CONDITIONS

Since we are mainly interested in bounded sources we shall demand the line element (28) to match with the Schwarzschild exterior metric on the boundary of the source for any possible choice of the functions $f(t)$, $h_1(r)$, and $h_2(t)$. We recall that two regions of the space-time are said to match across a separating hypersurface (say S) if the first and the second fundamental forms are continuous across S (Darmois conditions).

Now, the line element outside the source will be given, in Schwarzschild-like coordinates T,R,θ,ϕ, by

$$dS^2_E = (1 - 2M/R)\,dT^2 - (1 - 2M/R)^{-1}dR^2 - R^2 d\Omega^2, \quad (29)$$

where the subscript E stands for exterior, M is the total mass, and

$$d\Omega^2 \equiv (d\theta^2 + \sin^2\theta\, d\phi^2).$$

In these coordinates the equation of the boundary takes the form

$$R = R_b(T), \tag{30}$$

where b stands for boundary.

Then the induced metric on the boundary surface (from the outside) is

$$(dS^2)_b^+ = \left[\left(1-\frac{2M}{R_b}\right) - \left(1-\frac{2M}{R_b}\right)^{-1}\left(\frac{dR_b}{dT}\right)^2\right] dT^2 - R_b^2\, d\Omega^2 \tag{31}$$

and the corresponding line element on the boundary, from the inside, reads

$$(dS^2)_b^- = \frac{e^{-f(t)}}{[h_1(r_0)+h_2(t)]^2}\,[dt^2 - d\Omega^2], \tag{32}$$

where we have used (28) and the fact that the equation of the boundary in the comoving coordinates (t,r,θ,ϕ) reads

$$r = r_0 = \text{const}.$$

Then, demanding the first fundamental form to be continuous across the boundary, we get at once

$$R_b(T) = e^{-f(t)/2}/[h_1(r_0)+h_2(t)] \tag{33}$$

and

$$\left[\left(1-\frac{2M}{R_b}\right) - \left(1-\frac{2M}{R_b}\right)^{-1}\left(\frac{dR_b}{dT}\right)^2\right]^{1/2} dT = R_b\, dt, \tag{34}$$

or, using the function U_b, given by[10]

$$U_b = U^\mu \frac{\partial R_b}{\partial x^\mu} = e^{-\nu/2}\frac{\partial R_b}{\partial t}, \tag{35}$$

we get

$$R_b\, dt = dT\left(1-\frac{2M}{R_b}\right)\left[1-\frac{2M}{R_b} + U_b^2\right]^{-1/2}. \tag{36}$$

Next, it can be shown by a straightforward calculation that the continuity of the second fundamental form across the boundary surface is equivalent to the conditions[10–13]

$$m(r_0,t) = M, \tag{37}$$

$$-(T_1^1)_b = P_r(r_0,t) = 0, \tag{38}$$

where $m(r,t)$ is the mass function introduced by Misner and Sharp[10]

$$2m(r,t) = e^{\mu/2}\left[1 + e^{-\nu}\left(\frac{\partial e^{\mu/2}}{\partial t}\right)^2 - e^{-\lambda}\left(\frac{\partial e^{\mu/2}}{\partial r}\right)^2\right]. \tag{39}$$

Using (25), (28), and (39), Eqs. (37) and (38) become

$$\begin{aligned} P_r(r_0,t) &= 3h_1'^2(r_0) - 3\dot h_2^2(t)e^{f(t)} + [h_1(r_0)+h_2(t)] \\ &\quad\times[2\ddot h_2(t) - \dot h_2(t)\dot f(t)]\, e^{f(t)} + [h_1(r_0) \\ &\quad + h_2(t)]^2 e^{f(t)}[\ddot f(t) - \dot f^2(t)/4 - 1] = 0, \end{aligned} \tag{40}$$

$$\begin{aligned} 2M &= e^{-f(t)/2}/[h_1(r_0)+h_2(t)] \\ &\quad\times[1 + \tfrac14(\dot f(t) + 2\dot h_2(t)/[h_1(r_0)+h_2(t)])^2 \\ &\quad - e^{-f(t)}h_1'^2(r_0)/[h_1(r_0)+h_2(t)]^2]. \end{aligned} \tag{41}$$

Thus, the Darmois conditions (which are equivalent to Lichnerowicz conditions[14]) amount to satisfying Eqs. (33), (40), and (41). Now, we can express (40) and (41) in terms of the function R_b to obtain

$$2\ddot R_b R_b - \dot R_b^2 = 3h_1'^2(r_0)R_b^4 - R_b^2, \tag{42}$$

$$\dot R_b^2 = 2MR_b - R_b^2 + h_1'^2(r_0)R_b^4. \tag{43}$$

It is worthwhile to point out that as a consequence of the fulfillment of the junction conditions [expressed by Eqs. (33), (40), and (41)] the matter must be anisotropic. In fact, in the case of isotropic (perfect) fluid, the functions $h_1(r)$, $h_2(t)$, and $f(t)$ satisfy the following set of equations[1] (see the Appendix):

$$2h_1''(r) + C_1h_1(r) = 2C_2, \tag{44}$$

$$\dot h_2(t)\dot f(t)e^{f(t)} - C_1h_2(t) = 2C_2, \tag{45}$$

$$(2\ddot f(t) - \dot f^2(t) - 4)\, e^{f(t)} = 4C_1. \tag{46}$$

It is a simple matter to prove that the last two equations are incompatible with Eqs. (40) and (41).

Let us now turn to the integration of Eqs. (42) and (43), which govern the evolution of the boundary. First of all observe that (43) is nothing but the first integral of Eq. (42). Consequently we only have to integrate the equation

$$\dot R_b^2 = 2MR_b - R_b^2 + h_1'^2(r_0)R_b^4. \tag{47}$$

In order to exhibit the general features of the solutions of this late equation (surface equation), it is useful to introduce the auxiliary function

$$V(R_b) \equiv 1/R_b^2 - 2M/R_b^3, \tag{48}$$

in terms of which the surface equation reads

$$\dot R_b^2 = R_b^4[h_1'^2(r_0) - V(R_b)]. \tag{49}$$

Thus the region of allowed values of the R_b is given by the inequality (see Fig. 1)

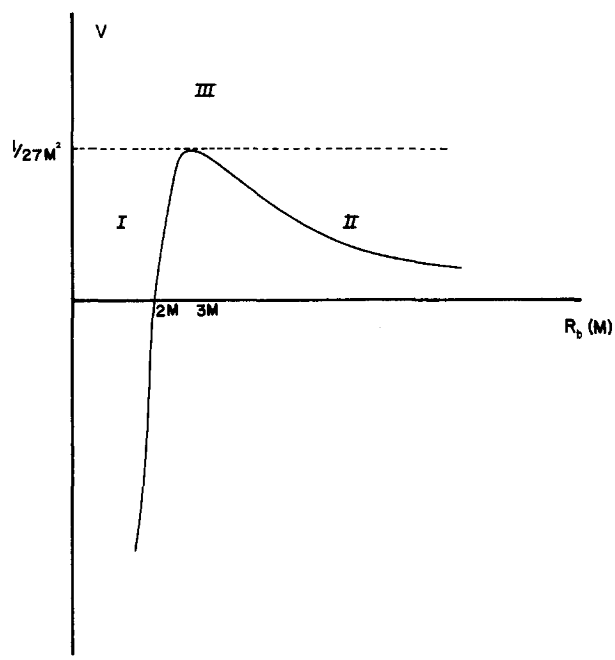


FIG. 1. V as a function of R_b. Regions I and II are separated from region III by the line $V = 1/27M^2$. There is a maximum of V for $R_b = 3M$.

$$h_1'^2(r_0) - V(R_b) \geqslant 0 . \tag{50}$$

Although the solutions of (49), in general, are expressed in terms of elliptic functions, it is possible, for some specific choices of $h_1'^2(r_0)$, to find solutions expressed solely in terms of elementary functions. Thus, for example, choosing $h_1'^2(r_0) = 1/27M^2$ we integrate (49) to give

$$R_b^{(1)} = 6M/(3\coth^2[(t-t_0)/2] - 1) \tag{51}$$

and

$$R_b^{(2)} = 6M/(3\tanh^2[(t-t_0)/2] - 1) . \tag{52}$$

The first solution [as given by (51)] represents an expanding boundary surface which asymptotically tends to a sphere with radius $R_b = 3M$ as $t \to \infty$. Furthermore we see that if we choose $t_0 = 0$, then the sphere is concentrated at the origin at the initial time $t = 0$. From these considerations it is obvious that the boundary surface should cross the horizon ($R_b = 2M$) for some finite value of the comoving time (say $t = t_g$), which is easily calculated from (51) to be

$$t_g = 2\,\mathrm{arccoth}(2/\sqrt{3}) . \tag{53}$$

The second solution [as given by (52)] represents a contracting boundary surface which also tends to a sphere of radius $R_b = 3M$ as $t \to \infty$. In order to exclude negative values of R_b in the interval $t \in (0,\infty)$, we may choose

$$t_0 = -2\,\mathrm{arctanh}(1/\sqrt{3}) ,$$

so that $R_b^{(2)}(0) = \infty$.

A third analytical solution of the surface equation may be found for the choice

$$h_1'^2(r_0) = 0 .$$

We get in this case

$$R_b^{(3)} = 2M\cos^2 \tfrac{1}{2}(t-t_0) . \tag{54}$$

This solution represents an oscillating sphere whose radius changes in the interval $[0,2M]$.

For other values of $h_1'^2(r_0)$ between zero and $1/27M^2$, there are, in general, two kinds of solutions. Solutions of region I (see Fig. 1) represent spheres oscillating between the singularity and some value of R_b [depending on the value of $h_1'^2(r_0)$] in the interval $[2M,3M]$. The solutions of region II represent spheres contracting from some initial value of R_b (which may be infinity), bouncing for some value of R_b [which depends on $[h_1'^2(r_0)]$ and returning to the initial configuration.

Finally, in region III for $h_1'^2(r_0) > 1/27M^2$, R_b may change in the interval $(0,\infty)$.

IV. THE SOLUTIONS

In this section we shall present some models which are constructed from the solutions of the surface equation of the preceding section. For the sake of simplicity we shall further restrict our solutions, with the choice $h_2(t) = 0$. With this condition, Eqs. (33), (42), and (43) read

$$e^{f(t)/2}h_1(r_0) = 1/R_b \equiv 1/R(r_0,t) , \tag{55}$$

$$e^{f(t)}\left[\ddot{f}(t) - \frac{\dot{f}^2(t)}{4} - 1\right] = -\frac{3h_1'^2(r_0)}{h_1^2(r_0)} , \tag{56}$$

$$\dot{f}^2(t) = 8Mh_1(r_0)e^{f(t)/2} + [4h_1'^2(r_0)/h_1^2(r_0)]\, e^{-f(t)} - 4 . \tag{57}$$

Using the above equations, we obtain from (25), (26), and (27)

$$8\pi P_r = 3h_1'^2(r) - 3\omega^2 h_1^2(r) , \tag{58}$$

$$8\pi P_\perp = 3h_1'^2(r) - 2h_1(r)h_1''(r) + \tfrac{1}{2}[\ddot{f}(t)/R^2] , \tag{59}$$

$$8\pi\rho = -3h_1'^2(r) + 2h_1(r)h_1''(r) + \omega^2 h_1^2(r) + \frac{2M}{R_b R^2} , \tag{60}$$

where

$$h_1'^2(r_0)/h_1^2(r_0) = \omega^2 ,$$

and as stated before

$$R(r,t) = e^{-f(t)/2}/[h_1(r) + h_2(t)] .$$

Now, in order to have a solution completely determined we have still to specify the function $h_1(r)$. The ideal approach to do that would be to know the relation between P_r and $P_\perp$ on physical grounds (an equation of state for the stresses). However, since this seems to be very difficult to carry out at the present we shall instead guess the function $h_1(r)$ from general (physical) considerations, e.g.,

$$h_1(r) > 0 , \quad \frac{\partial R}{\partial r} > 0 , \quad \rho > 0 , \quad \frac{\partial \rho}{\partial r} < 0 .$$

The first choice of $h_1(r)$ will be suggested by the condition of the positiveness of the energy density. A sufficient condition to meet this last requirement is

$$-3h_1'^2(r) + 2h_1(r)h_1''(r) = 0 , \tag{61}$$

from which

$$h_1(r) = 1/(Cr + B)^2 , \tag{62}$$

with C and B constants of integration. Next, imposing the regularity condition

$$R(0,t) = 0 , \tag{63}$$

we obtain

$$B = 0 , \quad R(r,t) = e^{-f(t)/2}C^2r^2 ,$$

$$\omega^2 = 4/r_0^2$$

and Eqs. (58)–(60) become

$$8\pi P_r = (12/C^4r^4r_0^2)[(r_0/r)^2 - 1] , \tag{64}$$

$$8\pi P_\perp = \frac{\ddot{f}(t)}{2R^2} = \frac{2}{C^4r^4r_0^2}\left[\frac{M}{2C^2}e^{3/2f(t)} - 2\right] , \tag{65}$$

$$8\pi\rho = (4/C^4r^4r_0^2)[1 + (M/2C^2)e^{3/2f(t)}] , \tag{66}$$

$$ds^2 = e^{-f(t)}C^4r^4[dt^2 - e^{f(t)}\,dr^2 - d\Omega^2] . \tag{67}$$

As a second example we take

$$h_1(r) = De^{-2(r/r_0)} ,$$

with $D = \text{const}$ and $\omega^2 = 4/r_0^2$. Then

$$R(r,t) = D^{-1}e^{-f(t)/2}e^{2(r/r_0)} , \tag{68}$$

and for the matter variables and the line element we get

$$8\pi P_r = 0 , \tag{69}$$

$$8\pi P_\perp = (1/R^2)[4D^2/r_0^2 + \tfrac{1}{2}\ddot{f}(t)] , \tag{70}$$

$$8\pi\rho = M/R^2R_b , \tag{71}$$

$$ds^2 = e^{-f(t)}D^{-2}e^{4(r/r_0)}[dt^2 - e^{f(t)}\,dr^2 - d\Omega^2] . \tag{72}$$

Two remarks are in order at this point.

(a) Observe that the function $R(r,t)$ as given by (68) does not satisfy the regularity condition (63). In other words we have excluded the center of symmetry $R=0$, and R varies in the interval $[R_b e^{-\omega r_0}, R_b]$.

(b) The configuration described by (69)–(71) represents spheres sustained only by tangential stresses. Solutions of this kind have been considered by Lemaitre.[15]

Finally we shall consider an example consistent with the third solution of the surface equation (unlike the two examples above), namely

$$h_1(r) = \cosh[\alpha(r_0 - r)]\,, \tag{73}$$

where α is a positive constant, obviously for this case

$$h_1'^2(r_0) = 0\,.$$

The expressions for the matter variables and the line element read

$$8\pi P_r = 3\alpha^2 \sinh^2[\alpha(r_0 - r)]\,, \tag{74}$$

$$8\pi P_\perp = \alpha^2\{\cosh^2[\alpha(r_0 - r)] - 3\} + \tfrac{1}{2}\ddot{f}(t)/R^2\,, \tag{75}$$

$$8\pi\rho = \alpha^2\{3 - \cosh^2[\alpha(r_0 - r)]\} + 2M/R_b R^2\,, \tag{76}$$

$$ds^2 = \{e^{-f(t)}/\cosh^2[\alpha(r_0 - r)]\}[dt^2 - e^{f(t)}dr^2 - d\Omega^2]\,. \tag{77}$$

Observe that in this example again the regularity condition is not satisfied (so that we exclude the center of symmetry $R=0$) and R changes in the interval $[R_b/\cosh \alpha r_0, R_b]$. The positiveness of the energy density is assured by the condition $\cosh \alpha r_0 \leqslant \sqrt{3}$.

Now, for each configuration (64)–(67) and (69)–(72) we have two different models depending upon the choice of the function $f(t)$ from the different possible solutions of the surface equation, namely

$$e^{f_1(t)} = (3/r_0^2)[3\coth^2[(t - t_0)/2] - 1]^2 \tag{78}$$

and

$$e^{f_2(t)} = (3/r_0^2)\,[3\tanh^2[(t - t_0)/2] - 1]^2\,. \tag{79}$$

It can be seen at once from (66) and (71) that the energy density is an increasing (decreasing) function of time for the contracting (expanding) models. The static limit for the configuration (64)–(67) (for both the expanding and the contracting models) is easily found to be

$$8\pi P_r = (1/9M^2)(r_0/r)^2[(r_0/r)^2 - 1]\,, \tag{80}$$

$$8\pi P_\perp = 0\,, \tag{81}$$

$$8\pi\rho = (1/9M^2)(r_0/r)^4\,, \tag{82}$$

$$ds^2 = (C^4 r_0^2 r^4/12)[dt^2 - (12/r_0^2)\,dr^2 - d\Omega^2]\,. \tag{83}$$

In Schwarzschild-like coordinates

$$R = C^2 r_0 r^2/2\sqrt{3}\,, \quad T = 3M\sqrt{3}\,t\,, \tag{84}$$

the line element (83) becomes

$$ds^2 = (R^2/27M^2)\,dT^2 - (R/M)\,dR^2 - R^2\,d\Omega^2\,, \tag{85}$$

where we have taken into account that

$$C^2 = 6M\sqrt{3}/r_0^3\,,$$

which follows from the fact that for both models (expanding and contracting)

$$h_1'^2(r_0) = 1/27M^2\,.$$

Now, for the configuration (69)–(72) the static limit gives

$$8\pi P_r = 0\,, \tag{86}$$

$$8\pi P_\perp = 4D^2/r_0^2\,R^2\,, \tag{87}$$

$$8\pi\rho = 2/3R^2\,, \tag{88}$$

$$ds^2 = (r_0^2 e^{2\omega r}/12D^2)[dt^2 - (12/r_0^2)\,dr^2 - d\Omega^2]\,. \tag{89}$$

Transforming to Schwarzschild-like coordinates

$$R = r_0 e^{\omega r}/2\sqrt{3}\,D\,, \quad T = 3\sqrt{3}\,Mt\,, \tag{90}$$

the line element (89) reads

$$ds^2 = (R^2/27M^2)\,dT^2 - 3\,dR^2 - R^2\,d\Omega^2\,. \tag{91}$$

Note that solutions given by (80)–(85) and (86)–(91) represent the two extreme cases of anisotropic configurations. Namely, $P_r = 0$, $P_\perp \neq 0$, and $P_r \neq 0$, $P_\perp = 0$.

Finally for the configuration (74)–(77) there is only one model, given by the third (oscillating) solution of the surface equation

$$e^{f_3(t)} = 1/4M^2\cos^4[(t - t_0)/2]\,.$$

V. CONCLUSIONS

We have seen so far that the inclusion of a one-parameter group of conformal motions, together with the spherical symmetry (and the additional restriction $\xi^\alpha U_\alpha = 0$) open the possibility to construct analytical models which may be of some interest in the study of the evolution of compact objects.

It is worth stressing the role played by the anisotropy in the matching of the solutions with the Schwarzschild exterior metric on the boundary of the matter.

We have explicitly displayed three families of solutions.

(a) Expanding solutions representing spheres growing out of a singularity, crossing the horizon and tending asymptotically to the static regime described by Eqs. (80–(85). We have here an example of a white hole, whose final configuration is characterized by the ratio $M/R_b = \frac{1}{3}$.

(b) Contracting solutions, which represent spheres shrinking from an initial highly diffuse state, toward static spheres of radius $R_b = 3M$ [Eqs. (80)–(85)]. These solutions describe the collapse of spheres whose final state is close to a black hole, but are still outside the horizon.

(c) Oscillating solutions, representing spheres whose boundary oscillates between the center and the horizon.

We would like to conclude with the following remarks: (a) the vector field ξ^α defines a motion whenever $h_1'(r) = 0$ (for all r), and (b) we recall that the models presented here correspond to the choice $h_2(t) = 0$.

ACKNOWLEDGMENTS

We would like to thank Professor A. H. Taub for helpful comments and suggestions.

APPENDIX: THE PERFECT FLUID CASE

If we demand the fluid to be locally isotropic, then we obtain from (25) and (26)

$$e^{\lambda/2}[\dot{h}_2(t)\dot{f}(t)e^{f(t)} - 2h_1''(r)] = \Phi(t)\,, \tag{A1}$$

with

$$4\Phi(t) = (2\ddot{f}(t) - \dot{f}^2(t) - 4)e^{f(t)}. \qquad \text{(A2)}$$

Taking derivatives of (A1) with respect to r, we get

$$(\lambda'/2)\Phi(t) - 2h_1'''(r)e^{\lambda/2} = 0 \qquad \text{(A3)}$$

and from (23)

$$-(\lambda'/2)e^{-\lambda/2} = h_1'(r). \qquad \text{(A4)}$$

Using (A3) and (A4) we are led to

$$-2h_1'''(r)/h_1'(r) = \Phi(t), \qquad \text{(A5)}$$

which implies at once that $\Phi(t) = C_1 =$ const. We can now integrate (A5) to obtain

$$2h_1''(r) + C_1 h_1(r) = 2C_2, \qquad \text{(A6)}$$

where C_2 is a constant of integration. Next we can use (A6) to rewrite (A1) in the form

$$\dot{h}_2(t)\dot{f}(t)e^{f(t)} - C_1 h_2(t) = 2C_2. \qquad \text{(A7)}$$

In addition, Eq. (A2) may be written as

$$[2\ddot{f}(t) - \dot{f}^2(t) - 4]\, e^{f(t)} = 4C_1. \qquad \text{(A8)}$$

Thus, the local isotropy leads to the system (A6)–(A8) [Eqs. (44)–(46)], which as stressed before is incompatible with the junction conditions.

[1] L. Herrera and J. Ponce de Léon, J. Math. Phys. **26**, 778 (1985).
[2] V. Canuto, "Neutron Stars: General Review," presented at the Solvay Conference on Astrophysics and Gravitation, Brussels, 1973.
[3] M. Ruderman, Annu. Rev. Astron. Astrophys. **10**, 427 (1972).
[4] S. J. Putterman, *Superfluid Hydrodynamics* (North-Holland, Amsterdam, 1974).
[5] R. Bowers and E. Liang, Astrophys. J. **188**, 657 (1974).
[6] L. Herrera, G. J. Ruggeri, and L. Witten, Astrophys. J. **234**, 1094 (1979).
[7] M. Cosenza, L. Herrera, M. Esculpi, and L. Witten, J. Math. Phys. **22**, 118 (1981).
[8] M. Cosenza, L. Herrera, M. Esculpi, and L. Witten, Phys. Rev. D **25**, 2527 (1982).
[9] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* (Addison–Wesley, Reading, MA, 1951), p. 311.
[10] C. W. Misner and D. M. Sharp, Phys. Rev. B **136**, 571 (1964).
[11] E. H. Robson, Ann. Inst. H. Poincaré **16**, 41 (1972).
[12] M. E. Cahill and A. H. Taub, Commun. Math. Phys. **21**, 1 (1971).
[13] M. E. Cahill and G. C. McVittie, J. Math. Phys. **11**, 1392 (1970).
[14] W. B. Bonnor and P. A. Vickers, Gen. Relativ. Gravit. **13**, 29 (1981).
[15] G. Lemaitre, Ann. Soc. Sci. Bruxelles Ser. 1 A **53**, 97 (1933).

Invariant decomposition of the retarded electromagnetic field

J. Graells, C. Martín, and J. M. Codina
Departament d'Electricitat i Electrònica, Facultat de Física, Universitat de Barcelona, Catalunya, Spain

(Received 17 December 1980; accepted for publication 5 April 1985)

The integral representation of the electromagnetic two-form, defined on Minkowski space-time, is studied from a new point of view. The aim of the paper is to obtain an invariant criteria in order to define the radiative field. This criteria generalizes the well-known structureless charge case. We begin with the curvature two-form, because its field equations incorporate the motion of the sources. The gauge theory methods (connection one-forms) are not suited because their field equations do not incorporate the motion of the sources. We obtain an integral solution of the Maxwell equations in the case of a flow of charges in irrotational motion. This solution induces us to propose a new method of solving the problem of the nature of the retarded radiative field. This method is based on a projection tensor operator which, being local, is suited to being implemented on general relativity. We propose the field equations for the pair {electromagnetic field, projection tensor}. These field equations are an algebraic differential first-order system of one-forms, which verifies automatically the integrability conditions.

I. INTEGRAL REPRESENTATION OF THE ELECTROMAGNETIC FIELD

It is well known that the vacuum Maxwell equations for the potential one-form A_μ are

$$A_{\mu,\alpha}{}^{\alpha} - A_{\alpha,}{}^{\alpha}{}_{\mu} = -4\pi j_\mu, \tag{1.1}$$

from which we can derive the wave equation for the electromagnetic two-form field $F_{\mu\nu} = A_{\nu,\mu} - A_{\mu,\nu}$,

$$F_{\mu\nu,\alpha}{}^{\alpha} = -4\pi(j_{\nu,\mu} - j_{\mu,\nu}). \tag{1.2}$$

We shall make use of this equation because it incorporates the motion of the sources. We must point out that although the law of charge conservation has been lost in (1.2), all our results are consistent with it.

Analyzing the motion of the sources by Fourier transformation, the integral representation of $F_{\mu\nu}$ is obtained in momentum space as

$$F_{\mu\nu}(x) = -\frac{4\pi i}{(2\pi)^2}\int d^4k\, e^{-ik^\sigma x_\sigma} \times [k_\mu j_\nu(k^\sigma) - k_\nu j_\mu(k^\sigma)]/k^\sigma k_\sigma. \tag{1.3}$$

This approach is in the spirit of Huygens and Fresnel. Like them, we consider the field as generated by a distribution of localized sources. The calculus of this integral leads to results that can be found in any advanced textbook,[1] but transforming back $j_\nu(k^\sigma)$ from momentum space to space-time, Eq. (1.3) becomes

$$F_{\mu\nu}(x) = -\frac{4\pi i}{(2\pi)^4}\int d^4\xi \int d^4k\, e^{-ik^\sigma(x-\xi)_\sigma} \times [k_\mu j_\nu(\xi) - k_\nu j_\mu(\xi)]/k^\sigma k_\sigma, \tag{1.4}$$

which is the retarded field, when suited boundary conditions are imposed. It may seem at a glance that this hybrid expression is not mathematically attractive, but if we look at it from a physical point of view, it can be interpreted as follows: the field generation process at the event ξ, and its propagation to the field event x, is analyzed in momentum space by means of the density

$$d^4\xi\, e^{-ik^\sigma(x-\xi)_\sigma}\{[k_\mu j_\nu(\xi) - k_\nu j_\mu(\xi)]/k^\sigma k_\sigma\}.$$

The integration over k^σ then gives the contribution to the total field generated at ξ^α, and finally the integration over the domain of definition of the current $j_\mu(\xi^\alpha)$ yields to $F_{\mu\nu}(x)$. The possibility of such an interpretation is apparently closely related to the linearity of the theory. This simple interpretation encourages us to go with our purpose to study geometrically the retarded radiative field and to generalize the expression of the electromagnetic field created by a structureless point charge in a given arbitrary motion.

Our starting point will be the integral on momentum space. With respect to a global inertial frame the volume element splits into $d^4k = d^3\mathbf{k}\, dk^0$. Now, we can analytically continue k^0 to the complex plane, and applying the residue theorem, by choosing the retarded prescription, it is obtained:

$$F_{\mu\nu}(x) = \frac{1}{(2\pi)^2}\int d^4\xi\, \theta(x^0 - \xi^0)\int \frac{d^3\mathbf{k}}{|\mathbf{k}|}\{e^{-i\overset{k^\sigma}{+}(x-\xi)_\sigma} \times (\underset{+}{k}\wedge j(\xi))_{\mu\nu} + e^{-i\overset{k^\sigma}{-}(x-\xi)_\sigma}(\underset{-}{k}\wedge j(\xi))_{\mu\nu}\}, \tag{1.5}$$

where we have defined the following.

(1) $\overset{k^\sigma}{+} \equiv (|\mathbf{k}|, \mathbf{k})$, $\overset{k^\sigma}{-} \equiv (-|\mathbf{k}|, \mathbf{k})$ are the lightlike vectors.

(2) $(k\wedge j(\xi))_{\mu\nu} \equiv k_\mu j_\nu(\xi) - k_\nu j_\mu(\xi)$ is the exterior product.

(3) $\theta(x^0 - \xi^0)$ is the Heaviside or step function. It is understood that the Minkowski metric has signature $+2$, i.e.,

$$k^\sigma k_\sigma = \eta_{\alpha\beta}k^\alpha k^\beta = -(k^0)^2 + \mathbf{k}^2.$$

Until now, this approach is equivalent to the propagator (Green's function) approach.

Taking into account that $dw = d^3\mathbf{k}/|\mathbf{k}|$ is the measure of the upper light cone on momentum space, Eq. (1.5) can be written geometrically

$$F(x) = \frac{1}{(2\pi)^2}\int d^4\xi\, \theta(x^0 - \xi^0)\int_{k^\sigma k_\sigma = 0} dw\, \epsilon(k^0) \times e^{ik^\sigma(x-\xi)_\sigma}\, k\wedge j(\xi), \tag{1.6}$$

where $\epsilon(k^0)$ is the signum function.

 0022-2488/85/082024-06$02.50

It is clear in Eq. (1.6) that the integral over the momentum light cone is the exterior differential of the invariant Jordan–Pauli distribution,[2] therefore we can reduce the integral to the upper light cone $k^\sigma k_\sigma \theta(k^0)=0$, by virtue of $\theta(x^0-\xi^0)$. In order to calculate it, we shall make use of the following orthonormal tetrad field: $\{e_i\}=\{e_0,e_1,e_2,e_3\}$, which is defined at the generic source event ξ according to the following.

(a) $e_0 \equiv v$ is the current four-vector velocity, i.e., if ρ_0 is the proper charge density, then $j^\mu=\rho_0 v^\mu, v^\mu v_\mu=-1$.

(b) e_1 is a unit spacelike vector $e_1^\alpha e_{1\alpha}=1$, orthogonal to e_0 and included in the plane determined by e_0 and the separation vector $R^\alpha=(x-\xi)^\alpha$.

(c) e_2 and e_3 are two unit orthogonal spacelike vectors $e_2^\alpha e_{2\alpha}=e_3^\alpha e_{3\alpha}=1, e_2^\alpha e_{3\alpha}=0$, the plane they determine being perpendicular to that defined by $\{e_0,e_1\}$.

The dual basis $\{\omega^i\}$ to $\{e_i\}$ is canonically defined by

$$\langle \omega^i, e_j\rangle=\delta^i_j,$$

where δ^i_j is the Kronecker delta. Care must be taken because the velocity v, considered as a one-form, is related to ω^0 with a minus sign: $\omega^0_\alpha=-v_\alpha$.

The separation vector $R=(x-\xi)$ between the field event x and the source one ξ is expressed with respect to the tetrad $\{e_i\}$ by

$$R \equiv x-\xi = re_0+r'e_1, \tag{1.7}$$

where $r=-e_0\cdot R$ and $r'=e_1\cdot R$ are the Lorentz-invariant tetrad components of R.

In order to compute the integral over the upper light cone

$$I(\xi)\equiv\int_{k^\sigma k_\sigma \theta(k^0)=0} dw\, e^{-ik^\sigma (x-\xi)_\sigma}\, k\wedge j(\xi), \tag{1.8}$$

we make use of standard calculations. For example, by choosing spherical coordinates at the simultaneity $\xi^0=c^{\Gamma_e}$, with polar axis e_1, it is easily obtained

$$I(\xi)=4\pi i\int_0^\infty d\,|\mathbf{k}|\, e^{i|\mathbf{k}|r}\times\left(|\mathbf{k}|\,\frac{\cos|\mathbf{k}|r'}{r'}-\frac{\sin|\mathbf{k}|r'}{r'}\right)e_1\wedge j(\xi).$$

Now, according to the Laplace–Carlson transform,[3]

$$p\int_0^\infty dt\, e^{-p^t}\sin at=\frac{ap}{p^2+a^2},$$

$$p\int_0^\infty dt\, e^{-p^t}\,t^{\nu-1}\cos at=\frac{\Gamma(\nu)}{2}p\left(\frac{1}{(p-ia)^\nu}+\frac{1}{(p+ia)^\nu}\right).$$

Insertion of $p=-ir$, $a=r'$, $\nu=2$, $\Gamma(2)=1$, $t=|\mathbf{k}|$ into the above integrals yields

$$I(\xi)=-4\pi i\,\frac{1}{r'}\left(\frac{r^2+r'^2}{(-r^2+r'^2)^2}+\frac{1}{-r^2+r'^2}\right)\times e_1(\xi)\wedge j(\xi). \tag{1.9}$$

Inserting the expression (1.9) into the electromagnetic field (1.6), we get the equation

$$F(x)=-\frac{4\pi i}{(2\pi)^2}\int d^4\xi\,\theta(x^0-\xi^0)\frac{1}{r'}\times\left(\frac{r^2+r'^2}{(-r^2+r'^2)^2}+\frac{1}{(-r^2+r'^2)}\right)e_1(\xi)\wedge j(\xi). \tag{1.10}$$

II. LOCALLY PROPER TIME SYNCHRONIZABLE FLOW OF CHARGES

At a glance it may seem that in order to derive consequences of expression (1.10), it is necessary to know $j(\xi)$, but on second thought if we dare try some general properties of the flow of charges that constitute the current density, an interesting expression of the electromagnetic two-form can be derived.

We first assume for a current density $j=\rho_0 v$, that its velocity one-form field v verifies Frobenius condition[4]

$$v\wedge dv=0. \tag{2.1}$$

From a physical point of view it means that the observers associated to the v timelike flow make a locally synchronizable frame. This condition result is too general, therefore, we furthermore assume the more restrictive one

$$dv=0, \tag{2.2}$$

which means that the flow is locally proper time synchronizable, i.e., the observers reference frame can experimentally correlate by "radar" their proper times. In other words, there exists a family of three-spaces to which the streamlines are orthogonal. This motion is an irrotational one, i.e., in decomposing that portion of the covariant derivative $v_{\alpha;\beta}$ which is perpendicular to the velocity, into its antisymmetric part, its symmetric trace-free part, and the trace itself[5]

$$v_{\alpha;\beta}=-v_{\alpha;\gamma}v^\gamma v_\beta+\omega_{\alpha\beta}+\sigma_{\alpha\beta}+\tfrac{1}{3}\vartheta h_{\alpha\beta}.$$

The rotation reduces to $\omega=a\wedge v$. Now, according to the global version of the Frobenius theorem, due to Chevalley and Ehresman, applied to the present case, it implies that Minkowski space-time is foliated by the integral manifolds (three-spaces above) of the vector distribution defined by the velocity field $v{:}x\rightarrow M_x$. This foliation can be labeled by the proper time τ coordinate ($\nu=d\tau$) and three spacelike coordinates η^i suited to the three-spaces.

Therefore, we refer Minkowski space-time to the inertial coordinate system (ξ^0,ξ^i) and to that induced by the foliation (τ,η^i)

$$\begin{cases}\xi^0=\xi^0(\tau,\eta^i),\\ \xi^i=\xi^i(\tau,\eta^i),\end{cases}\quad \begin{cases}\tau=\tau(\xi^0,\xi^i),\\ \eta_i=\eta^i(\xi^0,\xi^i).\end{cases} \tag{2.3}$$

We can bring the Minkowski line element with respect to the coordinates (τ,η^i) to the form

$$ds^2=-d\tau^2+g_{ij}(\tau,\eta^i)d\eta^i\,d\eta^j=-(d\xi^0)^2+(d\,\xi)^2, \tag{2.4}$$

because v^α is not a null vector. The volume element will be expessed as

$$d^4\xi=\sqrt{|g|}\,d\tau\,d^3\eta=\sqrt{g}\,d\tau\,d^3\eta. \tag{2.5}$$

With these premises, Eq. (1.10) can be split into

$$F(x)=\frac{1}{\pi i}\int d^3\eta\int d\tau\sqrt{g}\,\theta(x^0-\xi^0)\times\frac{1}{e_1\cdot R}\left(\frac{(v\cdot R)^2+(e_1\cdot R)^2}{(R_\alpha R^\alpha)^2}+\frac{1}{R_\alpha R^\alpha}\right)e_1\wedge j. \tag{2.6}$$

The integration over τ will be done by applying the residue theorem. The Heaviside function selects the retarded pole, but the difference from the previous calculation (integral over k^0) consists in the fact that the denominator $(R^\alpha R_\alpha)^2$ now vanishes to the second order (pole of order 2), so that we have to carry the expansion to the order $(\tau-\tau_r)^2$ (where "r" stands for "retarded")

$$R_\alpha R^\alpha=-(R_\alpha v^\alpha)_{\tau_r}(\tau-\tau_r)-(1+R_\alpha a^\alpha)_{\tau_r}\times(\tau-\tau_r)^2+O[(\tau-\tau_r)^3]. \tag{2.7}$$

Inserting the development (2.7) into the expression (2.6), it reduces to

$$F(x)=\frac{1}{\pi i}\int d^3\eta(-2\pi i)\lim_{\tau\to\tau_r}\left\{(\tau-\tau_r)\frac{\sqrt{g}}{\epsilon_1\cdot R}\frac{e_1\wedge j}{-2R_\alpha v^\alpha(\tau-\tau_r)+O[(\tau-\tau_r)^2]}+\frac{\partial}{\partial\tau}\left((\tau-\tau_r)^2\cdot\frac{\sqrt{g}}{e_1\cdot R}\frac{(v\cdot R)^2+(e_1\cdot R)^2}{-2R_\alpha v^\alpha(\tau-\tau_r)-(1+R_\alpha a^\alpha)(\tau-\tau_r)^2+O[(\tau-\tau_r)^2]}e_1\wedge j\right)\right\}. \tag{2.8}$$

This method of calculus follows the Sommerfeld approach to the residue theorem.[6] We now introduce the lightlike vector $L=e_1+v$, which verifies the following algebraic relations:

$$L\cdot L=0,\quad L\cdot e_1=1,\quad L\cdot v=-1.$$

Differentiating these relations along the flow worldline, i.e., with respect to τ, and making simple algebraic operations, we obtain

$$\frac{\partial e_1^\alpha}{\partial\tau}\equiv(\nabla_v e_1)^\alpha=-a^\alpha+(a\cdot e_1)L^\alpha. \tag{2.9}$$

Taking into account the law of charge conservation $\partial_\alpha j^\alpha=0$, we can easily derive

$$\frac{\partial\rho_0}{\partial\tau}=\nabla_v\rho_0=-\rho_0\nabla\cdot v, \tag{2.10}$$

and, finally, applying the equality

$$\frac{1}{g}\frac{\partial g}{\partial\tau}=\nabla\cdot v, \tag{2.11}$$

it allows us to express Eq. (2.8) as we desire

$$F(x)=-\int d^3\eta\left(\sqrt{g}\rho_0\frac{e_1\wedge L}{(-v\cdot R)^2}\right)_{\tau_r}+\frac{1}{2}\int d^3\eta\left(\sqrt{g}\nabla\cdot v\rho_0\frac{e_1\wedge L}{(-v\cdot R)}\right)_{\tau_r}+\int d^3\eta\left(\sqrt{g}\rho_0\frac{a+(a\cdot e_1)v}{(-v\cdot R)}\wedge L\right)_{\tau_r}. \tag{2.12}$$

Until now we have not dealt with the question of the integral domain of the η^i coordinates. The calculus of the residue has implied that this domain is defined by the equations

$$\begin{cases}\tau_r=\tau(x^0-|\mathbf{x}-\boldsymbol{\xi}|,\xi^j),\\ \eta^i=\eta^i(x^0-|\mathbf{x}-\boldsymbol{\xi}|,\xi^j),\end{cases}\Leftrightarrow\begin{cases}\xi^0=x^0-|\mathbf{x}-\boldsymbol{\xi}|,\\ \xi^j=\xi^j.\end{cases} \tag{2.13}$$

Therefore it is clearly seen that the integration domain is the retarded light cone of the field event: $I_r(x)$.

Now, following Newman and Penrose,[7] a complex null tetrad $\{z\}=\{l,n,m,\bar m\}$ is introduced for expressing Eq. (2.12), because of its adequacy and internal adaptability for studying the solution and the equations of massless fields possessing certain algebraic properties.

The Newman–Penrose tetrad is defined by the standard prescription

$$\begin{aligned}&l=(1/\sqrt{2})(v+e_1)\equiv L/\sqrt{2},\quad v=(1/\sqrt{2})(l+n),\\&n=(1/\sqrt{2})(v-e_1),\quad e_1=(1/\sqrt{2})(l-n),\\&m=(1/\sqrt{2})(e_2+ie_3),\quad e_2=(1/i\sqrt{2})(m-\bar m),\\&\bar m=(1/\sqrt{2})(e_2-ie_3),\quad e_3=(1/\sqrt{2})(m+\bar m).\end{aligned} \tag{2.14}$$

With respect to this tetrad the acceleration a is expressed as

$$a=a^1e_1+a^2e_2+a^3e_3=(1/\sqrt{2})(a^1l-a^1n+\bar Am+A\bar m), \tag{2.15}$$

where we have defined $A=(a^3+ia^2)$; $\bar A$ is the complex conjugate of A.

Therefore, Eq. (2.12) will reduce to

$$F(x)=\int_{I_r(x)}d^3\eta\sqrt{g}\rho_0\frac{n\wedge l}{r^2}+\frac{1}{2}\int_{I_r(x)}d^3\eta\sqrt{g}\rho_0\nabla\cdot v\frac{n\wedge l}{r}+\int_{I_r(x)}d^3\eta\sqrt{g}\rho_0\frac{(\bar Am+A\bar m)}{r}\wedge l. \tag{2.16}$$

At this stage we must point out that the Newman–Penrose (NP) tetrad is associated with the flow of charges, but in the one-point structureless charge case is directly associated to F. In fact, for this particular and well-studied problem, the integration is made trivially because it reduces to the intersection of the q charge worldline with the retarded light cone $I_r(x)$, and taking into account that $\nabla\cdot v=0$, Eq. (2.16) is in this case

$$F_q(x)=q(n\wedge l/r^2)_{\tau_r}+q((\bar Am+A\bar m)\wedge l/r)_{\tau_r}. \tag{2.17}$$

We remember here that every magnitude refers to the retarded event $z^\alpha(\xi^0_r)=z^\alpha(x^0-|\mathbf{x}-\boldsymbol{\xi}|)$ of the x field event, $z^\alpha(\xi^0)$ being the trajectory of the q-point charge. Now the Newman–Penrose tetrad is associated with the light congruence of F, whose tangent null vector field is l.

The matrix representation of Eq. (2.17) with respect to the n.p. tetrad is

$$(F)=\frac{q}{r^2}\left(\begin{array}{cc|cc} 0 & -1 & & \\ -1 & 0 & \multicolumn{2}{c}{0} \\ \hline & & & \\ \multicolumn{2}{c|}{0} & \multicolumn{2}{c}{0} \end{array}\right) +\frac{q}{r}\left(\begin{array}{cc|cc} \multicolumn{2}{c|}{0} & -\bar{A} & A \\ & & 0 & 0 \\ \hline \bar{A} & 0 & & \\ -A & 0 & \multicolumn{2}{c}{0} \end{array}\right). \tag{2.18}$$

As was shown by Newman and Penrose, this approach is equivalent to the spinor formalism. The consequence for the one-point charge have been widely studied in the literature. For this reason, we now turn back to the much more general and not-studied case, represented by Eq. (2.16). We must remember that the integrals extend to the lower light cone of the field event, parametrized by a generic spacelike slice of the foliation induced by v. The splitting of the two-form $F=F_{0i}\,dt\wedge dx^i+F_{ij}\,dx^i\wedge dx^j$, which lets us identify $(F_{0i})=\mathbf{E}$ as the electric field and $(F_{ij}\epsilon^{ijk})=\mathbf{B}$ as the magnetic field, only has meaning with respect to an *inertial frame*! Therefore, we must refer Eq. (2.16) to an inertial frame. Then, writing down the vector components by Greek letters, we obtain, returning back to source kinematical variables

$$\begin{aligned} F_{\alpha\beta}(x)=&\int_{I_r(x)} dw\,\rho_0\frac{v_\alpha e_{1\beta}-v_\beta e_{1\alpha}}{r}+\frac{1}{2}\int_{I_r(x)} dw\,\rho_0 \\ &\times\nabla\cdot v(v_\alpha e_{1\beta}-v_\beta e_{1\alpha})+\int_{I_r(x)} dw\,\rho_0[(a_\alpha v_\beta-a_\beta v_\alpha) \\ &-(e_{1\alpha}(a^1v_\beta+a_\alpha)-e_{1\beta}(a^1 v_\alpha+a_2))], \end{aligned} \tag{2.19}$$

where we have taken into account that $dw=\sqrt{g}\,(d^3\eta/-v\cdot R)$ is the measure or absolute Lorentz-invariant two-content of the retarded null cone of the field event.

The meaning of the three-integral splitting of Eq. (2.19) is as follows:

generalized coulomb field,

$$C_{\alpha\beta}\equiv\int_{I_r(x)} dw\,\rho^0\,\frac{v_\alpha e_{1\beta}-v_\beta e_{1\alpha}}{r};$$

intermediate-longitudinal field,

$$I_{\alpha\beta}\equiv\int_{I_r(x)} d\omega\,\frac{1}{2}\rho_0\nabla\cdot v(v_\alpha e_{1\beta}-v_\beta e_{1\alpha}); \tag{2.20}$$

radiation field,

$$\begin{aligned} R_{\alpha\beta}\equiv\int_{I_r(x)} d\omega\,\rho_0\,&[(a_\alpha v_\beta-a_\beta v_\alpha)-(e_{1\alpha}(a^1v_\beta+a_\beta) \\ &-e_{1\beta}(a^1v_\alpha+a_\alpha))]. \end{aligned}$$

The justification of $C_{\alpha\beta},I_{\alpha\beta},R_{\alpha\beta}$ becomes apparent when one goes to the rest system of the integrand two-forms. In this system, it is verified at the fixed source event we work on: $v_\alpha=(-1,0),e_1=(0,\mathbf{r}/|\mathbf{r}|)$; $\mathbf{r}$ being the three-vector which points from the retarded position of the flow of charges to the field point. The Coulomb and radiation fields are direct generalizations of the one-point structureless charge.

The intermediate-longitudinal field is a new one, which has a structure, being longitudinal, similar to the Coulomb field, but it depends on r, like a radiation field. Then the Lorentz-invariant separation between velocity and acceleration-radiation fields is only possible for the one-point charge case.

III. A NEW APPROACH: PROJECTION OPERATOR

The electromagnetic radiation field appears in a number of different forms: acceleration-retarded fields (kinematical source criteria); free fields (dynamical criteria), e.g., $F=F_{\text{retarded}}-F_{\text{advanced}}$; null fields (algebraic criteria); asymptotic developments (Goldberg–Keer theorem, B.M.S. group); etc. None of these methods scarcely may be completely implemented in general relativity, and they are not always equivalent among themselves. Then we consider it quite reasonable to ask ourselves: What do we mean by "radiative field"? Perhaps this question is a methodological one, but even in this case, we must remember what Ginzburg[8] said: "In Physics there are many 'perpetual problems' the discussion of which continues for decades... . On the other hand, however, neglect of such methodological types of problems sometimes incurs vengeance!" Our approach to dealing with the problem, i.e., to defining what the radiative field is for us, is a dynamical geometrical one, which takes into account that when the electromagnetic field produced by a given source is measured, one always finds the retarded field, and it shares the maximum number of common features of all the above criteria.

For the moment we shall not invoke "Occam's razor," and shall assume the existence of a projection tensor operator $P^{\alpha\beta}{}_{\gamma\delta}$, such that when applied to the total electromagnetic field gives us the desired radiative part

$$R_{\alpha\beta}\equiv P_{\alpha\beta}{}^{\gamma\delta}F_{\gamma\delta}. \tag{3.1}$$

It is easily found that $P_{\alpha\beta\gamma\delta}$ is antisymmetric in each of the index pairs $\alpha\beta$ and $\gamma\delta$, but is symmetric under interchange of the pairs.

It is clear that the problem now has been shifted to find $P_{\alpha\beta\gamma\delta}$. What are the advantages to introducing this new entity? First of all, if we can find a reasonable system of first-order partial differential equations for it, we shall have a criteria based on dynamical-geometric equations that when applying the strong equivalence principle will be valid in general relativity. Second, it is therefore a local field criteria. Third, it implies a new concept of the electromagnetic field, which we view now as the pair $\{P,F\}$, the contraction $\langle P,F\rangle$ being the radiative part.

The algebraic-partial differential system that we propose is as follows:

$$*F_{[\alpha\beta,\gamma]}-4\pi*j_{[\alpha\beta\gamma]}=0, \tag{3.2a}$$

$$F_{[\alpha\beta,\gamma]}=0, \tag{3.2b}$$

$$P_{\alpha\beta\mu\nu}-P_{\alpha\beta\gamma\delta}P^{\gamma\delta}{}_{\mu\nu}=0, \tag{3.2c}$$

$$P_{\alpha\beta\mu\nu}F^{\alpha\beta}F^{\mu\nu}=0, \tag{3.2d}$$

$$\epsilon_{\alpha\beta\gamma\delta}P^{\alpha\beta\mu\nu}P^{\gamma\delta\tau\sigma}F_{\mu\nu}F_{\tau\sigma}=0, \tag{3.2e}$$

$$P_{[\alpha\beta}{}^{\gamma\delta}F_{\gamma\delta,\nu]}+P^{\gamma\delta}{}_{[\alpha\beta,\nu]}F_{\gamma\delta}=0, \tag{3.2f}$$

and boundary conditions in the Sommerfeld sense. We have to mention that, because there is some confusion in the literature. Sommerfeld said quite explicitly that his aim is to exclude incoming radiation. Retarded solutions for spatially bounded sources satisfy his conditions automatically in future null directions. ($\mathscr{I}^+$ is the "future null infinity," the region $t + |\mathbf{r}| \to +\infty$ at finite $t - |\mathbf{r}|$.)

In system (3.2), the first two equations are Maxwell equations. The third equation simply states that P is a projection tensor. The fourth and fifth equations mean that the radiative field has to be degenerate or algebraically special (they are equivalent to $F_{\alpha\beta}F^{\alpha\beta}=0, *F_{\alpha\beta}F^{\alpha\beta}=0$). The sixth equation states that the radiative field is closed. As far as the boundary conditions are concerned, we adopt the Sommerfeld point of view, and explicitly we exclude the homogeneous-free solutions of the system. Then as Fock has remarked, care must be taken at past infinity, because as one recedes along null straight lines coming in from the past, the retarded field reflects source behavior at even earlier times; a condition on the time dependence of the sources in the past infinity is required in order that the retarded field satisfies Sommerfeld's condition at past null infinity.[9]

We have studied the system (3.2) from the point of view of the differential ideal of Frobenius–Cartan theory, i.e., as an exterior differential system. After a lengthy but easy calculation we have proved that the system verifies the required integrability conditions, and consequently has solutions in any given Riemannian spacetime, i.e., in the presence of gravitational fields. (In the Appendix to the present article we sketch the proof.)

For the one-point structureless charge problem, we have found the solution. The projection tensor referring to the dual bases of the orthonormal tetrad defined in Sec. I is

$$P = -(\omega^0\wedge\omega^2)(\omega^0\wedge\omega^2) + (\omega^1\wedge\omega^2)(\omega^1\wedge\omega^2), \qquad (3.3)$$

being expressed with respect to the global inertial frame as follows:

$$P_{\alpha\beta\mu\nu} = [1/(a^2)^2]\,\delta^{\alpha'\beta'}_{\alpha\beta}\,\delta^{\mu'\nu'}_{\mu\nu}\{-(v_{\alpha'}a_{\beta'} - a^1 v_{\alpha'}e_{1\beta'}) \times(v_{\mu'}a_{\nu'} - a^1 v_{\mu'}e_{1\alpha'}) + (e_{1\alpha'}a_{\beta'}e_{1\mu'}a_{\nu'})\}, \qquad (3.4)$$

where $\delta^{\alpha'\beta'}_{\alpha\beta} = \delta^{\alpha'}_{\alpha}\delta^{\beta'}_{\beta} - \delta^{\alpha'}_{\beta}\delta^{\beta'}_{\alpha}$, and the orthonormal basis has been chosen so that $a = a^1e_1 + a^2e_2$. Functionally P must be considered as depending on the field event x through the retarded prescription, i.e., $P(x) = P(\tau_r(x))$.

The physical and mathematical structure of the projection tensor which corresponds to the electromagnetic field derived in Sec. II of this article can be easily determined if we is taken into account its similarity with the one-point charge solution; in fact, the Coulomb and intermediate two-form integrands are always bivector orthogonal to the radiation two-form integrand.

According to the existence theorem for the system (3.2), it shall be possible to find a tetrad field of one-forms $\{W^i(x)\}$ to refer to the cotangent bundle associated to Minkowski space-time. With respect to this tetrad field, Eq. (2.19) may be expressed as

$$F(x) = (CI)(x)\{W^0(x)\wedge W^1(x)\} + R(x)\{(W^0(x) - W^1(x))\wedge W^2(x)\}, \qquad (3.5)$$

where

$$(CI)(x)W^0\wedge W^1 \equiv -\int_{I_r(x)} d\omega\,\rho_0\left[\frac{1}{r} + \frac{1}{2}\nabla\cdot v\right]\omega^0\wedge\omega^1, \qquad (3.6)$$

$$R(x)(W^0 - W^1)\wedge W^2 \equiv \int_{I_r(x)} d\omega\,\rho_0(\omega^0 - \omega^1)\wedge\omega^2.$$

Therefore the projection tensor operator will be

$$P(x) = -(W^0\wedge W^2)(W^0\wedge W^2) + (W^1\wedge W^2)(W^1\wedge W^2). \qquad (3.7)$$

It is interesting to remark that the radiative solution implied by Eqs. (3.6) and (3.7) obeys at least the same properties of usual null fields. In fact, with respect to a global inertial frame $\{x^\alpha\}$, where it is defined $W^0(x) - W^1(x) \equiv l_\alpha\, dx^\alpha$, the radiative field will be expressed as

$$R_{\alpha\beta} = R(x) l_\alpha \wedge W^2_\beta(x). \qquad (3.8)$$

Evidently, here l_α is a null one-form: $l^\alpha l_\alpha = 0$, which belongs to the congruence associated to $R_{\alpha\beta}(x)$. Its impulse-energy tensor reduces to

$$T^{(R)}_{\mu\nu} = (1/4\pi)(R_{\mu\lambda}R^{\lambda}_{\ \nu} - \tfrac{1}{4}\eta_{\mu\nu}R_{\alpha\beta}R^{\alpha\beta}) = \frac{R^2(x)}{4\pi}\, l_\mu(x) l_\nu(x), \qquad (3.9)$$

which in view of the lightlike nature of l_α, satisfies the known relations for null fields

$$l_\nu T^{\nu\mu} = 0, \quad T^{\mu}_{\ \mu} = 0, \quad T^{\mu\nu}T_{\mu\nu} = 0. \qquad (3.10)$$

An inertial observer characterized by its four-velocity u^α will measure a flux of energy or density of four-momentum given by

$$N^\alpha(x) = T^{\alpha\beta}u_\beta = [R^2(x)/4\pi]\, l_\beta u^\beta l^\alpha. \qquad (3.11)$$

As N^α is proportional to l^α, it is also a null vector, whose zero component $N^\alpha u_\alpha$ gives the energy density as measured by this inertial observer.

APPENDIX: EXISTENCE OF SOLUTIONS

The algebraic-partial differential system (3.2) is analyzed from the point of view of the differential ideal of Frobenius–Cartan.[4] Consequently, it will be considered as an exterior differential system, in which $G^{\mu\nu\sigma\tau}$ is a representation of the metric in the exterior algebra

$$\begin{aligned}
&\overset{0}{\omega} \equiv F_{[\alpha\beta\gamma\delta]} = 0,\\
&\overset{0}{\alpha} \equiv *F_{[\alpha\beta\gamma]} - 4\pi *j_{[\alpha\beta\gamma]},\\
&\overset{0}{\omega}_{\alpha\beta\gamma\delta} \equiv P_{\alpha\beta\gamma\delta} - G^{\mu\nu\sigma\tau}P_{\alpha\beta\mu\nu}P_{\sigma\tau\gamma\delta} = 0,\\
&\overset{0}{\beta} \equiv G^{\alpha\beta\gamma\delta}G^{\mu\nu\rho\sigma}\epsilon^{\lambda\pi\tau0}P_{\lambda\pi\alpha\beta}P_{\tau0\mu\nu}F_{\gamma\delta}F_{\rho\sigma} = 0,\\
&\overset{0}{\gamma} \equiv G^{\lambda\pi\alpha\beta}G^{\tau0\mu\nu}P_{\lambda\pi0\tau}F_{\alpha\beta}F_{\mu\nu} = 0, \qquad (\mathrm{A1})\\
&\overset{0}{\omega}_{\lambda\pi\nu} \equiv G^{\tau0\alpha\beta}P_{[\lambda\pi\tau0\nu]}F_{\alpha\beta} + P_{[\lambda\pi\tau0}G^{\tau0\alpha\beta}{}_{,\nu]}F_{\alpha\beta}\\
&\qquad + P_{[\lambda\pi0\tau}G^{0\tau\alpha\beta}F_{\alpha\beta\nu]} = 0,\\
&\overset{1}{\omega}_{\lambda\pi} \equiv dF_{\lambda\pi} - F_{\lambda\pi\nu}\,dx^\nu = 0,\\
&\overset{1}{\omega}_{\lambda\pi\tau0} \equiv dP_{\lambda\pi\tau0} - P_{\lambda\pi\tau0\nu}\,dx^\nu,
\end{aligned}$$

plus the closure, i.e., the exterior differential of the system (A1), that we symbolically write down

$$d\,(\mathrm{A}1). \tag{A2}$$

Equations (A1) and (A2) are defined on a 139-dimensional manifold $N = N(\kappa^{\alpha}{}_{\gamma}, F_{\mu\nu}, P_{\alpha\delta\sigma\tau}, F_{\alpha\beta\gamma}, P_{\gamma\delta\mu\nu\tau})$, because the electromagnetic field source and the geometry of the base space (space-time) are considered to be given.

After a lengthy but easy calculation it is proven that we recover system (3.2) when the complete exterior system (A1) and (A2) is restricted to the original manifold, and its integrability conditions are automatically satisfied!

Therefore, the equations we propose to define the radiative electromagnetic field as the contraction $P_{\alpha\beta\gamma\delta}F^{\gamma\delta}$ have solutions. It remains to prove mathematically its uniqueness.

[1]E. Konopinski, *Electromagnetic Fields and Relativistic Particles* (McGraw–Hill, New York 1981).
[2]F. Rohrlich, *Classical Charged Particles* (Addison–Wesley, Reading, MA, 1965).
[3]V. Ditkine and A. Provdnikov, *Transformations Intégrals et Calcul Opérationnel* (Mir, Moscow, 1978).
[4]Y. Choquet-Bruhat and C. De Witt-Morette, with M. Dillar-Bleick, *Analysis, Manifolds and Physics* (North-Holland, Amsterdam, 1982), revised edition.
[5]G. F. R. Ellis, "Relativistic Cosmology," in *Proceedings of the International School of Physics "Enrico Fermi,"* Course XLVII, edited by B. K. Sachs (North-Holland, Amsterdam, 1971).
[6]A.Sommerfeld, *Electrodynamics* (Academic, New York, 1952); *Partial Differential Equations in Physics* (Academic, New York, 1967).
[7]V. P. Frolov, "The Newman–Penrose method in the theory of General Relativity," in *Proceeding (Trudy) of the P. N. Lebedev Physics Institute*, Vol. 96, edited by N. G. Basov (Consultants Bureau, New York, 1979).
[8]V. L. Ginzburg, Sov. Phys. Usp. **98**, 569 (1969).
[9]M. Walker, in *Isolated Gravitating Systems in General Relativity*, edited by J. Ehlers (North-Holland, Amsterdam, 1979).

Generalized Weyl-type gauge geometry

Peter G. Bergmann
Department of Physics, Syracuse University, Syracuse, New York 13210 and Department of Physics, New York University, 4 Washington Place, New York, New York 10003
Arthur B. Komar
Physics Department, Yeshiva University, New York, New York 10003 and Department of Physics, New York University, 4 Washington Place, New York, New York 10003

(Received 19 October 1984; accepted for publication 5 April 1985)

Weyl-type gauge geometry based on gauging by GL(4, R) and presented in an earlier paper is developed from an alternative point of view, which emphasizes independent geometric objects as the building blocks. This approach avoids most of the elaborate calculations of the earlier paper, and thus contributes to an intuitive understanding. The new building blocks, which are the metric and two different affine connections, uniquely determine the tensorial structures of the earlier paper, and vice versa.

I. INTRODUCTION

In a previous paper[1] one of us has proposed a generalization of Weyl's gauging of the metric field by subjecting the metric at each world point of space-time to a bilinear tensorial transformation

$$g_{\mu\nu}{}' = \Omega_\mu{}^\rho \Omega_\nu{}^\sigma g_{\rho\sigma}, \tag{1}$$

where $\Omega_\mu{}^\rho$ is some element of GL (4, R). The transformation preserves the symmetry of the metric and its signature, and the determinant of the metric will continue to be nonzero. In that paper,[1] a central role was assigned to a structure $A_\alpha{}^\beta{}_\gamma$, which was interpreted as the generalization of Weyl's "compensating" vector potential. In particular, $A_\alpha{}^\beta{}_\gamma$ is required to obey certain integrability conditions if there is to exist a gauge transformation (1) that reduces the geometry to a conventional Riemannian geometry. Here, $A_\alpha{}^\beta{}_\gamma$ possesses 64 algebraically independent components. Its transformation law, Eq. (4.5) of Ref. 1, involves the components of the metric tensor.

The purpose of the present paper is to describe a somewhat different approach, which leads to the identical geometric structure as in Ref. 1, but by its complementary perspective may afford additional intuitive insight. We shall start from a set of two distinctive fiber bundles, each with its own affine connection. It will turn out that, with the metric given, one of these connections has 40 algebraically independent components, the other 24, and that these two connections determine unambiguously the structure $A_\alpha{}^\beta{}_\gamma$. Conversely, given $A_\alpha{}^\beta{}_\gamma$, these two connections can be constructed unambiguously. Both connections are geometric objects, that is to say, given its components in one coordinate system *cum* gauge frame, its components in another coordinate system (with another gauge frame) are unambiguously defined, and without reference to any other structure, such as the metric.

II. THE SYMMETRY GROUP

The complete symmetry group of Ref. 1 consists of the concatenation of coordinate transformations with (generalized) gauge transformations. Thus a gauge-sensitive vector U^μ would transform as follows:

$$U^{\nu'} = \frac{\partial x^{\nu'}}{\partial x^\kappa} \theta_\rho{}^\kappa U^\rho. \tag{2}$$

The matrix $\theta_\rho{}^\kappa$ is the reciprocal of $\Omega_\mu{}^\rho$, thus

$$\Omega_\mu{}^\rho \theta_\rho{}^\kappa = \delta_\mu{}^\kappa. \tag{3}$$

Here the gauge transformation is applied first, followed by the coordinate transformation. This sequence may be reversed,

$$U^{\nu'} = \theta_{\kappa'}{}^{\nu'} \frac{\partial x^{\kappa'}}{\partial x^\rho} U^\rho, \tag{4}$$

and the result will be the same as in Eq. (2) provided the following relation holds:

$$\theta_{\lambda'}{}^{\nu'} = \frac{\partial x^{\nu'}}{\partial x^\kappa} \frac{\partial x^\rho}{\partial x^{\lambda'}} \theta_\rho{}^\kappa. \tag{5}$$

Just as in other fiber bundles involving gauge groups and based on either a Minkowski or a general-relativistic space-time, the gauge group is a normal subgroup of the full symmetry group, with the coordinate transformations forming the factor group. The modification *vis-à-vis* gauge groups that act on strictly internal degrees of freedom is described by Eqs. (3)–(5), which detail what are to be considered the "same" gauge transformations in different coordinate systems.

In addition to gauge-sensitive vectors (2) there exists, of course, the standard tangent bundle, which defines gauge-invariant vectors. Gauge-sensitive covariant vectors are gauged with the help of the matrices $\Omega_\mu{}^\rho$, and again there exist gauge-invariant covariant vectors as well.

III. THE TWO AFFINE CONNECTIONS

In order to permit any sensible analysis dealing with gauge-sensitive vectors and tensors, there must be an affine connection defining the parallel displacement of gauge-sensitive vectors

$$dU^\nu = -k_\mu{}^\nu{}_\kappa U^\mu dx^\kappa. \tag{6}$$

This equation implies the existence of a covariant derivative of such a vector field

$$U^\nu{}_{|\kappa} \equiv U^\nu{}_{,\kappa} + k_\mu{}^\nu{}_\kappa U^\mu, \tag{7}$$

and determines the transformation law of $k_\mu{}^\nu{}_\kappa$ under com-

 0022-2488/85/082030-02$02.50

bined coordinate and gauge transformations

$$k_{\mu}{}^{\nu}{}_{\kappa}{}' = \frac{\partial x^{\lambda}}{\partial x^{\kappa'}} \frac{\partial x^{\rho}}{\partial x^{\mu'}} \left[\frac{\partial x^{\nu'}}{\partial x^{\sigma}} \theta_{\beta}{}^{\sigma} \Omega_{\rho}{}^{\alpha} k_{\alpha}{}^{\beta}{}_{\lambda} - \frac{\partial^2 x^{\nu'}}{\partial x^{\rho} \partial x^{\lambda}} - \frac{\partial x^{\nu'}}{\partial x^{\sigma}} \Omega_{\rho}{}^{\alpha} \theta_{\alpha}{}^{\sigma}{}_{,\lambda} \right]. \tag{8}$$

The superscript ν and the subscript μ are "gauge sensitive," whereas the subscript κ identifies the component of the one-form and is related to the standard cotangent bundle. Thus $k_{\mu}{}^{\nu}{}_{\kappa}$'s two subscripts play entirely different roles. A requirement either of symmetry or of antisymmetry with respect to the two subscripts could not be maintained under gauge transformations; no such requirement may be imposed.

However, given the metric (1) of this paper, it makes sense to restrict the affine connection (6) by requiring that its covariant derivative vanish, that it be covariantly constant

$$g_{\mu\nu|\kappa} \equiv g_{\mu\nu,\kappa} - k_{\mu}{}^{\rho}{}_{\kappa} g_{\rho\nu} - k_{\nu}{}^{\rho}{}_{\kappa} g_{\rho\mu} = 0. \tag{9}$$

This requirement reduces the number of algebraically independent components to those skew symmetric in the subscripts, the torsion, according to standard derivations. With the help of Christoffel symbols and the torsion $t_{\mu}{}^{\nu}{}_{\kappa}$, $k_{\mu}{}^{\nu}{}_{\kappa}$ can be written in the form

$$\begin{aligned} k_{\mu}{}^{\nu}{}_{\kappa} &= \{{}_{\mu}{}^{\nu}{}_{\kappa}\} + t^{\nu}{}_{\mu\kappa} + t_{\mu\kappa}{}^{\nu} + t_{\kappa\mu}{}^{\nu}, \\ t^{\nu}{}_{\mu\kappa} &\equiv k_{[\mu}{}^{\nu}{}_{\kappa]}. \end{aligned} \tag{10}$$

There is no implication that the torsion has simple transformation properties under gauge transformations. With respect to pure coordinate transformations it behaves as a tensor.

From the affine connection $k_{\mu}{}^{\nu}{}_{\kappa}$ one can construct a curvature tensor $K^{\nu}{}_{\mu\kappa\lambda}$, which arises as the commutator of two covariant differentiations

$$\begin{aligned} & U^{\mu}{}_{|\kappa\lambda} - U^{\mu}{}_{|\lambda\kappa} \equiv K^{\mu}{}_{\nu\kappa\lambda} U^{\nu}, \\ & K^{\mu}{}_{\nu\kappa\lambda} \equiv k_{\nu}{}^{\mu}{}_{\kappa,\lambda} - k_{\nu}{}^{\mu}{}_{\lambda,\kappa} - k_{\rho}{}^{\mu}{}_{\kappa} k_{\nu}{}^{\rho}{}_{\lambda} + k_{\rho}{}^{\mu}{}_{\lambda} k_{\nu}{}^{\rho}{}_{\kappa}. \end{aligned} \tag{11}$$

In its covariant version,

$$K_{\mu\nu\kappa\lambda} \equiv g_{\mu\rho} K^{\rho}{}_{\nu\kappa\lambda}, \tag{12}$$

this curvature tensor is gauge sensitive with respect to the first two indices, skew symmetric with respect to them (because the affine connection is metric), and gauge invariant as well as skew symmetric with respect to the last index pair. It does not satisfy the usual triple (Jacobi) relation, which, if valid, would involve three indices with different transformation properties. However, the Bianchi differential identities are satisfied.

The curvature tensor $K^{\mu}{}_{\nu\kappa\lambda}$ is but one example of tensors arising naturally, some of whose indices are gauge sensitive, others of which are gauge invariant. To be able to carry on a general analysis one needs a second affine connection, the latter permitting parallel displacement of gauge-invariant vectors

$$dV^{\mu} = -\lambda_{\nu}{}^{\mu}{}_{\kappa} V^{\nu} dx^{\kappa}. \tag{13}$$

The affine connection $\lambda_{\nu}{}^{\mu}{}_{\kappa}$ cannot be tied to the metric. As it is to be connected with gauge-invariant operations, its transformation law is that of an ordinary affine connection

$$\lambda_{\nu}{}^{\mu}{}_{\kappa}{}' = \frac{\partial x^{\sigma}}{\partial x^{\nu'}} \frac{\partial x^{\lambda}}{\partial x^{\kappa'}} \left(\frac{\partial x^{\mu'}}{\partial x^{\rho}} \lambda_{\sigma}{}^{\rho}{}_{\lambda} - \frac{\partial^2 x^{\mu'}}{\partial x^{\sigma} \partial x^{\lambda}} \right), \tag{14}$$

even in the presence of gauge transformations. With this transformation law $\lambda_{\nu}{}^{\mu}{}_{\kappa}$ can be restricted by the requirement that it be symmetric with respect to its two subscripts, that its torsion vanish. With this assumption made, $\lambda_{\nu}{}^{\mu}{}_{\kappa}$ has but 40 algebraically independent components.

Again, the affine connection permits the construction of a curvature tensor, in precise analogy to Eq. (11). This curvature tensor $\Delta^{\mu}{}_{\nu\kappa\lambda}$ will be skew symmetric with respect to its last two indices, and it will satisfy the algebraic triple relations with respect to its three subscripts, as well as the differential Bianchi identities. As there is no way to lower its superscript, no relation analogous to the skew symmetry of $K^{\mu}{}_{\nu\kappa\lambda}$, Eq. (12), with respect to μ and ν, can be postulated in an invariant manner.

IV. RELATIONSHIP TO $A_{\nu}{}^{\mu}{}_{\kappa}$

With respect to pure coordinate transformations the two affine connections $k_{\nu}{}^{\mu}{}_{\kappa}$ and $\lambda_{\nu}{}^{\mu}{}_{\kappa}$ satisfy the same transformation law (14), hence their difference, though not a geometric object under gauge transformations, is a tensor under pure coordinate transformations. This difference is the structure $A_{\nu}{}^{\mu}{}_{\kappa}$ of Ref. 1

$$A_{\nu}{}^{\mu}{}_{\kappa} = \lambda_{\nu}{}^{\mu}{}_{\kappa} - k_{\nu}{}^{\mu}{}_{\kappa}. \tag{15}$$

Given the two affine connections, $A_{\nu}{}^{\mu}{}_{\kappa}$ is fully determined.

Conversely, the two affine connections are uniquely determined if the metric and $A_{\nu}{}^{\mu}{}_{\kappa}$ are known. Because by assumption $\lambda_{\nu}{}^{\mu}{}_{\kappa}$ is symmetric in its subscripts, it follows from Eq. (15) that

$$t_{\nu}{}^{\mu}{}_{\kappa} = -A_{[\nu}{}^{\mu}{}_{\kappa]}. \tag{16}$$

With this information the affine connection $k_{\nu}{}^{\mu}{}_{\kappa}$ can be constructed on the basis of Eq. (10). Once $k_{\nu}{}^{\mu}{}_{\kappa}$ has been determined, $\lambda_{\nu}{}^{\mu}{}_{\lambda}$ is obtained from Eq. (15).

V. CONCLUSION

The route described in this paper starts from three geometric objects under the combined group of coordinate and gauge transformations, the metric and the two affine connections. It recovers the results of Ref. 1 without elaborate calculations. By contrast, the earlier paper[1] took as its point of departure the structure $A_{\nu}{}^{\mu}{}_{\lambda}$, whose transformation law under gauge transformations explicitly involves the metric as well, but which is a tensor under pure coordinate transformations. Either way, the number of algebraically independent field variables in this formalism is 10 at the differential level of the metric, and 64 at the level of the affine connections. Possibilities of physical theories based on this formalism will be discussed elsewhere.

ACKNOWLEDGMENTS

The authors gratefully acknowledge partial support of their work by the National Science Foundation under grants No. PHY-8209355 and No. PHY-8318350 to Syracuse University and Grant No. PHY-8406114 to New York University.

[1] A. Komar, J. Math. Phys. **26**, 831 (1985).

The harmonic mapping character of Rosen's bimetric theory of gravity and the geometry of its harmonic mapping space

W. R. Stoeger
Vatican Observatory, I-00120 Città del Vaticano, Rome, Italy

A. P. Whitman
Vatican Observatory, I-00120 Città del Vaticano, Rome, Italy and Pontifícia Universidade Católica do Rio de Janeiro, Rua Marquês de São Vicente 293, 22.451 Rio de Janeiro, Brazil

R. J. Knill
Department of Mathematics, Tulane University, New Orleans, Louisiana 70118

(Received 19 October 1984; accepted for publication 8 March 1985)

After showing that Rosen's bimetric theory of gravity is a harmonic map, the geometry of the ten-dimensional harmonic mapping space (HMS), and of its nine-dimensional symmetric submanifolds, which are the leaves of the codimension one foliation of the HMS, is detailed. Both structures are global affinely symmetric spaces. For each, the metric, connections, and Riemann, Ricci, and scalar curvatures are given. The Killing vectors in each case are also worked out and related to the "conserved quantities" naturally associated with the harmonic mapping character of the theory. The structure of the Rosen HMS is very much like that determined by the DeWitt metric on the six-dimensional Wheeler superspace of all positive definite three-dimensional metrics. It is clear that a slight modification of the Rosen HMS metric will yield the corresponding metric on the space of all four-dimensional metrics of Lorentz signature. Finally, interesting avenues of further research are indicated, particularly with respect to the structure and comparison of Lagrangian-based gravitational theories which are similar to Einstein's general relativity.

I. INTRODUCTION

One of us recently showed that Rosen's bimetric theory of gravity[1] is a harmonic map.[2] Because of the interest in harmonic maps themselves and in their application as mathematical models for physical theories,[3] it is important to work out the details of this unusual case. This is particularly true since the bimetric theory constitutes an example involving a hyperbolic manifold—the domain of the harmonic map is four-dimensional space-time. Very little has been done on such instances.[4] Furthermore, the harmonic mapping space (HMS) in the Rosen case turns out to be a very interesting ten-dimensional space, a generalization of Wheeler's superspace of all positive-definite three-metrics to one of all Lorentz four-dimensional metrics. The metric on this ten-dimensional HMS is a natural generalization of the DeWitt metric[5] on superspace.

In Sec. II we give the basic structure of Rosen's bimetric theory of gravity, indicating its harmonic mapping character. After discussing the geometry of the DeWitt metric very briefly in Sec. III, we work out the geometric characteristics of its Rosen harmonic mapping generalization in Sec. IV. We calculate the connection coefficients and curvature components for it and for its nine-dimensional leaves, which are symmetric submanifolds, and then in Sec. V determine the number and form of the Killing vectors, comparing these results with the corresponding ones obtained in the DeWitt case. In Sec. VI, we discuss briefly the "conserved quantities" associated with the Killing vectors and, in the final section, indicate two avenues for further work, the examination of other related gravitational theories and the detailed study of harmonic maps defined on noncompact hyperbolic spaces, particularly the existence, uniqueness, and construction of their solutions.

II. ROSEN'S BIMETRIC THEORY AS A HARMONIC MAPPING

As its name indicates, Rosen's theory employs two metrics

$$ds^2 = g_{\mu\nu}\, dx^\mu\, dx^\nu \tag{1}$$

and

$$d\sigma^2 = \gamma_{\mu\nu}\, dx^\mu\, dx^\nu\,. \tag{2}$$

Here $g_{\mu\nu}$ is the metric potential characterizing the interaction between mass-energy and gravity. It has as its source localized distributions of mass-energy. The $\gamma_{\mu\nu}$, in contrast, is a background metric, usually considered flat. It constitutes, therefore, a "prior geometry"—and, as such, may have various physical interpretations.

The gravitational Lagrangian density is

$$\mathscr{L}_g = (-\gamma)^{1/2} g^{\alpha\beta} g^{\lambda\rho} \gamma^{\sigma\tau} (\tfrac{1}{4} g_{\alpha\lambda|\sigma} g_{\beta\rho|\tau} - \tfrac{1}{8} g_{\alpha\beta|\sigma} g_{\nu\rho|\tau})\,. \tag{3}$$

Here a vertical bar ("|") denotes covariant differentiation with respect to $\gamma_{\mu\nu}$, and a colon (":") that with respect to $g_{\mu\nu}$. Let $\gamma \equiv \det \gamma_{\mu\nu}$. The usual variation of the full Lagrangian action yields the field equations

$$N_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} N = -8\pi\kappa T_{\mu\nu}\,, \tag{4a}$$

where

$$N_{\mu\nu} \equiv \tfrac{1}{2}\gamma^{\alpha\beta} g_{\mu\nu|\alpha\beta} - \tfrac{1}{2}\gamma^{\alpha\beta} g^{\lambda\rho} g_{\mu\lambda|\alpha} g_{\nu\rho|\beta}\,. \tag{4b}$$

We also have the conservation laws

 0022-2488/85/082032-07$02.50

$$T_{\mu}{}^{\nu}{}_{;\nu} = 0 \quad (5)$$

and the empty-space field equations

$$N_{\mu\nu} = 0. \quad (4c)$$

There are two principle reasons for introducing theories like Rosen's as alternatives to general relativity. One is to provide a reasonable devil's advocate to black holes[6–8]—and a remedy for the existence of singularities in other theories.[6] In Rosen's bimetric theory, for instance, a study of the spherically symmetric solution to the field equations shows that, although highly collapsed objects exist, they cannot be black holes, unless they have totally collapsed to a point. Event horizons do not form around extended configurations of mass-energy in Rosen's theory, no matter how concentrated they are. The second motivation for such theories is to incorporate the fundamental rest frame into gravitational theory in a more integral way than general relativity is able to do through the imposition of boundary conditions.[6] For the purposes of this paper we make no commitment to either motivation. We wish only to indicate why the theory has been proposed and studied.

If the bimetric theory is subjected to the rigors of the PPN tests,[9] it is found to be a viable alternative to general relativity at that level. However, it fails to predict the "correct" change in period for the binary pulsar PSR 1913 + 16,[10] due to the existence of gravitational dipole radiation in the theory, and it manifests other anomalies in the structure and characteristics of its gravitational waves.[9,7] Our primary interest in Rosen's theory here is, though, as a harmonic map. If we have a mapping[3,11] $\phi: M \to M', x \mapsto \phi x \equiv \phi(x)$, where M and M' are two pseudo-Riemannian manifolds with coordinates x^{μ} and ϕ^{A} and metrics $\gamma_{\mu\nu}(\mathrm{x})$ and $G_{AB}(\phi)$, respectively, it is harmonic if we have an action integral, or energy, relating the two spaces

$$I = \tfrac{1}{2} \int \sqrt{|\gamma|}\, d^{4}x\, \gamma^{\mu\nu} \frac{\partial \phi^{A}}{\partial x^{\mu}} \frac{\partial \phi^{B}}{\partial x^{\nu}} G_{AB}(\phi) . \quad (6)$$

By varying this, of course, we obtain the Euler–Lagrange, or field, equations. For further details regarding the theory of harmonic maps and their application to physical theories, see Misner.[3]

Now it can be easily seen[2] that Rosen's theory is a *bona fide* harmonic map. A simple change in indices puts the gravitational Lagrangian (3) into the harmonic map form

$$\mathscr{L}_{g} = G_{AB} \gamma^{\sigma\tau} \phi^{A}{}_{|\sigma} \phi^{B}{}_{|\tau} , \quad (7)$$

where

$$G_{AB} \equiv \tfrac{1}{4} g^{\alpha\lambda} g^{\beta\rho} - \tfrac{1}{8} g^{\alpha\beta} g^{\lambda\rho} , \quad (8)$$

and $\phi^{A} \equiv g_{\alpha\beta}$. It is clear that $G_{AB} = G_{BA}$. However, it is more convenient to use the symmetrized form

$$G_{AB} \equiv \tfrac{1}{4} g^{(\alpha|\lambda|} g^{\beta)\rho} - \tfrac{1}{8} g^{\alpha\beta} g^{\lambda\rho} . \quad (9)$$

The inverse metric is

$$G^{AB} \equiv 4 g_{(\alpha|\lambda|} g_{\beta)\rho} - 2 g_{\alpha\beta} g_{\lambda\rho} , \quad (10)$$

so that

$$G^{AB} G_{BC} = G_{\alpha\beta\lambda\rho} G^{\lambda\rho\mu\nu} = \delta^{\mu}_{(\alpha} \delta^{\nu}_{\beta)} . \quad (11)$$

In Eq. (7) the Lagrangian contains covariant derivatives with respect to the background metric, instead of simple partial derivatives, as given in Eq. (6). This does not alter its harmonic mapping character (the important properties associated with harmonic mappings continue to apply) but prevents the field equations from being written in a simple form with respect to G_{AB} and its connection, except in Minkowski coordinates (see Sec. IV below), for which covariant derivatives are equivalent to simple partial derivatives. From Eqs. (7) and (9) we see that Rosen's theory defines a harmonic mapping from four-dimensional space-time, with signature $(-+++)$, to a ten-dimensional HMS $\mathscr{M}$, in which each component of $g_{\mu\nu}$ is a generalized coordinate. That is, $\mathscr{M}$ is the space of all symmetric matrices with signature $(-+++)$, S(1,3). This space has a signature $(----++++++)$.

It is clearly important to investigate the geometry of this HMS and relate it to the underlying base space and to the theory. There may be insights into the structure of the theory which stem from its harmonic mapping character. For instance, it is known (Misner, private communication) that the Killing vectors in the HMS correspond to certain "conserved quantities" in the theory (see Sec. VI below). It turns out that our investigation here will not lead immediately to any profound insights into Rosen's theory itself. But it will reveal some very intriguing characteristics of the HMS, similar to those of Wheeler–DeWitt superspace, and point to other classes of gravitational theories which may prove interesting both in themselves and in their relationship with general relativity.

Before proceeding to discuss the geometry of the ten-dimensional HMS in Rosen's theory, we summarize the results concerning the six-dimensional Wheeler superspace on which the DeWitt metric lives.

III. THE ROSEN HMS AND THE DeWITT METRIC

As pointed out above, the metric (9), apart from a factor, is the natural four-dimensional hyperbolic generalization of the DeWitt metric, which is the metric on the superspace of all (spatial) three-dimensional metrics of positive-definite signature.[12] The geometry of that six-dimensional space has been fully worked out by DeWitt.[5]

The DeWitt metric is given by

$$G^{ijkl} = \gamma^{1/2} [\gamma^{i(|k|} \gamma^{j)l} - \gamma^{ij} \gamma^{kl}] , \quad (12a)$$

$$G_{ijkl} = \gamma^{-1/2} [\gamma_{(i|k|} \gamma_{j)l} - \tfrac{1}{2} \gamma_{ij} \gamma_{kl}] , \quad (12b)$$

where here γ_{ij} is a positive definite three-dimensional metric. Then G^{ijkl}, as we have mentioned, is a metric on a six-dimensional space M', having the signature $(-+++++)$, in which the γ_{ij} are generalized coordinates. If we use the fact that a change in the modulus, or determinant, of γ_{ij} constitutes a typical "timelike" displacement in M' and introduce ζ as a function of $\gamma = \det \gamma_{ij}$ as the "timelike" coordinate, with any other five coordinates ζ^{A} orthogonal to it, we discover that M' consists of a set of "nested" five-dimensional submanifolds all having the same intrinsic shape. These leaves are specified by ζ, $\{\zeta\} \times \overline{M}'$, where $\overline{M}'$ is the leaf which passes through the identity, and their geometry is described by the positive-definite metric $\overline{G}^{(DW)}{}_{AB}$. It turns out that $\overline{M}'$ is a symmetric Einstein space ($\overline{R}^{(DW)}{}_{AB} = K\overline{G}^{(DW)}{}_{AB}$, K constant) of negative definite Ricci curvature, upon which

SL(3,ℝ) acts [with respect to the "squaring action" $kg\tilde{k}$ for g in $\overline{M}'$, k in SL(3,ℝ), and $\tilde{k}$ the transpose of k] as the isometry group and SO(3) as the isotropy group. Futhermore, the natural symmetry on SL(3,ℝ), $\tilde{k}^{-1}$, induces a symmetry on $\overline{M}'$, which is also an isometry. Therefore, $\overline{M}' = \mathrm{SL}(3,\mathbb{R})/\mathrm{SO}(3)$ is a symmetric homogeneous space. Furthermore, although $\overline{M}'$ is geodesically complete, in M' itself there is a frontier of infinite curvature at $\zeta = 0$, since $R^{(DW)} = -20/\zeta^2$, which all geodesics will hit. For further details, see DeWitt.[5]

In performing our own calculations for the Rosen HMS, we have also checked those of DeWitt. They are essentially correct except that in his expression for the Riemann curvature of $\overline{M}'$ there should be a factor of $\frac{1}{2}$. His equation (A26) should read

$$\overline{R}^{(DW)}{}_{ABC}{}^{D} = \tfrac{1}{2}\mathrm{tr}\left[\gamma_{,B}\gamma^{-1}\gamma_{,A}\gamma^{-1}\gamma_{,C}\frac{\partial\zeta^{D}}{\partial\gamma}\right] - \tfrac{1}{2}\mathrm{tr}\left[\gamma_{,A}\gamma^{-1}\gamma_{,B}\gamma^{-1}\gamma_{,C}\frac{\partial\zeta^{D}}{\partial\gamma}\right].$$

Succeeding calculations are affected by this change. The related curvatures should be

$$\overline{R}^{(DW)}{}_{ABCD} = \tfrac{1}{2}\gamma^{1/2}\,\mathrm{tr}[\gamma^{-1}\gamma_{,D}\gamma^{-1}\gamma_{,C}\gamma^{-1} \times(\gamma_{,A}\gamma^{-1}\gamma_{,B} - \gamma_{,B}\gamma^{-1}\gamma_{,A})],$$
$$\overline{R}^{(DW)}{}_{AB} = -\tfrac{3}{4}\overline{G}^{(DW)}{}_{AB},\quad \overline{R}^{(DW)} = -40\zeta^{-2},$$
$$R^{(DW)}{}_{AB} = -\tfrac{3}{8}\overline{G}^{(DW)}{}_{AB},\quad R^{(DW)} = -20\zeta^{-2}.$$

These corrections do not substantially alter any of DeWitt's results. Furthermore, looking at the covariant derivatives of the Riemann curvature tensors for $\overline{M}'$ and M', respectively, we find that

$$\overline{R}^{(DW)}{}_{ABC}{}^{D}{}_{;E} = 0,\quad R^{(DW)}{}_{ABC}{}^{D}{}_{;E} = 0,$$
$$R^{(DW)}{}_{ABC}{}^{D}{}_{;0} = -2\zeta^{-1}R^{(DW)}{}_{ABC}{}^{D},$$
$$R^{(DW)}{}_{0BC}{}^{D}{}_{;E} = -\zeta^{-1}R^{(DW)}{}_{EBC}{}^{D},$$
$$R^{(DW)}{}_{A0C}{}^{D}{}_{;E} = -\zeta^{-1}R^{(DW)}{}_{AEC}{}^{D},$$
$$R^{(DW)}{}_{AB0}{}^{D}{}_{;E} = -\zeta^{-1}R^{(DW)}{}_{ABE}{}^{D},$$
$$R^{(DW)}{}_{ABC}{}^{0}{}_{;E} = -\zeta^{-1}R^{(DW)}{}_{ABCE},$$
$$R^{(DW)}{}_{ABC}{}^{D}{}_{,0} = 0.$$

IV. THE GEOMETRY OF THE ROSEN TEN-DIMENSIONAL HMS

In calulating the connection and curvature components of our HMS metric (9), we employ the matrix techniques introduced by DeWitt. We an expose the natural nine-dimensional leaves $\{\zeta\}\times\overline{\mathscr{M}}$ of the foliations of our ten-dimensional HMS $\mathscr{M}$ by choosing $\zeta = \ln g^{1/4}$ as the timelike coordinate and moding this out of the rest of the space. Here $g = |\det g_{\alpha\beta}|$. Thus we have

$$\mathscr{M} = \mathbb{R}\times\overline{\mathscr{M}},\quad \overline{\mathscr{M}} = \{g_{\alpha\beta}\in\mathscr{M} \,|\, g = 1\}. \tag{13}$$

Then the metric (9) can be written

$$G = G^{\alpha\beta\gamma\delta}\,dg_{\alpha\beta}\otimes dg_{\gamma\delta} = -\tfrac{1}{16}d(\ln g)\otimes d(\ln g) + \tfrac{1}{4}\mathrm{tr}[\overline{g}^{-1}d\overline{g}\otimes\overline{g}^{-1}d\overline{g}] = -d\zeta\otimes d\zeta + \tfrac{1}{4}\mathrm{tr}[\overline{g}^{-1}d\overline{g}\otimes\overline{g}^{-1}d\overline{g}],$$

or

$$G = -d\zeta\otimes d\zeta + \tfrac{1}{4}\overline{G},\quad \overline{G} \equiv \mathrm{tr}[\overline{g}^{-1}d\overline{g}\otimes\overline{g}^{-1}d\overline{g}]. \tag{14a}$$

If $\zeta = \ln g^{1/4}$ is the new time coordinate in the HMS, then leaves of constant ζ have orthogonal trajectories, as in the DeWitt case, whose tangent vectors are proportional to

$$\frac{\partial\zeta}{\partial g_{\alpha\beta}} = \frac{1}{4}\frac{1}{g}\frac{\partial g}{\partial g_{\alpha\beta}} = \frac{1}{4}g^{\alpha\beta} \tag{15}$$

and

$$G_{\alpha\beta\sigma\tau}\frac{\partial\zeta}{\partial g_{\alpha\beta}} = -g_{\sigma\tau}. \tag{16}$$

On the leaves themselves we label a set of coordinates orthogonal to the ζ coordinate by $\zeta,^{A} A = 1,\dots,9$. From Eq. (15) we have

$$g^{\alpha\beta}\frac{\partial g_{\alpha\beta}}{\partial\zeta^{A}} = 4\frac{\partial\zeta}{\partial\zeta^{A}} = 0, \tag{17}$$

and $\partial g_{\alpha\beta}/\partial\zeta$ must satisfy

$$G^{\alpha\beta\sigma\tau}\frac{\partial g_{\alpha\beta}}{\partial\zeta}\frac{\partial g_{\sigma\tau}}{\partial\zeta^{A}} = 0, \tag{18}$$

$$G^{\alpha\beta\gamma\delta}\frac{\partial g_{\alpha\beta}}{\partial\zeta}\frac{\partial g_{\gamma\delta}}{\delta\zeta} = -1. \tag{19}$$

From these relationships, we also have, using Eq. (16),

$$\frac{\partial g_{\alpha\beta}}{\partial\zeta} = g_{\alpha\beta} \tag{20}$$

and

$$g_{\alpha\beta}\frac{\partial\zeta^{A}}{\partial g_{\alpha\beta}} = \frac{\partial\zeta^{A}}{\partial\zeta} = 0. \tag{21}$$

Thus,

$$G^{\alpha\beta\gamma\delta}\frac{\partial g_{\alpha\beta}}{\partial\zeta^{A}}\frac{\partial g_{\gamma\delta}}{\partial\zeta^{B}} = \frac{1}{4}g^{\gamma\alpha}g^{\delta\beta}\frac{\partial g_{\alpha\beta}}{\partial\zeta^{A}}\frac{\partial g_{\gamma\delta}}{\partial\zeta^{B}}. \tag{22}$$

Finally, using (18), (19), and (22), our metric (14a) takes the form

$$G_{AB} = \begin{pmatrix} -1 & 0 \\ 0 & \frac{1}{4}\overline{G}_{AB}\end{pmatrix},\quad G^{AB} = \begin{pmatrix} -1 & 0 \\ 0 & 4\overline{G}^{AB}\end{pmatrix}, \tag{14b}$$

where

$$\overline{G}_{AB} = g_{\gamma\alpha}^{-1}g_{\delta\beta}^{-1}\frac{\partial g_{\alpha\beta}}{\partial\zeta^{A}}\frac{\partial g_{\gamma\delta}}{\partial\zeta^{B}}$$

or

$$\overline{G}_{AB} = \mathrm{tr}[g^{-1}g_{,A}g^{-1}g_{,B}] \tag{23}$$

and

$$\overline{G}^{AB} = \mathrm{tr}\left[g\left(\frac{\partial\zeta^{A}}{\partial g}\right)g\left(\frac{\partial\zeta^{B}}{\partial g}\right)\right]. \tag{24}$$

With the appropriate changes, all the relations DeWitt derives for his metric carry over in the Rosen case. We obtain the working identities

$$\mathrm{tr}\left(g_{,A}\frac{\partial\zeta^{B}}{\partial g}\right) = \delta_{A}{}^{B}, \tag{25}$$

$$\mathrm{tr}\left(g\frac{\partial \zeta^A}{\partial g}\right)=0, \tag{26}$$

$$\mathrm{tr}\left(g^{-1}\frac{\partial g}{\partial \zeta^A}\right)=0, \tag{27}$$

$$\mathrm{tr}\left[g^{-1}(g_{,AB}-g_{,A}g^{-1}g_{,B})\right]=0, \tag{28}$$

and

$$\frac{\partial g_{\alpha\beta}}{\partial \zeta^A}\frac{\partial \zeta^A}{\partial g_{\gamma\delta}}=\delta_{(\alpha}{}^{\gamma}\delta_{\beta)}{}^{\delta}-\tfrac{1}{4}g_{\alpha\beta}g^{\gamma\delta}. \tag{29}$$

And for arbitrary 4×4 matrices $\mathbf{M}$ and $\mathbf{N}$ we have

$$\mathrm{tr}(g_{,A}\mathbf{M})\mathrm{tr}\left(\mathbf{N}\frac{\partial \zeta^A}{\partial g}\right)=\tfrac{1}{2}\mathrm{tr}(\mathbf{MN}+\mathbf{M}\widetilde{\mathbf{N}})-\tfrac{1}{4}\mathrm{tr}(g\mathbf{M})\mathrm{tr}(g^{-1}\mathbf{N}), \tag{30}$$

$$\mathrm{tr}(g_{,AB}\mathbf{M})\mathrm{tr}\left(\mathbf{N}\frac{\partial \zeta^B}{\partial g}\right)+\mathrm{tr}(g_{,B}\mathbf{M})\mathrm{tr}\left[\mathbf{N}\left(\frac{\partial \zeta^B}{\partial g}\right)_{,A}\right]=-\tfrac{1}{4}\,\mathrm{tr}(g_{,A}\mathbf{M})\mathrm{tr}(g^{-1}\mathbf{N})+\tfrac{1}{4}\,\mathrm{tr}(g\mathbf{M})\mathrm{tr}(g^{-1}g_{,A}g^{-1}\mathbf{N}), \tag{31}$$

and

$$\mathrm{tr}\left[\frac{\partial g}{\partial \zeta^A}\mathbf{M}\frac{\partial \zeta^A}{\partial g}\mathbf{N}\right]=\tfrac{1}{2}\,\mathrm{tr}\,\mathbf{M}\,\mathrm{tr}\mathbf{N}+\tfrac{1}{2}\,\mathrm{tr}(\mathbf{M}\widetilde{\mathbf{N}})-\tfrac{1}{4}\,\mathrm{tr}[g\mathbf{M}g^{-1}\mathbf{N}]. \tag{32}$$

As in DeWitt[5] these relationships are used to give the connection coefficients and the curvature components of both $\mathscr{M}$ and $\{\zeta\}\times\overline{\mathscr{M}}$ from the metrics (14), and (23) and (24). For the nine-dimensional submanifold $\{\zeta\}\times\overline{\mathscr{M}}$, with metric $\frac{1}{4}\overline{G}_{AB}$ we have

$$\overline{G}_{AC}\overline{G}^{CB}=\delta_A{}^B, \tag{33}$$

$$\overline{\Gamma}_{ABC}=\tfrac{1}{8}(\overline{G}_{AC,B}+\overline{G}_{BC,A}-\overline{G}_{AB,C})=\tfrac{1}{8}\,\mathrm{tr}[g_{,C}g^{-1}(-g_{,A}g^{-1}g_{,B}-g_{,B}g^{-1}g_{,A}+2g_{,AB})g^{-1}], \tag{34}$$

$$\overline{\Gamma}_{AB}{}^{C}=\mathrm{tr}\left[(-g_{,A}g^{-1}g_{,B}+g_{,AB})\frac{\partial \zeta^C}{\partial g}\right], \tag{35}$$

$$\overline{R}_{ABC}{}^{D}=\tfrac{1}{2}\mathrm{tr}\left[g_{,C}g^{-1}(-g_{,A}g^{-1}g_{,B}-g_{,B}g^{-1}g_{,A})\frac{\partial \zeta^D}{\partial g}\right], \tag{36}$$

$$\overline{R}_{ABCD}=\tfrac{1}{8}\,\mathrm{tr}[g^{-1}g_{,D}g^{-1}g_{,C}g^{-1}(g_{,A}g^{-1}g_{,B}-g_{,B}g^{-1}g_{,A})], \tag{37}$$

$$\overline{R}_{ABC}{}^{D}{}_{;E}=0, \tag{38}$$

$$\overline{R}_{AB}=-4(\tfrac{1}{4}\overline{G}_{AB})=-\overline{G}_{AB},\quad \overline{R}=-36. \tag{39}$$

Therefore, each $\{\zeta\}\times\overline{\mathscr{M}}$ is a symmetric Einstein space. However, it does *not* have constant Gaussian curvature, and is *not* maximally symmetric, contrary to what was incorrectly stated in Stoeger.[2] In fact, we shall show below that each $\{\zeta\}\times\overline{\mathscr{M}}$ has 15 Killing vectors; if it were maximally symmetric it would have 45, $N(N+1)/2$. Likewise, the five-dimensional DeWitt submanifolds are symmetric spaces, but, contrary to what DeWitt[5] states and Stoeger[2] quotes in his preliminary report, they do not have constant Gaussian curvature either, and therefore are not maximally symmetric. They possess just eight Killing vectors, instead of the 15 maximum allowable. In both cases, for there to be constant Gaussian curvature, we would have to have

$$\overline{R}_{ABCD}=Ka^2(\overline{G}_{DA}\overline{G}_{CB}-\overline{G}_{DB}\overline{G}_{CA}),\qquad K\equiv[1/N(N-1)]\overline{R}. \tag{40}$$

In DeWitt's case, $a=\gamma^{1/2}$; in Rosen' bimetric case, $a=\frac{1}{2}$. Equation (39) is not enough. It is impossible to break Eq. (37) into the form (40). The trace there cannot be separated into a difference of a product of traces.

Moving on to consider the full HMS $\mathscr{M}$, keeping A, B, C, etc. $=1,2,\ldots,9$, and denoting components in the ζ direction by 0 indices, we have, similar to the DeWitt case, but with some notable differences

$$\Gamma_{AB}{}^{C}=\overline{\Gamma}_{AB}{}^{C}, \tag{41}$$

$$\Gamma_{AB}{}^{0}=0, \tag{42}$$

$$\Gamma_{A0}{}^{B}=0, \tag{43}$$

$$\Gamma_{A0}{}^{0}=\Gamma_{00}{}^{A}=\Gamma_{00}{}^{0}=0, \tag{44}$$

$$R_{ABC}{}^{D}=\overline{R}_{ABC}{}^{D} \tag{45}$$

$$R_{ABC}{}^{0}=R_{0AB}{}^{D}=R_{A0B}{}^{D}=R_{AB0}{}^{D},\ \text{etc.}=0 \tag{46}$$

(all components with 0's vanish),

$$R_{BC}=-\tfrac{3}{4}\overline{G}_{BC},\quad R_{A0}=0,\quad R_{00}=0, \tag{47}$$

$$R=-\tfrac{3}{4}(4)\overline{\delta}^B{}_B=-27. \tag{48}$$

Finally, we have the derivatives of the Riemann tensor

$$R_{ABC}{}^{D}{}_{;E}=0,\quad E=1,2,\ldots,9, \tag{49}$$

$$R_{ABC}{}^{D}{}_{;0}=0, \tag{50a}$$

$$R_{0BC}{}^{D}{}_{;E}=R_{A0C}{}^{D}{}_{;E}=R_{AB0}{}^{D}{}_{;E}=R_{ABC}{}^{0}{}_{;E}=0, \tag{50b}$$

$$R_{ABC}{}^{D}{}_{,0}=0. \tag{50c}$$

Thus, $\mathscr{M}$ is not an Einstein space [see Eq. (47)], but is locally affine symmetric with respect to the connection defined by the metric (9), as well as its nine-dimensional leaves, with respect to the induced metric connection. In fact, as may be seen, $\mathscr{M}$ and $\overline{\mathscr{M}}$ are globally affine symmetric spaces when equipped with the squaring action of $GL^+(4,\mathbf{R})$ and with the involution taking $g\in\mathscr{M}$ to $I_{1,3}\tilde{g}^{-1}I_{1,3}\in\mathscr{M}$. Here $I_{1,3}=\mathrm{diag}\,(-1,1,1,1)$.

There are several obvious differences between the geometry of superspace, on which the DeWitt metric lives, and the HMS of Rosen's bimetric theory of gravity. Besides the difference in signature—$(-\,-\,-\,-\,+\,+\,+\,+\,+\,+)$ for Rosen and $(-\,+\,+\,+\,+\,+)$ for DeWitt—and the abovementioned locally affine symmetric space structure of $\mathscr{M}$, which is lacking on M' (although both show a globally affine symmetric space structure on the leaves), there is no frontier in the Rosen case where the curvature goes to infinity. As pointed out above, this is at $\zeta=0$ in Wheeler–DeWitt superspace M'. The source of this difference, which also accounts for the lack of a locally affine symmetric space structure on M', is the presence of $\gamma^{1/2}$ in the DeWitt metric—so that the timelike coordinate in superspace is of the form $\zeta\equiv(32/3)^{1/2}\gamma^{1/4}$. The Rosen HMS metric lacks a corresponding factor of $g^{1/2}$, enabling the timelike coordinate in $\mathscr{M}$ to be written $\zeta\equiv\ln g^{1/4}$.

Finally, in the DeWitt case the leaf which passes through the identity $\overline{M}'$ can be identified with the coset space

$$\overline{M}' = \mathrm{SL}(3,\mathbb{R})/\mathrm{SO}(3),$$

with the canonical connection. We might remark that in the DeWitt case M' does have the coset space structure

$$M' = \mathrm{GL}^+(3,\mathbb{R})/\mathrm{SO}(3),$$

but the dilation factor in the metric (12a) prevents the induced metric connection from being isometric with the canonical connection on M' as a coset space. This results in M' not being a locally as well as a globally affine symmetric space. However, similar identifications do hold in the Rosen case. As Misner observed (cf. Ref. 3, p. 4517), $\mathscr{M}$ with metric (9) can be identified with the coset space

$$\mathscr{M} = S(1,3) = \mathrm{GL}^+(4,\mathbb{R})/\mathrm{SO}(1,3),$$

with the canonical connection; and its restriction to $\overline{\mathscr{M}}$ is identified with

$$\overline{\mathscr{M}} = \mathrm{SL}(4,\mathbb{R})/\mathrm{SO}(1,3).$$

We conclude that in the Rosen case $\mathscr{M}$, as well as all of the leaves isometric with $\overline{\mathscr{M}}$, are globally affine symmetric spaces. This result is consistent with the curvature caculations (49), (50a), (50b), and (38).

V. KILLING VECTORS OF THE ROSEN HMS

As we have implied above, the metric (23) on $\overline{\mathscr{M}}$ is invariant under transformations by the matrices $\mathrm{SL}(4,\mathbb{R})$, with effective action carried only by $\mathrm{SL}(4,\mathbb{R})/\{I, -I\}$. SO(1,3) is the isotropy subgroup referenced with respect to $I_{1,3}$. Thus, we are sure from this that the number of Killing vectors admitted by $\overline{\mathscr{M}}$ is at least 15, the dimension of $\mathrm{SL}(4,\mathbb{R})$.

It can be shown that the metric on the full space $\mathscr{M}$ is invariant under $\mathrm{GL}^+(4,\mathbb{R})$, yielding a minimum of 16 Killing vectors on $\mathscr{M}$. In the DeWitt case[5] the situation is, of course, similar. $\mathrm{SL}(3,\mathbb{R})$ provides effective action on $\overline{M}'$ as the isometry group, giving eight Killing vectors. But, because the dilation destroys the isometry of the metric in the ξ direction for this case, $\mathrm{GL}^+(3,\mathbb{R})$ is not the isometry group for M', which also is left with just eight Killing vectors.

However, we want to assure ourselves that there are not more "hidden" Killing vectors on $\overline{\mathscr{M}}$ and $\mathscr{M}$. We can do this by looking at the integrability conditions for the Killing equation, which are

$$L_\xi(R_{ABC}{}^{D}{}_{;E_1\ldots E_n}) = 0, \quad n = 0,1,2\ldots. \tag{51}$$

That is, the Lie derivative of successive covariant derivatives of the Riemann tensor must equal zero. Then, we can apply the theorem.[13] If the rank of the linear algebraic equations (51) for ξ_a and $\xi_{a;b}$ is q, where ξ_a are Killing vectors, then the maximal group G_r of motions of the V_N has $r = \frac{1}{2}N(N+1) - q$ parameters, i.e., the maximum number of Killing vectors is r. Application of this theorem in the Rosen and DeWitt cases is particularly simple because of Eqs.(38), (49), and (50) in the Rosen instance and Eqs. (12) for the DeWitt metric. Let us have a brief look at Eqs. (51) for the Rosen HMS.

First, we consider the nine-dimensional submanifold $\overline{\mathscr{M}}$. From Eq. (38) is it obvious that Eqs. (51) will be just

$$\mathscr{L}_\xi \overline{R}_{ABC}{}^{D} = 0. \tag{52}$$

Furthermore, it is clear that these equations will be algebraic equations for $\xi_{a;b}$ only—of which there are $\frac{1}{2}N(N-1)$ or 36. (The equation $\overline{R}_{ABC}{}^{D}{}_{;E} = 0$ automatically assures us, therefore, of at least nine Killing vectors.) How many of these are linearly independent?

Studying Eqs. (52), we discover that there are 252 separate equations corresponding to independent components of $\overline{R}_{ABC}{}^{D}$ with no two indices the same. Each of these equations has 32 terms. It is only these which affect the rank of the system; equations corresponding to components of $\overline{R}_{ABC}{}^{D}$ with two or more indices the same will have at most 24 terms and, as we shall see below, will not be able to decrease the rank of the system. Combining the 252 equations we can find the number of the 36 unknowns which are linearly independent. This will give us the rank.

The number of equations needed to reduce the rank by n will be just 2^n. Thus, to reduce it by 8, we would need 256 equations. But there are only 252 available to us. Therefore, we can reduce the rank by 7; it is then $q = 36 - 7 = 29$. Because of the gap between 29 and 24, the maximum number of unknowns contained in the other equations of the system, it is clear that there will not be enough of them to reduce the rank further.

Thus, according to the theorem above, the number of Killing vectors admitted by $\overline{\mathscr{M}}$ will be $\frac{1}{2}N(N+1) - q = 45 - 29 = 16$. However, the dilation invariance condition defining the manifold $\overline{\mathscr{M}}$ gives us another relation which reduces the number to 15, just the number given by the dimension of $\mathrm{SL}(4,\mathbb{R})$.

When we examine the full space $\mathscr{M}$, we find that the extra equations arising from the covariant derivatives (50a) will not alter the rank from that given by the $\overline{\mathscr{M}}$ system. So the rank will again be 29; and the number of Killing vectors will be $\frac{1}{2}N(N+1) - q = 55 - 29 = 26$ for the ten-dimensional HMS. However, from the same set of equations we can show that the $\xi_{0;E}$, $E = 1,\ldots,9$, and ξ_0 are zero, thus reducing the total number of Killing vectors in $\mathscr{M}$ to 16, the number dictated by the dimension of $\mathrm{GL}^+(4,\mathbb{R})$.

In the DeWitt case, a similar analysis leads to the conclusion that the number of Killing vectors is 8 for $\overline{M}'$, corresponding to $\mathrm{SL}(3,\mathbb{R})$.

Finally, we can work out the form of these Killing vectors. They will be the infinitesimal transformations $\xi_{\alpha\beta}$ of the metric $g_{\alpha\beta}$,

$$g'_{\alpha\beta} = g_{\alpha\beta} + \xi_{\alpha\beta},$$
$$\text{or } g' = \tilde{A}gA, \quad A = (e^{-\beta\lambda})_j{}^l, \tag{53}$$

which leave the HMS metric G_{AB} invariant. Therefore,

$$\xi_{\alpha\beta} = \frac{dg_{\alpha\beta}}{d\lambda}. \tag{54}$$

Carrying out this differentiation, we obtain

$$\xi_{\alpha\beta} = -g_{(\alpha|\gamma|}\beta^{\gamma}{}_{\beta)}, \tag{55}$$

where $\beta^\gamma_\beta \in \mathrm{sl}(4,\mathbb{R})$ for $\overline{\mathscr{M}}$, and $\beta^\gamma_\beta \in \mathrm{gl}(4,\mathbb{R})$ for $\mathscr{M}$. Here, $\mathrm{sl}(4,\mathbb{R})$ and $\mathrm{gl}(4,\mathbb{R})$ are the Lie algebras of $\mathrm{SL}(4,\mathbb{R})$ and $\mathrm{GL}(4,\mathbb{R})$, respectively. Any bases for these two Lie algebras will give the 15 Killing vectors for $\overline{\mathscr{M}}$ and the 16 for $\mathscr{M}$

through Eq. (55). In operator form, they will be, of course, just

$$X^{(i)} = -g_{(\alpha|\gamma|}\beta^{\gamma}{}_{\beta)}{}^{(i)}\frac{\partial}{\partial g_{\alpha\beta}}. \tag{56}$$

This form of the Killing vectors in the DeWitt case has already been given by Jantzen[14] and in a rather unclear expression by Ryan.[15]

In the DeWitt case, in which $\bar{M}'$, $=\mathrm{SL}(3,\mathbb{R})/\mathrm{SO}(3)$, all the information concerning the maximum number of Killing vectors and their particular form may be obtained by determining the automorphisms and inner automorphisms of SO(3) which extend to the automorphisms of SL(3,$\mathbb{R}$) and using a theorem of Cartan.[16] It is not yet clear whether this approach can be extended to the Rosen case, in which $g_{\mu\nu}$ has a Lorentz signature.

These Killing vector fields on the HMS we have been discussing can be used to construct new solutions to the field equations.[17] If ϕ^A is a known solution, then there are new solutions $\tilde{\phi}^B$, where $\tilde{\phi}^B$ is obtained from ϕ^A by an isometry of the HMS.

VI. HARMONIC MAPPING "CONSERVED QUANTITIES"

In unpublished work, Misner (private communication) pointed out that, within the context of harmonic mapping theory, each Killing vector in HMS generates a conserved quantity in the given field theory. How does this observation apply to the Rosen bimetric theory? It happens that one must be quite careful in interpreting these quantities, particularly in the Rosen case. The conserved quantities indicated by Misner arise essentially from both the Killing equation and from the harmonic mapping field equations. The former is, of course,

$$\xi_{A;B} + \xi_{B;A} = 0. \tag{57}$$

The field equations, when the harmonic mapping action is expressed in the form (6), can be written simply as[3]

$$\phi^{A}{}_{\mu}{}^{;\mu} = 0, \tag{58}$$

where $\phi^A{}_\mu \equiv \partial\phi^A/\partial x^\mu$ and where the covariant derivative [which we here express using a semicolon (";")] is with respect to the induced connection $\Gamma^A{}_{B\mu} = \Gamma^A{}_{BC}\,\partial\phi^C/\partial x^\mu$, determined by the metric G_{AB}. Then it is clear that we have "conserved quantities"

$$J_\mu{}^{(\xi)} = \xi_A\phi^A{}_\mu, \tag{59}$$

such that

$$J_\mu{}^{(\xi);\mu} = 0, \tag{60}$$

since

$$J_\mu{}^{(\xi);\mu} = \xi_A{}^{;\mu}\phi^A{}_\mu + \xi_A\phi^A{}_\mu{}^{;\mu}.$$

The second term vanishes because of the field equations (58), and $\xi_A{}^{;\mu} = \xi_{A;C}\phi^{C\mu}$, which vanishes by virtue of the Killing equation (57).

In Rosen's bimetric theory, however, we cannot represent the field equations, generally speaking, in the form (58), because of the presence of the covariant derivatives with respect to the background metric in the gravitational Lagrangian (7). When the harmonic action contains covariant derivatives, that is, when

$$I = \tfrac{1}{2}\int \sqrt{|\gamma|}\, d^4x\, \gamma^{\mu\nu}(x)\phi^A{}_{|\mu}\phi^B{}_{|\nu}G_{AB}(\phi), \tag{61}$$

where the vertical bar ("|") indicates covariant differentiation with respect to the metric $\gamma_{\mu\nu}(x)$, the field equations take the general form

$$-\gamma^{\mu\nu}\phi_{B|\mu\nu} + \gamma^{\mu\nu}\Gamma^D{}_{AB}\phi^A{}_{|\mu}\phi_{D|\nu} = 0, \tag{62}$$

where $\Gamma^D{}_{AB}$ is the connection corresponding to the metric G_{AB}. We can only write the field equations in a form like (58), i.e., in terms of a covariant divergence of some quantity with respect to the connection $\Gamma^A{}_{BC}$, when the action contains only partial derivatives. There appears to be no *simple* generalization of (58) when the action contains covariant derivatives.

Therefore, we can obtain the conserved quantities associated with Rosen's theory by writing $\gamma_{\mu\nu}$ in Minkowski coordinates, so that the covariant derivatives in the action are equivalent to partial derivatives and the field equations can then be written in the form (58). Then we shall have the 16 conserved quantities

$$J^{(\xi)}_{\mu(R)} = \xi_A\phi^A{}_\mu = -g_{(\alpha|\gamma|}\beta^\gamma{}_{\beta)}\frac{\partial g^{\alpha\beta}}{\partial x^\mu}, \tag{63}$$

where x^μ are Minkowski coordinates and $\beta^\gamma{}_\beta \in \mathrm{gl}(4,\mathbb{R})$. There does not seem to be any special significance connected with these quantities, other than that the

$$J^{(\xi);\mu}_{\mu(R)} = 0 \tag{64}$$

are precisely the projection of the field equations along the 16 Killing vector directions in the HMS. It should be noted that these conserved quantities have an entirely different status than those specified by $T_\mu{}^\nu{}_{;\nu}$ [Eq. (5)]. The connections with respect to which the covariant derivatives are taken are different; in (5) it is with respect the one determined by the dynamical metric $g_{\mu\nu}$. Secondly, the conserved quantities in (62) are strictly gravitational quantities. They do not set any constraints on the mass-energy of the sources, whereas those of Eq. (5) obviously do.

In fact, in discussing Rosen's bimetric theory as a harmonic map, we have limited ourselves to a consideration of the gravitational action, which yields the vacuum field equations (4c). The matter Lagrangian would not, generally speaking, preserve the harmonic mapping form.

VII. DISCUSSION

Rosen's bimetric theory was first constructed as a gravitational theory and only then recognized as a harmonic map. And so it is intrinsically worthwhile to study its harmonic mapping structure, especially because it is a complete and consistent theory of gravity, and at least marginally viable at the PPN level. Moreover, it is a very complicated and unusual harmonic map—from the Lorentz four-manifold of space-time into a ten-dimensional manifold of signature $(- - - - + + + + + +)$—and one of the few instances of a harmonic map defined on a noncompact hyperbolic manifold.

Finally, as we have seen, the HMS itself is very intriguing and closely related in structure to the well-studied

Wheeler–DeWitt superspace. It is easy to see from both cases what the structure of the closely related space of all Lorentz four-metrics would be like.

Though our analysis of the harmonic mapping character of Rosen's theory has so far yielded few new physical insights—most of what we have done has, instead, highlighted its mathematical structure—there are related avenues of research which now seem promising. Primary among these would be to look at the structure and observational characteristics of the several obvious "relatives" of Rosen's theory, those with similar gravitational actions. Some of these are also *bona fide* harmonic mappings, with two metrics; several others are not, and, of those, a couple possess only one metric. Examination of this class of theories, which we have already begun, is oriented towards better understanding the mathematical roots of physical characteristics in Lagrangian-based theories, including general relativity.

It is hoped that by looking at a number of closely related theories, which possess varying similarities and differences, we may obtain a better understanding of how certain observable characteristics, e.g., the existence of event horizons or the gravitational-wave structure, originate in differing mathematical structures.

There are, finally, several other areas of investigation which may prove fruitful. One is the possible use of the ten-dimensional superspace of all four-dimensional Lorentz metrics, which is closely related to Rosen's HMS. Another on the more purely mathematical side is to work at extending the results of harmonic mapping theory—particularly with respect to the uniqueness, existence, and stability of solutions—to cases involving hyperbolic manifolds. Rosen's theory provides a nontrivial example.

ACKNOWLEDGMENTS

We gratefully acknowledge important discussions with R.T. Jantzen during the course of this work. W.R.S. expresses his appreciation for the help and encouragement of C.W. Misner at the beginning of the research.

[1]N. Rosen, Gen. Relativ. Gravit. **4**, 435 (1973); Ann. Phys. (NY) **84**, 455 (1974); Gen. Relativ. Gravit. **6**, 259 (1975).

[2]W. R. Stoeger, in *Proceedings of the Third Marcel Grossman Meeting on Recent Developments in General Relativity*, Shanghai, edited by Hu Ning (North-Holland, Amsterdam, 1983), part B, pp. 921–925.

[3]C. W. Misner, Phys. Rev. D **18**, 4510 (1978).

[4]But see C.-H. Gu, Commun. Pure Appl. Math **33**, 727 (1980) and T. Klotz Milnor, Proc. Nat. Acad. Sci. U.S.A. **79**, 2143 (1982).

[5]B. S. DeWitt Phys. Rev. **160**, 1113 (1967).

[6]Cf. N. Rosen, Found. Phys. **10**, 673 (1980).

[7]W. R. Stoeger, Gen. Relativ. Gravit. **9**, 165 (1978).

[8]W. R. Stoeger Gen. Relativ. Gravit. **10**, 671 (1979); Mon. Not. R. Astron. Soc. **190**, 715 (1980).

[9]D. L. Lee, C. M. Caves, W.-T. Ni, and C. M. Will, Astrophys. J. **206**, 555 (1976); see also C. M. Will, *Theory and Experiment in Gravitational Physics* (Cambridge U. P. Cambridge, England, 1981), p. 131.

[10]C. M. Will and D. M. Eardley, Astrophys. J. **212**, L91 (1977); C. M. Will, in Ref. 9, pp. 307–309.

[11]J. Eells and J. H. Sampson, Am. J. Math. **86**, 109 (1964); J. Eells and L. Lemaire, Bull. London Math. Soc. **10**, 1 (1978); "Selected Topics in Harmonic Maps," in *CBMS Regional Conference Lecture Notes* No. 50 (1984).

[12]This was first pointed out to us by R. T. Jantzen. Our discussions with him and his checking of some of our work was essential to its completion. We are most appreciative of his contribution.

[13]Compare with D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Cambridge U. P., Cambridge, England, 1980), p. 100.

[14]R. T. Jantzen, In *Cosmology of the Early Universe*, edited by L. Z. Fang and R. Ruffini (World Scientific, Singapore, 1984), p. 257, Eq. 2.41.

[15]M. Ryan, *Hamiltonian Cosmology* (Springer New York, 1972), p. 137.

[16]Compare with J. Wolf, *Spaces of Constant Curvature* (McGraw–Hill, New York, 1967), p. 264, Theorem 8.8.1.

[17]M. Halilsoy, Phys. Lett. A **84**, 404 (1981).

Boson approximants for lattice Fermi systems

Piotr Garbaczewski
Institute of Theoretical Physics, University of Wrocław, 50-205 Wrocław, Poland

(Received 24 July 1984; accepted for publication 14 December 1984)

It is demonstrated that Bose systems can be used to simulate properties of their Fermi partners in thermal bath, provided the spectral property $H_B = PH_BP + (1-P)H_B(1-P)$, $H_F = PH_BP$ can be proved in the state space of the Bose system. The approximation accuracy can be made arbitrarily good by varying the (free) coupling parameter $\lambda \in (0,\infty)$, and permits studying of fermionic partition or correlation functions in a finite volume by means of the standard techniques (Bose:path integral, Monte Carlo, etc).

I. MOTIVATION

Among many researchers studying numerical (Monte Carlo) simulation of Fermi systems,[1–9] the so-called pseudofermion method has gained some popularity.[3–5] Though called pseudofermions, the objects used to replace the traditional Grassmann algebra elements are the commuting c-number functions. Since the commuting function ring is used to quantize Bose systems via path integration, in the pseudofermion approach we have in fact the Bose system attributed to the Fermi one of interest.

The customary prejudice is that if no Bose–Fermi (e.g., Coleman's) equivalence can be established, then Bose and Fermi systems are considered disjointedly. On the other hand, our investigations on quantization of spinor fields[10] have shown that in principle each fermion system (irrespective of the space-time dimensionality) can be embedded in the related (mother) Bose system. In application to lattice models, this embedding observation suggests looking for such Bose systems and such physical situations (temperature and coupling constant regimes) which allow for an unambiguous separation of Fermi contributions from all relevant characteristics of the Bose system. This idea underlies the use of boson expansion methods in the study of finite spin or Fermi lattices,[11,12] and this of the spin-$\frac{1}{2}$ approximation concept for lattice bosons.[13]

Our aim is to investigate a family of Fermi models in one space dimension (for which the Monte Carlo tests are usually performed) to prove that their partition and correlation functions can be arbitrarily well approximated by means of those for the related Bose systems. Since this Bose approximation scheme is considered for a family of solvable models, the reliability of the method can be easily tested against the exact results.

II. BOSE–FERMI INTERPLAY FOR THE SIMPLEST FERMION MODELS ON A ONE-DIMENSIONAL LATTICE

Let us consider the family of fermion models for which the Monte Carlo simulation was attempted.[8,9] We shall be interested in the spinless fermion hopping problem

$$H = -J\sum_k (c^*_{k+1}c_k + c^*_k c_{k+1}),$$
$$[c_k, c^*_l]_+ = \delta_{kl}, \quad [c_k, c_l]_+ = 0 \tag{2.1}$$

and its modifications received from (2.1) by adding the different density–density interaction terms

$$V_1 = V\sum_k n_{k+1}n_k, \quad n_k = c^*_k c_k,$$
$$V_2 = V\sum_k \left(n_k - \frac{1}{2}\right)\left(n_{k+1} - \frac{1}{2}\right), \tag{2.2}$$

plus (one-flavor case) the lattice Gross–Neveu model

$$H = \sum_k \{-J(c^*_k c_{k+1} + c^*_{k+1}c_k) + \Delta(-1)^k c^*_k c_k - \tfrac{1}{2}V(n_k - n_{k+1})^2\}. \tag{2.3}$$

The particle number operator N is conserved for each of these models $[H,N]=0$, $N = \Sigma_j n_j$, hence the general form of the eigenvectors is immediate

$$N|f) = n|f),$$
$$|f) = \sum_{(i)} f_{i_1,\ldots,i_n} c^*_{i_1}\cdots c^*_{i_n}|0), \quad c_j|0) = 0, \quad \forall_j. \tag{2.4}$$

The eigenvalue equations for the problem (2.1) can be written as follows:

$$\begin{aligned}
&n=0, \quad H|0) = 0,\\
&n=1, \quad H|f) = -J\sum_{j=1}^{M}(f_j c^*_{j+1} + f_{j+1}c^*_j|0)\\
&\qquad\qquad = -2J\sum_j \frac{1}{2}(f_{j-1} + f_{j+1})c^*_j|0),\\
&n=2, \quad H|f) = -2J\sum_{i<j}\frac{1}{2}[(f_{i-1j} + f_{i+1j})\\
&\qquad\qquad + (f_{ij-1} + f_{ij+1})]c^*_i c^*_j|0),\\
&\vdots\\
&n=k, \quad H|f) = -2J\sum_{i_1<\cdots<i_k}\sum_{j=i_1}^{i_k}\frac{1}{2}(f_{i_1\cdots j-1\cdots i_k}\\
&\qquad\qquad + f_{i_1\cdots j+1\cdots i_k})c^*_{i_1}\cdots c^*_j\cdots c^*_{i_k}|0).
\end{aligned} \tag{2.5}$$

We assume that our fermions are embedded in the canonical commutation relations (CCR) algebra generated[10] by operators

 0022-2488/85/082039-06$02.50

$$[a_k,a_l^*]_- = \delta_{kl}\,, \quad [a_k,a_l]_- = 0 = [a_k^*,a_l^*]_-\,,$$
$$a_k|0) = c_k|0) = 0\,, \quad \forall k\,, \quad a_k^*|0) = c_k^*|0)\,, \tag{2.6}$$

which, according to Refs. 10 and 11, implies that all Fermi states of the Bose system belong to a proper subspace [including $|0)$] $\mathscr{H}_F$ of the Bose state space $\mathscr{H}_B$, so that

$$|f) = |f)_F = \sum_{i_1<\cdots<i_m} f_{i_1\cdots i_m}\, c_{i_1}^*\cdots c_{i_n}^*|0)$$
$$= \sum_{i_1<\cdots<i_m} f_{i_1\cdots i_m}\cdot \epsilon_{i_1\cdots i_m}\cdot a_{i_1}^*\cdots a_{i_m}^*|0)\,, \tag{2.7}$$

where $\epsilon_{i_1\cdots i_m}$ is the completely antisymmetric (Levi-Civita) tensor taking values 0, $\pm$ 1. Consequently the tensor coefficient $\overset{1}{f} = f_{i_1\cdots i_n}\cdot \epsilon_{i_1\cdots i_n}$ is symmetric and vanishes if any two indices coincide.

Let us now enact the following Bose Hamiltonian:

$$H_B = -J\sum_k (a_k^* a_{k+1} + a_{k+1}^* a_k)\,, \tag{2.8}$$

on vectors of the form $|f)$. For $n=2$ we have

$$H_B|f)$$
$$= -J\sum_{i<j} \overset{1}{f}_{ij}(\overset{*}{a}_{i+1}\overset{*}{a}_j + \overset{*}{a}_{j+1}a_i + \overset{*}{a}_{i-1}\overset{*}{a}_j + \overset{*}{a}_{j-1}\overset{*}{a}_i)|0)$$
$$= -J\sum_{i<j}(\overset{1}{f}_{i-1j} + \overset{1}{f}_{ij-1} + \overset{1}{f}_{i+1j} + \overset{1}{f}_{ij+1})\overset{*}{a}_i\overset{*}{a}_j|0)\,, \tag{2.9}$$

where $\overset{1}{f}_{ik} = \epsilon_{ik} f_{ik}$. Now it is enough to observe that given $(i,j), i<j$ implies either $\epsilon_{ik} = \epsilon_{i-1j} = \epsilon_{ij-1} = \epsilon_{i+1j} = \epsilon_{ij+1}$, which corresponds to $i<j$, $i+1<j$, $i<j-1$, or $\epsilon_{ik} = \epsilon_{i-1j} = \epsilon_{ij+1}$, which corresponds to $i+1=j$, i.e., $i=j-1$ (then $\overset{1}{f}_{ij-1} = 0 = \overset{1}{f}_{i+1j}$). Consequently, the coefficient function $\overset{1}{f}_{ij}$, if multiplied by ϵ_{ij} (note that $\epsilon_{ij}^2 = 0,1$), exactly coincides with this appearing in $\Sigma_{i<j} f_{ij} c_i^* c_j^*|0) = |f)$ provided we exploit the identity

$$\epsilon_{ij} a_i^* a_j^*|0) = c_i^* c_j^*|0) \tag{2.10}$$

implied by the "bosonization" discussions in Refs. 10 and 11.

A generalization to arbitrary n is obvious with the result that eigenstates of $H = H_F$ may happen to be those of H_B (the joint Bose–Fermi spectral problem of Ref. 14)

$$f_{iM+1} = f_{i1}\,, \quad f_{M+1,j} = f_{1j}$$
$$\Rightarrow H_F|f) = \epsilon|f) \equiv H_B|f) = H_F|f) = \epsilon|f)\,. \tag{2.11}$$

It is useful to observe that the operator unit 1_F of the Fermi algebra plays in $\mathscr{H}_B$ the role of the projection operator

$$1_F\mathscr{H}_B = \mathscr{H}_F\,,$$
$$1_F|f) = |f) \Leftrightarrow |f) \in \mathscr{H}_F\,, \tag{2.12}$$

and because of (2.11) we may expect to have

$$[H_B, 1_F]_- = 0\,,$$
$$H_B = 1_F H_B 1_F + (1-1_F)H_B(1-1_F)$$
$$= H_F + (1-1_F)H_B(1-1_F)\,, \quad H_F = 1_F H_B 1_F\,, \tag{2.13}$$

which is an orthogonal decomposition since $1_F(1-1_F) = 0$, 1 being the operator unit of the Bose algebra. As a straightforward consequence, we realize then that

$$\mathscr{Z}_B = \mathrm{tr}\exp(-\beta H_B)$$
$$= \mathrm{tr}\exp(-\beta H_F) + \mathrm{tr}\exp[-\beta(1-1_F)H_B(1-1_F)]$$
$$\doteq \mathscr{Z}_F + R\,. \tag{2.14}$$

Hence the Bose trace formula includes the Fermi trace formula as a well-defined, but to be extracted, contribution.

Quite analogously, in the case of the bosonic correlation functions

$$\mathscr{Z}_B\,\rho_k[(m_1,\beta_1),\dots,(m_k,\beta_k)]$$
$$= \mathrm{tr}[\phi_{m_1}(\beta_1)\cdots\phi_{m_k}(\beta_k)\exp(-\beta H_B)]\,, \tag{2.15}$$
$$\beta_1 \leqslant \cdots \leqslant \beta_k \leqslant \beta\,, \quad \phi_m(\beta) = \exp(\beta H_B)\cdot\phi_m\cdot\exp(-\beta H_B)\,,$$

we arrive at

$$(\mathscr{Z}_F + R)\rho_k[(m_1\beta_1),\dots,(m_k,\beta_k)]$$
$$= \mathrm{tr}[\sigma_{m_1}^1(\beta_1)\cdots\sigma_{m_k}^1(\beta_k)\exp(-\beta H_F)]$$
$$+ R\,[(m_1,\beta_1),\dots,(m_k,\beta_k)]\,, \tag{2.16}$$
$$\sigma_m^1(\beta) = \exp\beta H_F\cdot\sigma_m^1\cdot\exp(-\beta H_F)\,,$$
$$\sigma_m^1 \doteq 1_F\phi_m 1_F\,, \quad \phi_m = 2^{-1/2}(a_m^* + a_m)\,.$$

Hence, upon dividing both sides of (2.16) by Z_F, the fermionic contribution explicitly arises in the general formula

$$(\mathscr{Z}_B/\mathscr{Z}_F)\rho_B = \rho_F + R/\mathscr{Z}_F\,. \tag{2.17}$$

One problem which remains is that the observation (2.17) is not that useful, unless the spin-$\frac{1}{2}$ lattice (fermionic) contribution can be viewed as dominant.

There is also another big problem (I would like to thank the referee for pointing out the issue): the discussion (2.9)–(2.11) does not yet provide the guarantee that the projection P with the properties $H_F = PH_BP$ and $H_B = PH_BP + (1-P)H_B(1-P)$ exists in the state space of the general Bose system. At this point it is not useless to mention that the Kac–Moody algebra which is developed in Refs. 15 and 16 is the infinite-dimensional generalization of the Lie algebra. This gives a recursive procedure for writing equations like Eq. (2.17) in terms of initial data, say q and r and all derivatives $q_x, q_{xx},\dots,r_x,r_{xx},\dots$ (see, e.g., Sec. 7 of Ref. 16, where the original bosonization of Skyrme[17,18] and others[19] is shown for Kac–Moody algebras). Presumably this observation can be used to argue that the projection operator P must exist when the Kac–Moody algebra is well defined.

In the next section we shall give an explicit construction of the projection P for the particular lattice Bose model. From the technical point of view one essential difference, if compared with Refs. 16 and 20, must be emphasized. Namely, it is that in our construction of the Fermi system in the (mother) Bose system, each lattice Bose degree is mapped into the corresponding lattice Fermi degree. This is not the case in Refs. 16 and 20, where it is essential that the so-called integral lattice is subdivided into the odd and even sublattices. Then, while the lattice boson is defined everywhere, the Fermi operators are attributed to the odd sublattice merely. In connection with different aspects of Fermi–Bose relationships see, e.g., Ref. 21.

III. EXACT SPECTRAL SOLUTION FOR THE LATTICE HOPPING

We study (alternative procedure) the model (2.1), both in its Fermi and Bose versions $\{H_F, c_k^*, c_k\}$ and $\{H_B, a_k^*, a_k\}$, respectively. Our aim is to relate them through considering the joint Bose–Fermi spectral problem in the same state space (this of the Bose system). We assume the periodic boundary conditions, which implies that both hopping Hamiltonians can be rewritten in the form

$$H = -J \sum_{ij=1}^{n} A_i^* W_{ij} A_j ,$$
$$W_{ij} = \delta_{ij-1} + \delta_{ij+1z}, \quad i,j = 1,\dots,n , \tag{3.1}$$

where the square $n \times n$ matrix W can be given as follows:

$$W = \sum_{l=0}^{n-1} c_l \gamma^{l-1} ,$$
$$c_l = \delta_{l1} + \delta_{ln-1} ,$$
$$\gamma^n = 1 ,$$
$$\gamma = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \cdots & \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}, \tag{3.2}$$

where n indicates the number of sites in the chain.

Since $W = \gamma + \gamma^{n-1}$ and $\gamma^n = 1$ the spectral problem for W is solved immediately by making use of

$$\gamma f_k = \lambda_k f_k ,$$
$$\lambda_k = \exp i(2\pi/n)k = \varphi^k , \quad \varphi = \exp i(2\pi/n) ,$$
$$k = 0,1,\dots,n-1 , \tag{3.3}$$
$$f_k = \{ f_{k\alpha} \} , \quad \alpha = 1,2,\dots,n ,$$
$$f_{k1} = 1 , \quad f_{k2} = \varphi^k , \quad f_{k3} = \varphi^{2k}, \dots, \quad \varphi_{kn} = \varphi^{(n-1)k} ,$$

which yields

$$W f_k = w_k f_k ,$$
$$W_k = \lambda_k + \lambda_k^{n-1} = \varphi^{-k}(1 + \varphi^{2k}) = \varphi^{-k} + \varphi^k$$
$$= 2\cos(2\pi/n)k . \tag{3.4}$$

Vectors $\{g_k = f_k/\sqrt{n}\}$ form an orthonormal eigensystem for W

$$\frac{1}{n}(f_k, f_l) = \frac{1}{n} \sum_{q=0}^{n-1} \exp i \frac{2\pi}{n}(k-l)q = \delta_{kl} , \tag{3.5}$$

hence a passage to a new set of conjugate variables is possible

$$\xi_k = \sum_{\alpha=1}^{n} \bar{g}_{k\alpha} a_k , \quad \xi_k^* = \sum_{\alpha=1}^{n} g_{k\alpha} a_\alpha^* ,$$
$$[\xi_k, \xi_l^*]_- = \delta_{kl} , \quad [\xi_k, \xi_l]_- = 0 ,$$
$$\eta_k = \sum_{\alpha=1}^{n} \bar{g}_{k\alpha} c_\alpha , \quad \eta_k^* = \sum_{\alpha=1}^{n} g_{k\alpha} c_\alpha^* ,$$
$$[\eta_k, \eta_l^*]_+ = \delta_{kl} , \quad [\eta_k, \eta_l]_+ = 0 , \tag{3.6}$$

which implies

$$H_B = \sum_k \left(-2J \cos \frac{2\pi}{n} k \right) \xi_k^* \xi_k ,$$
$$H_F = \sum_k \left(-2J \cos \frac{2\pi}{n} k \right) \eta_k^* \eta_k . \tag{3.7}$$

The respective eigenvectors belong to the n-body Fock space of the Bose and Fermi chains, respectively,

$$|m_1,\dots,m_n)_B = \overset{*}{\xi}{}_1^{m_1} \cdots \overset{*}{\xi}{}_n^{m_n} |0)_B ,$$
$$|p_1,\dots,p_n)_F = \overset{*}{\eta}{}_1^{p_1} \cdots \overset{*}{\eta}{}_n^{p_n} |0)_F . \tag{3.8}$$

In the boson case we shall compose the product of two level projections

$$P = \prod_k p_k , \quad p_k = :\exp(-\overset{*}{\xi}_k \xi_k): + \overset{*}{\xi}_k :\exp(-\overset{*}{\xi}_k \xi_k): \xi_k , \tag{3.9}$$

which has the following properties:

$$[H_B, P]_- = 0 , \quad H_B = P H_B P + (1-P) H_B (1-P) ,$$
$$P \xi_k^* P \equiv \sigma_k^+ , \quad P \xi_k P \equiv \sigma_k^- , \quad [\sigma_k^-, \sigma_k^+]_+ = p_k ,$$
$$[\sigma_k^\#, \sigma_l^\#]_- = 0 , \quad k \neq l , \tag{3.10}$$
$$P \overset{*}{\xi}{}_1^{k_1} \cdots \overset{*}{\xi}{}_n^{k_n} |0)_B = (\sigma_1^+)^{k_1} \cdots (\sigma_n^+)^{k_n} |0)_B , \quad \text{or } 0 ,$$
$$(\sigma_j^\pm)^k = 0 , \quad k > 1 .$$

One should here notice the identity

$$(\sigma_1^+)^{k_1} \cdots (\sigma_n^+)^{k_n} |0)_B = \overset{*}{\xi}{}_1^{k_1} \cdots \overset{*}{\xi}{}_n^{k_n} |0)_B , \quad k_i \leqslant 1 . \tag{3.11}$$

On the other hand, if we start from the fermion case, then either by using the Jordan–Wigner transformation or by exploiting the embedding of the Fermi (CAR) algebra in the Bose (CCR) algebra (see, e.g., Refs. 10–14), we are able to identify all fermion eigenvectors in the state space of the Bose system (as its Fermi states)

$$\overset{*}{\eta}{}_1^{p_1} \cdots \overset{*}{\eta}{}_n^{p_n} |0)_F = (\sigma_1^+)^{p_1} \cdots (\sigma_n^+)^{p_n} |0)_B$$
$$= \overset{*}{\xi}{}_1^{p_1} \cdots \overset{*}{\xi}{}_n^{p_n} |0)_B , \quad |0)_F = |0)_B . \tag{3.12}$$

It automatically follows that

$$H_B \overset{*}{\eta}{}_1^{p_1} \cdots \overset{*}{\eta}{}_n^{p_m} |0)_B = P H_B P \overset{*}{\eta}{}_1^{p_1} \cdots \overset{*}{\eta}{}_n^{p_n} |0)_B$$
$$= H_F \overset{*}{\eta}{}_1^{p_1} \cdots \overset{*}{\eta}{}_n^{p_n} |0)_B , \tag{3.13}$$

and in the range of the projection P there holds $P H_B P = H_F$.

The relevant observation at this point is that the projection P, though defined in terms of $\{\xi^*_k, \xi_k\}$ (and only through a^*, a expansions of $\xi^*_k \xi_k$, in terms of the initial operators $\{a^*_k, a_k\}$), is nevertheless a projection on the state (sub-)space $\mathscr{H}_F$ in $\mathscr{H}_B$, including all possible Fermi states of the Bose (CCR) algebra constructed about the Bose vacuum. In fact, either $\{a^*_k, a_k\}_{1<k<n}$ or $\{\xi^*_k, \xi_k\}_{1<k<n}$ can be used to construct the basis system in the (very same) subspace of $\mathscr{H}_B : P(\xi^*, \xi)\mathscr{H}_B = \mathscr{H}_F = P(a^*, a)\mathscr{H}_B$.

IV. FERMI STATES OF THE BOSE SYSTEM: ON SYMMETRY PROPERTIES OF THE FERMI GROUND STATE

It should be emphasized that in the above construction Fermi states of the Bose system correspond to the lowest excitation levels of the latter (i.e., the mother one). This property is quite important with respect to the Bose approximation idea[13] in application to Fermi systems. The ground state for the Bose and Fermi systems is the same here. On the other hand, the traditional way of thinking requires that the ground state of the Fermi system exhibit an antisymmetry property, while that for bosons is symmetric. Before embarking on the Bose approximant problem, let us clarify this point, by analyzing properties of the harmonic chain

$$H = \sum_j \frac{p_j^2}{2m} + \sum_{i<j} \frac{m\omega^2}{2}(q_i - q_j)^2 , \qquad [q_i, p_j]_- = i\delta_{ij} , \quad [q_i, q_j]_- = 0 = [p_i, p_j]_- , \tag{4.1}$$

where

$$P = \sum_{i=1}^N p_i , \qquad Q = \frac{1}{N}\sum_{i=1}^N q_i \Rightarrow \sum_{1<i<j<N} (q_i - q_j)^2 = N\sum_{i=1}^N q_i^2 - N^2Q^2 , \tag{4.2}$$

$[H, P]_- = 0$

allow us to replace H by the new operator ($m = 1$)

$$H_0 = H - (2N)^{-1}P^2 = \sum_{i=1}^N \left(\frac{p_i^2}{2} + \frac{N}{2}\omega^2 q_i^2\right) - \left(\frac{1}{2N}P^2 + \frac{1}{2}N^2\omega^2Q^2\right) . \tag{4.3}$$

We use the N-site harmonic oscillator basis

$$\Omega = \prod_{i=1}^{N}{}^{\otimes} (f_0)_i ,$$

$$n_k \Omega = f_0 \otimes \cdots \otimes (nf_0)_k \otimes \cdots \otimes f_0 = 0 , \quad \forall k , \qquad n_k = a^*_k a_k , \quad nf_0 = a^* a f_0 = 0 , \tag{4.4}$$

$$q_j = (a^*_j + a_j)(2\omega\sqrt{N})^{-1/2} , \quad p_j = i\left(\frac{\omega\sqrt{N}}{2}\right)^{1/2}(a^*_j - a_j) ,$$

so that

$$\omega\sqrt{N}(\hat\eta + \tfrac{1}{2}) = \frac{1}{2N}\sum_{ij} p_i p_j + \frac{\omega^2}{2}\sum_{ij} q_i q_j = \omega\sqrt{N}\left\{\frac{1}{2} + \frac{1}{N}\left[\sum_{i=1}^N n_i + \sum_{i<j}(a^*_i a_j + a^*_j a_i)\right]\right\} \tag{4.5}$$

and

$$\hat\eta = \frac{1}{N}\left[\sum_{i=1}^N n_i + \sum_{i<j}(a^*_i a_j + a^*_j a_i)\right] , \tag{4.6}$$

$$[a_i, a^*_j]_- = \delta_{ij} , \quad [a_i, a_j]_- = 0 , \quad a_j\Omega = 0 , \quad \forall j ,$$

where

$$\left[\sum_i n_i , \sum_{i<j}(a^*_i a_j + a^*_j a_i)\right]_- = 0 .$$

We observe that (compare, e.g., the lattice-hopping problem)

$$\sum_{i<j}(a^*_i a_j + a^*_j a_i) = \sum_{i,j} a^*_i \omega_{ij} a_j , \quad \omega_{ij} = (1 - \delta_{ij}) , \qquad \omega = \sum_{l=0}^{N-1} c_l \gamma^l , \quad c_l = (1 - \delta_{0l}) , \tag{4.7}$$

$$\gamma f_k = \lambda_k f_k , \quad g_k = (1/\sqrt{n}) f_k ,$$

which implies

$$\xi_k = \sum_{\alpha=1}^N \bar g_{k\alpha} a_\alpha , \quad \xi^*_k = \sum_{\alpha=1}^N g_{k\alpha} a^*_\alpha ,$$

$$[\xi_k, \xi^*_l]_- = \delta_{kl} , \quad \sum_{\alpha\beta} a^*_\alpha \omega_{\alpha\beta} a_\beta = \sum_k \left(\sum_l c_l \varphi^{k\cdot l}\right)\xi^*_k \xi_k , \tag{4.8}$$

$$k \neq 0 \Rightarrow \sum_l c_l \varphi^{k\cdot l} = \sum_{l\neq 0} \varphi^{k\cdot l} = -\varphi^0 = -1 ,$$

$$k = 0 \Rightarrow \sum_l c_l \varphi^{k\cdot l} = N - 1 ,$$

i.e.,

$$\sum_{\alpha\beta} a^*_\alpha \omega_{\alpha\beta} a_\beta = -\sum_{k\neq 0} \xi^*_k \xi_k + (N-1)\xi^*_0 \xi_0 = -\sum_{k=0}^{N-1} \xi^*_k \xi_k + N\xi^*_0 \xi_0 .$$

After accounting for

$$\sum_{k=0}^{N-1} \xi^*_k \xi_k = \sum_{\alpha=1}^N a^*_\alpha a_\alpha = \sum_{\alpha=1}^N n_\alpha , \tag{4.9}$$

we arrive at a complete spectral solution for $H_0|E) = E|E)$ in terms of

$$H_0 = \omega\sqrt{N}\left[\sum_{k=0}^{N-1}\left(\xi^*_k \xi_k + \frac{1}{2}\right) - \left(\xi^*_0 \xi_0 + \frac{1}{2}\right)\right] ,$$

$$|n_1, \ldots, n_N) = \xi^{*n_1}_0 \xi^{*n_2}_1 \cdots \xi^{*n_N}_{N-1}\Omega , \tag{4.10}$$

$$H_0\Omega = \omega\sqrt{N}\tfrac{1}{2}(N-1)\Omega ,$$

the eigenvectors being given up to normalization and the so-called permutation degeneracy (it is a real surprise to the author that this complete solution was not produced in papers devoted to the harmonically coupled chains[22–24]).

In the form

$$H_0=\omega\sqrt{N}\left[\sum_{k=0}^{N-1}\left(\xi^*_k\xi_k+\frac{1}{2}\right)-\left(\xi^*_0\xi_0+\frac{1}{2}\right)\right]$$

or $H_0=H_0(a^*,a)$ [(4.3) and (4.5)] the harmonic chain can be considered either in its Bose or Fermi version by using the methods of Sec. III. Indeed, by using a product of two-level projections as done previously, we can replace the Bose problem H_0, (4.10) by its spin-$\frac{1}{2}$ (Fermi) relative

$$PH_0P=\omega\sqrt{N}\left[\sum_{k=0}^{N-1}\left(\sigma_k^+\sigma_k^-+\frac{1}{2}\right)-\left(\sigma_0^+\sigma_0^-+\frac{1}{2}\right)\right]\doteq H_F\,, \tag{4.11}$$

$$P\xi^*_kP=\sigma_k^+\,,\quad P\xi_kP=\sigma_k^-\,,\quad [P,H_0]_-=0\,.$$

On the other hand, if we recall that each eigenvector of H_0 can be expanded with respect to the N-site harmonic oscillator basis, we find out that the choice of the Schrödinger representation converts $|n_1,...,n_N)$ into the N-point function of space variables $f_{n_1,...,n_N}(x_1,...,x_N)$. It is obvious that with respect to symmetrization or antisymmetrization of such a function, the eigenvalue problem is quite insensitive except that the permutation degeneracy of energy levels is removed. In symbolic notation we have the commutation relations $[H_0,S]_-=0=[H_0,A]_-$. However, the antisymmetrization cannot be applied blindly since the lowest-energy eigenvector which persists in this operation is the well-known one

$$|0,1,2,...,N-1)=\xi^*_1\xi^{*2}_2\cdots\xi^{*N-1}_{N-1}|0)\,,$$

$$H_0|0,1,...,N-1)=\frac{\omega\sqrt{N}}{2}(N^2-1)|0,1,...,N-1)\,. \tag{4.12}$$

The respective vector after antisymmetrization is traditionally identified as the ground state of the Fermi oscillators subject to the harmonic couplings. But in our opinion this Fermi chain notion acquires a meaning only if we refer to the field theory model

$$H=-\frac{1}{2m}\int dx\,\nabla\phi^*\nabla\phi+\frac{1}{2}\int dx\int dy\,\phi^*(x)\phi^*(y)V(x,y)\phi(y)\phi(x)\,,$$

$$[\phi(x),\phi^*(y)]_-=\delta(x-y)\,,\quad [\phi(x),\phi(y)]_-=0\,, \tag{4.13}$$

$$\phi(x)|0)=0\,,\quad \forall x\in R'\,,$$

where, depending on the (anti-) commutation relations choice, either symmetric or antisymmetric wave functions are necessary

$$|f)=\int dx_1\cdots\int dx_N\,f(x_1,...,x_N)\phi^*(x_1)\cdots\phi^*(x_N)|0)\,.$$

V. LATTICE FERMIONS IN TERMS OF BOSONS

When dealing with Fermi systems, except for $\mathscr{Z}_F$, one tries to compute thermal averages of distinct quantities, e.g., the simplest correlation functions

$$\langle c_k^*(\beta_1)c_j(\beta_2)\rangle=(1/\mathscr{Z}_F)\mathrm{tr}\{c_k^*\exp(\beta_1-\beta_2)H_F\times c_j\exp(-\beta_1+\beta_2-\beta)H_F\}\,,$$

$$\langle c_k^*c_j\rangle=(1/\mathscr{Z}_F)\mathrm{tr}\{c_k^*c_j\exp(-\beta H_F)\}\,, \tag{5.1}$$

$$\langle c_{k_1}^*\cdots c_{k_p}^*c_{j_1}\cdots c_{j_p}\rangle=\frac{1}{\mathscr{Z}_F}\mathrm{tr}\{c_{k_1}^*\cdots c_{k_p}^*c_{j_1}\cdots c_{j_p}\exp(-\beta H_F)\}\,,$$

the computation of which, according to the standard methods, involves Grassmann algebra functional integrals.

A passage from Fermi variables to Bose variables can be accomplished by first "defermionizing" the system by means of the Jordan–Wigner formulas and then constructing boson approximants for so-received spin-$\frac{1}{2}$ lattice quantities. The procedure is easy for the equal temperature correlations, since then, for example,

$$\langle c_k^*c_{k+1}\rangle=\langle\sigma_k^+\sigma_{k+1}^-\rangle\,, \tag{5.2}$$

while for distant correlations we have

$$\langle c_k^*c_l\rangle=\left\langle\sigma_k^+\left(\exp i\pi\sum_{j=k+1}^{l-1}\sigma_j^+\sigma_j^-\right)\cdot\sigma_j^-\right\rangle=\left\langle\sigma_k^+\prod_{j=k+1}^{l-1}(1-2n_j)\cdot\sigma_j^-\right\rangle\,, \tag{5.3}$$

$$n_j=c_j^*c_j=\sigma_j^+\sigma_j^-\,,$$

which implies that the fermion correlations can be established by using the sequence of boson approximants

$$\begin{aligned}&\langle a_k^*a_l\rangle\,,\quad \langle n_{k+p}a_k^*a_l\rangle\,,\quad k<p<l\,,\\&\langle n_{k+p}n_{k+q}a_k^*a_l\rangle\,,\quad k<p<q<l\,,\\&\langle n_{k+p}n_{k+q}a_k^*a_l\rangle\,,\quad k<p<q<r<l\,,\ldots\,.\end{aligned} \tag{5.4}$$

The underlying boson approximation procedure amounts to modifying the Bose Hamiltonian

$$H_B=PH_BP+(1-P)H_B(1-P)\,,\qquad H_F=PH_BP\,,\quad P\equiv 1_F\,, \tag{5.5}$$

by adding to it the operator λL

$$H_B\rightarrow H_B(\lambda)=H_B+\lambda L;\lambda\gg 1\,,\quad \lambda\in(0,\infty)\,,\qquad L=\sum_j n_j(n_j-1)\,,\quad n_j=a_j^*a_j\,. \tag{5.6}$$

If we insert $H_B(\lambda)$ in the place of H_B in all the thermal formulas, we realize that because of $PLP=0$ and $\lambda\gg 1$, the spectral problem for $H_B(\lambda)$ is determined by solving the stationary state perturbation problem with the degenerate spectrum for the operator

$$H'_B(\lambda)\doteq(1/\lambda)H_B(\lambda)=L+(1/\lambda)H_B\,, \tag{5.7}$$

where L plays the role of the initial (unperturbed) operator. The eigenvalues of L we denote $l=\Sigma_j n_j(n_j-1)\geqslant 0$ and the respective eigenvectors we denote $|l,\alpha)$, where α enumerates the pairwise orthogonal eigenvectors of L corresponding to the eigenvalue l. In their linear span, we can always find another orthogonal set $\{|l,a)\}$ such that $(l,a|H_B|l,a')=0$, $a\neq a'$ and $|l,a)=\Sigma_\alpha f_{a\alpha}|l,\alpha)$.

The stationary state perturbation theory says then that in the presence of the perturbation, the group of states $\{|l,a)\}$ is replaced by the new group $\{|\mathbf{l},\mathbf{a})\}$, whose elements in the first order have the well-known form

$$|\mathbf{l},\mathbf{a}) = |l,a) + \frac{1}{\lambda}\sum_k{}' \frac{|k)(k|H_{\rm B}|l,a)}{l-k}, \tag{5.8}$$

where Σ' indicates that summations run over all eigenstates of L except for those which form the l th eigenvector.

The (la)th energy level to the second order reads

$$\epsilon'_{la} = l + \frac{1}{\lambda}(l,a|H_{\rm B}|l,a) + \left(\frac{1}{\lambda}\right)^2 \sum_k{}' \frac{|(k|H_{\rm B}|l,a)|^2}{l-k}, \tag{5.9}$$

and ϵ'_{la} is the eigenvalue of $H'_{\rm B}(\lambda)$: $\epsilon_{la} = \lambda\epsilon'_{la}$. Because of (5.8) and $\lambda \gg 1$, the spectrum of $H_{\rm B}(\lambda)$ is characterized by a large energy gap opening between the lowest ($l=0$) group of eigenlevels and the others. In the $l=0$ sector we have $H_{\rm B}(\lambda) = H_{\rm B} = PH_{\rm B}P = H_{\rm F}$. As a consequence, the replacement of $H_{\rm B}$ by $H_{\rm B}(\lambda)$ in the thermal formulas allows us to view the $l=0$ contribution as dominant, and hence to approximate the thermal characteristics of the spin-$\frac{1}{2}$ (Fermi) system by using the Bose formulas with $H_{\rm B}(\lambda)$ instead of $H_{\rm B}$. The approximation accuracy can be made arbitrarily good with the growth of λ. For $\lambda \gg 1$ there holds

$$[1/\mathscr{Z}_{\rm B}(\lambda)]\,{\rm tr}[\phi_{m_1}(\beta_1)\cdots\phi_{m_k}(\beta_k)\exp(-\beta H_{\rm B}(\lambda))] \cong \frac{1}{\mathscr{Z}_{\rm F}}\,{\rm tr}[\sigma^1_{m_1}(\beta_1)\cdots\sigma^1_{m_k}(\beta_k)\exp(-\beta H_{\rm F})], \tag{5.10}$$

$$\beta_1 \leqslant \beta_2 \leqslant \cdots \leqslant \beta_k \leqslant \beta.$$

Let us notice that the Jordan–Wigner transformation makes it possible to get a Bose approximation scheme for the genuine Fermi variables (see, e.g., the previous discussion). In the case of the short range order, the simple formula

$$\langle c^*_k c_{k+1}\rangle = \langle \sigma^+_k \sigma^-_{k+1}\rangle \cong \langle a^*_k a_{k+1}\rangle \tag{5.11}$$

holds true.

[1] J. Kogut, Rev. Mod. Phys. **55**, 755 (1983).
[2] F. Fucito, E. Marinari, G. Parisi, and C. Rebbi, Nucl. Phys. B **180**, 369 (1981).
[3] F. Fucito and E. Marinari, Nucl. Phys. B **190**, 369 (1981).
[4] G. Bhanot and U. Heller, Phys. Lett. B **129**, 440 (1983).
[5] D. Zwanziger, Phys. Rev. Lett. **50**, 1886 (1983).
[6] S. Otto and M. Randeria, Nucl. Phys. B **220**, 479 (1983).
[7] T. Burkitt, Nucl. Phys. B **220**, 4314 (1983).
[8] R. Blankenbecler and R. Sugar, Phys. Rev. D **27**, 1304 (1983).
[9] J. E. Hirsch, R. Sugar, D. Scalapino, and R. Blankenbecler, Phys. Rev. B **26**, 5033 (1983).
[10] P. Garbaczewski, J. Math. Phys. **19**, 642 (1978); **22**, 442 (1982); **24**, 341 (1983).
[11] P. Garbaczewski, Phys. Rep. C **36**, 65 (1978).
[12] J. Dobaczewski, Nucl. Phys. A **369**, 219 (1981).
[13] P. Garbaczewski, J. Math. Phys. **21**, 2670 (1980); **22**, 574 (1981); **24**, 651 (1983).
[14] P. Garbaczewski, J. Math. Phys. **25**, 862 (1984).
[15] D. I. Olive and N. Turok, Nucl. Phys. B **220**, 491 (1983).
[16] P. Goddard and D. I. Olive, "Algebras, lattices and strings," DAMPT83 22, Cambridge preprint.
[17] T. H. R. Skyrme, Proc. R. Soc. London, Ser. A **247**, 260 (1958).
[18] T. H. R. Skyrme, Proc. R. Soc. London, Ser. A **262**, 237 (1961).
[19] R. F. Streater and I. Wilde, Nucl. Phys. B **24**, 561 (1970).
[20] A. Luther and K. D. Schotte, Nucl. Phys. B **242**, 407 (1984).
[21] P. Garbaczewski, *Classical and Quantum Field Theory of Exactly Soluble Nonlinear Systems* (World Scientific, Singapore, 1985).
[22] J. M. Levy-Leblond, Phys. Lett. A **26**, 540 (1968).
[23] H. R. Post, Proc. Phys. Soc. London, Ser. A **66**, 649 (1953).
[24] F. Calogero, J. Math. Phys. **12**, 419 (1971); A. Perelomov, Teor. Mat. Fiz. **6**, 364 (1971).

ζ-function method and the evaluation of fermion currents

R. E. Gamboa Saraví, M. A. Muschietti, F. A. Schaposnik, and J. E. Solomin
Facultad de Ciencias Exactas, Departamento de Física, Universidad Nacional de La Plata, C.C. No. 67, 1900 La Plata, Argentina

(Received 23 November 1984; accepted for publication 15 February 1985)

Using the ζ-function method, a prescription for the evaluation of fermion currents in the presence of background fields is given. The method preserves gauge invariance at each step of the computation and yields to a finite answer showing the relevant physical properties for arbitrary background configurations. Examples for $n = 2,3$ dimensions are worked out, emphasizing the connection between preservation of gauge invariance and violation of other symmetries (chiral symmetry for $n = 2$, parity for $n = 3$).

I. INTRODUCTION

Evaluation of composite current operators (CCO) is of great interest for the analysis of different field theoretical models. As it is well known, the path-integral formulation of quantum field theory provides a natural framework for these calculations, in particular in connection with its topological aspects. Not only can the role of solitons be more easily understood through semiclassical approximations of the path integral, but, in addition, index theorems and other powerful mathematical tools can be put to work very simply.

When gauge theories are concerned, CCO computations are not only faced with divergencies—and hence with the necessity of regularization—but also with the need of preserving gauge invariance in the sense of Schwinger, who developed, in his 1951 classic paper,[1] a regularization procedure, the so-called fictitious-time method.

Schematically, this method consists of a point-splitting regularization of the originally divergent vacuum currents; since this spoils gauge invariance, a compensating phase factor is introduced *ad hoc*, in order to maintain this invariance at every stage of the computation. In some cases, like the exactly soluble QED_2 model[2] or when only constant or static fields in three dimensions are considered,[3–5] the method is very simple. On the other hand, when non-Abelian gauge theories are investigated and arbitrary field configurations considered, it may become more involved.

It is the purpose of this work to propose an alternative method based on techniques developed by Seeley[6] from his definition of complex powers of pseudodifferential operators. Not only is the application of this method very simple, but it is also automatically gauge invariant, thus avoiding, in particular, the introduction of phase factors. Moreover, isolation of relevant contributions to ground-state currents in general cases (i.e., nonconstant, nonstatic field strengths) is possible, as it will become clear from the examples we discuss.

II. PROCEDURES

Consider the generating functional (in Euclidean n-dimensional space) for fermions in the presence of a background field

$$Z[A] = \int \mathscr{D}\bar\psi\, \mathscr{D}\psi \exp\Big(-\int \bar\psi \not{D}(A)\psi\, d^n x\Big) \equiv \det \not{D}(A), \tag{1}$$

with

$$\not{D}(A) = \gamma_\mu(i\,\partial_\mu + A_\mu). \tag{2}$$

Here, $\mathscr{D}\bar\psi\, \mathscr{D}\psi$ is some measure in the space of fermion field configuration, to be defined below. The background field A_μ takes values in the Lie algebra of some gauge group G.

Of course, since the product of eigenvalues of the Dirac operator grows with no bound, Z needs a regularization in order to make sense. Disregarding this fact for a moment, one can try to define the ground-state current in the presence of the background field

$$J_\mu(x) = \langle \bar\psi(x)\gamma_\mu\psi(x)\rangle = -\frac{\delta}{\delta A_\mu(x)}\log Z[A] = -\operatorname{tr}\gamma_\mu G(x,x), \tag{3}$$

where the Green's function $G(x,y)$ satisfies

$$\not{D}_x G(x,y) = \delta^{(n)}(x-y). \tag{4}$$

Expression (3) is also ill defined and it has to be regularized if we want to get a sensible result for J_μ. The method proposed by Schwinger[1,2] makes use of a point splitting in order to deal with the problem of multiplication of distributions at the same point

$$J_\mu(x)_{\rm reg} = \lim_{y\to x} \operatorname{tr}\gamma_\mu G(x,y). \tag{5}$$

Here the limit has to be taken so that in Minkowski space $y \to x$ from the future and from the past. As noted by Schwinger, the appropriate gauge-invariant expression for the current is not (5), but

$$J_\mu(x)_{\rm pt} = \lim_{y\to x} -\operatorname{tr}\gamma_\mu G(x,y)\exp\Big(-i\int_x^y A_\mu(z)\,dz\Big). \tag{6}$$

The phase factor in Eq. (6) has been included in order to restore gauge invariance.

Let us now describe our alternative regularization method. As we stated above, the generating functional defined by Eq. (1) (the determinant of the Dirac operator $\not{D}$) is an ill-defined object (the eigenvalues of $\not{D}$ grow with no bound). This problem can be handled once and for all using the explicitly gauge-invariant ζ-function method[7] (valid for any elliptic operator $\not{D}$, not necessary Hermitian, defined on a compact manifold with no boundary[8]) in order to make sense from Z

$$Z_{\rm reg}[A] = \exp\Big(-\frac{d\zeta}{ds}(\not{D}(A),s)\Big)\Big|_{s=0}. \tag{7}$$

 0022-2488/85/082045-05$02.50

As expression (7) is well defined, the quantity

$$J_\mu(x)_{\rm reg} \equiv -(\delta/\delta A_\mu(x))\log Z_{\rm reg}[A], \tag{8}$$

if the $Z_{\rm reg}$ proves to be differentiable, can be taken as the definition of the regularized ground-state current. As we shall see below, (7) is indeed differentiable and then one can easily obtain, from Eqs. (7) and (8), a regularized version of the last equality in (3). Indeed, let us write for the functional derivative

$$\frac{\delta Z_{\rm reg}}{\delta A_\mu(x)}[A](a) = \frac{d}{dt}(Z_{\rm reg}[A + ta_\mu] - Z_{\rm reg}[A]), \tag{9}$$

where a_μ indicates the direction (in the space of A_μ) in which the functional derivative is taken. Now, from expression (7) and the results about differentiability of the ζ function,[9] we can write

$$\frac{\delta}{\delta A_\mu}\log Z_{\rm reg}[A](a) = \frac{d}{dt}\left(-\frac{d\zeta}{ds}(\not{D}(A + ta_\mu),s)|_{s=0}\right)_{t=0} \tag{10}$$

or

$$\frac{\delta}{\delta A_\mu}\log Z_{\rm reg}[A](a) = \frac{d}{ds}\{s\ {\rm tr}(\not{D}(A)^{-s-1}\gamma_\mu a_\mu)\}_{s=0}, \tag{11}$$

and hence the current $J^a_{\mu_{\rm reg}}$ can be written as

$$J^a_{\mu_{\rm reg}}(x) = -\frac{d}{ds}\left\{s\ {\rm tr}\left[\gamma_\mu t^a \int d^n x \times K_{-s-1}(x,x;\not{D}(A))\right]\right\}\Bigg|_{s=0}, \tag{12}$$

where $K_s(x,x;\not{D})$ is the meromorphic continuation of the evaluation on $x = y$ of the kernel $K_s(x,y;\not{D})$ associated to the operator $\not{D}^s$. This meromorphic extension is a continuous function, even for $x = y$ if $\operatorname{Re} s < -n$ and it has, at the most, a simple pole at $s = -1$. Note that $K_{-1}(x,y;\not{D}) = G(x,y)$ for $x \neq y$. The diagonal $(x = y)$ is precisely the singular support of $G(x,y)$. Whenever m, the order of the operator $\not{D}$, satisfies $m > n$, its Green's function will be continuous everywhere and coincide in the diagonal with $K_{-1}(x,y;\not{D})$. In Eq. (12) the t^a are the generators of the Lie algebra of G and tr denotes both γ matrices' and G matrices' traces.

We shall now state the main result in this work as a proposition which provides the finite part of $K_s(x,x;\not{D})$ as an analytic function of s through the relation

$$\text{Finite part } K_s(x,x;\not{D})|_{s=-1} = \frac{d}{ds}(K_{s-1}(x,x;\not{D}))|_{s=0}. \tag{13}$$

It is just this expression that appears in Eq. (12).

It remains then to give a method to evaluate (13) and then use relation (12) in order to get the regularized current. It is precisely the following proposition which can be used to this end.

Proposition: Let D be an elliptic invertible operator of order m on an n-dimensional compact manifold M without boundary, with a ray of minimal growth[6] and $n \geqslant m$. Then the following identity holds:

$$\frac{d}{ds}(sK_{s-1}(x,x;D))|_{s=0} = -\frac{1}{m}\int_{|\xi|=1}\left(\frac{d}{ds}C_{-m+n}(x,\xi;s)\right)\Bigg|_{s=-1}\frac{d\Omega_\xi}{(2\pi)^n} + \left\{G(x,y) - \sum_{j=0}^{n-m-1}\int\frac{d\xi}{(2\pi)^n}[C_j(x,\xi/|\xi|;-1) \times |\xi|^{-m-j}]e^{i\xi(x-y)} - \not{h}_0(x,x-y) + M(x)(\log|x-y| + \mathbb{C}(x-y))\right\}\Bigg|_{x=y}. \tag{14}$$

Here

$$C_j(x,\xi;s) = \frac{i}{2\pi}\int_\Gamma \lambda^s b_{-m-j}(x,\xi;\lambda)d\lambda, \tag{15}$$

with Γ being a curve in the complex plane, beginning at ∞, passing along the ray of minimal growth to a small circle about the origin, then clockwise along the circle and back to ∞ along the ray. In (15), b_j are the Seeley's coefficients[6] of D, $G(x,y)$ is its Green's function, and $\not{h}_0(x,z)$ is a homogeneous function of degree zero defined as follows:

$$\not{h}_0(x,z) = \int\frac{d\xi}{(2\pi)^n}\,{\rm P.V.}[C_{n-m}(x,\xi/|\xi|;-1) - M(x)]|\xi|^{-n}e^{i\xi z}. \tag{16}$$

Finally,

$$M(x) = \frac{1}{\omega_n}\int_{|\xi|=1}C_{n-m}(x,\xi;-1)d\Omega_\xi, \tag{17}$$

$$\omega_n\left(\log\frac{1}{|z|} - \mathbb{C}(z)\right) = \int\frac{d\xi}{(2\pi)^n}\chi|\xi|^{-n}e^{-i\xi z}, \tag{18}$$

with χ the characteristic function of $|\xi| \geqslant 1$ and $\mathbb{C}(z)$ regular in a neighborhood of $z = 0$. All Fourier transforms are taken in the sense of distributions and P.V. means principal value.

One should not be discouraged by the apparent complexity of relation (14). One can think of it as giving the finite part of $K_{-1}(x,y;D)$ at the diagonal in terms of a local part depending on the Seeley's coefficients, which can be evaluated very easily for arbitrary D, plus a contribution arising from the regular part of the Green's function (at $x = y$) which, given a particular D, one will be able to or not be able to solve closely. As we shall see below, Eqs. (12)–(14) can be used to obtain currents with great simplicity, and allow for relevant analysis for the case of general field-strength configurations. The term between brackets in (14) is the one where the nonregular part of $G(x,y)$ on the diagonal is subtracted. Concerning the first term, it is generated by the regularization method and it adds to the eventually nonzero regular part of $G(x,x)$ to give the final result. As we shall see in our $n = 3$ example, even without computing the regular part of $G(x,x)$ for arbitrary A_μ, we will manage to extract the relevant physical properties (in particular, parity violation) from (14).

The proof of the Proposition is given as the Appendix. In what follows, examples on how relation (14) can be used are discussed.[10] First consider the very simple case of two-dimensional quantum electrodynamics (discussed by Schwinger in Ref. 2) in order to shed some light on the different contributions to (14).

As we stated above, within the Schwinger approach, one starts from Eq. (6) and the knowledge (in this particular case) of the complete Green's function for an arbitrary background field A_μ

$$G(x,y) = (-i/2\pi)\exp[\gamma_5(\phi(x)-\phi(y)) + i(\eta(x)-\eta(y))]\gamma_\mu[(x_\mu - y_\mu)/|x-y|^2], \tag{19}$$

where ϕ and η are related to A_μ through the equation

$$A_\mu = -\epsilon_{\mu\nu}\,\partial_\nu\,\phi + \partial_\mu\eta. \tag{20}$$

Inserting relations (19) and (20) in Eq. (6) one gets the gauge-invariant expression

$$J_{\mu_{\rm pt}}(x) = -(1/\pi)[A_\mu - \partial_\mu \Delta^{-1}\,\partial_\nu A_\nu] \tag{21}$$

(with pt we indicate that the current has been obtained using the proper-time method).

It is interesting to note that the first term in (21) arises from the phase factor while the second one comes from the η exponential in (19). The relation between J_μ and the axial anomaly[11] can be understood from (21) if one notes that the (regularized) fermion determinant in $n=2$ dimensions can be computed as the (anomalous) Jacobian arising from a "decoupling chiral change" of the fermionic variables

$$\psi = e^{\gamma_5\phi + i\eta}\chi, \quad \bar\psi = \bar\chi e^{\gamma_5\phi - i\eta}. \tag{22}$$

One gets[12]

$$\log\frac{\det \not{D}}{\det i\not{\partial}} = \log J = -\frac{1}{2\pi}\int A_\mu[\delta_{\mu\nu} - \partial_\mu\Delta^{-1}\partial_\nu]A_\nu\, d^2x, \tag{23}$$

and (21) can be derived from (23) by simple differentiation. One can easily see that due to the presence of the first term in Eq. (21), the Ward identity for γ_5 transformations is anomalous since in two dimensions,

$$J^5_\mu = i\epsilon_{\mu\nu}J_\nu \tag{24}$$

and hence

$$\partial_\mu J^5_\mu = -i\epsilon_{\mu\nu}\,\partial_\mu A_\nu. \tag{25}$$

Let us now reobtain this result using the method summarized in the Proposition. One has to compute some terms in the asymptotic expansion of the symbol of $\not{D}^{-s}$ with $s=1$. In particular, for the $n=2$, $m=1$ case we are dealing with, only $(dC_1/ds)(x,\xi;s)$ is needed since all other terms containing C_j can be easily seen to vanish for $s=0$. Using the identity

$$\int_{|\xi|=1}\frac{dC_1}{ds}(x,\xi;s)\Big|_{s=-1}\frac{d\Omega_\xi}{(2\pi)^2} = \frac{i}{2\pi}\int_\Gamma\frac{d\lambda}{\lambda}\log\lambda\int_{|\xi|=1}b_{-2}(x,\xi;\lambda)\frac{d\Omega_\xi}{(2\pi)^2} \equiv \mathbb{C}_1, \tag{26}$$

our problem reduces to the evaluation of the Seeley's coefficient b_{-2}, which can be easily computed (see, for example, Ref. 8)

$$b_{-2}(x,\xi;\lambda) = [1/(\lambda^2-|\xi|^2)^2][2\lambda\xi_\mu A_\mu - (\lambda^2-|\xi|^2)\not{A} - 2\xi_\mu A_\mu\not{\xi}]. \tag{27}$$

The contribution from $\mathbb{C}_1$ to $J_{\mu_{\rm reg}}$ can be computed using Eqs. (26), (27), and (14)

$$\operatorname{tr}\gamma_\mu\mathbb{C}_1 = +(1/2\pi)A_\mu. \tag{28}$$

From (17) one can show that in this case $M(x)=0$. Concerning $\not{h}_0(x,z)$ one gets

$$\not{h}_0(x,z) = -(1/4\pi)(\not{A} - 2A_\mu z_\mu(\not{z}/z^2)), \quad z = x-y, \tag{29}$$

and hence

$$(G(x,y) - \not{h}_0(x,x-y))|_{x=y} = \frac{1}{2\pi}\gamma_\mu\epsilon_{\mu\nu}\,\partial_\nu\phi + \frac{1}{4\pi}\not{A}. \tag{30}$$

Collecting all these results, one finally gets

$$J_{\mu_{\rm reg}}(x) = -(1/\pi)\epsilon_{\mu\nu}\,\partial_\nu\phi, \tag{31}$$

which obviously coincides with (21) if one uses the relation $\eta = \Delta^{-1}\,\partial_\nu A_\nu$. Note that in this last evaluation of J_μ neither a phase factor nor a symmetric limit has to be taken in order to get a finite, gauge-invariant result.

As a second example, let us study the case of fermions coupled to SU(N) gauge fields in $n=3$ dimensions. This theory has been recently studied in connection with the spontaneous breakdown of a space-time symmetry[3–5]: the computation of J_μ for constant or static field strengths using Schwinger's method has shown a parity-violating contribution to the current (here parity means coordinate reflection invariance).

As we shall see, the use of the regularization method we propose shows that parity is violated for arbitrary (that is, not necessarily constant or static) field strengths. Indeed, we show for general A_μ that

$$J^a_{\mu_{\rm reg}}(x) = \pm(i/16\pi)\epsilon_{\mu\nu\rho}F^a_{\nu\rho} + \cdots, \tag{32}$$

where we indicate with dots normal parity terms. We have chosen the Pauli matrices as the 2×2 Dirac matrices and, as usual, we work in Euclidean space. The A_μ takes values in the Lie algebra of SU(N). (Concerning the $\pm$ sign in J_μ, see the discussion below.)

It will be clear from the explicit calculations that the parity-violating term is a product of the (gauge-invariant) regularization method exactly as the preservation of gauge invariance was at the origin of the breakdown of chiral invariances in the $n=2$ case. That is, preservation of gauge invariance introduces, via regularization, violation of other symmetries (like chiral or parity symmetries). In both cases this violation appears as a topological term. In the present $n=3$ case, integration of (32) gives a topological mass term (Chern–Simons secondary invariant) for the effective action. Even if absent at the classical level, the necessity of regularization induced this term, first studied by Jackiw and Templeton[13] (see also Refs. 3, 4, 14, and 15).

Equation (32) is obtained from relations (12)–(14) following the same steps described for the $n=2$ example. Indeed, one can easily compute the contribution to J_μ from $\mathbb{C}_2$, defined as

$$\mathbb{C}_2 = \int_{|\xi|=1}\frac{d}{ds}C_2(x,\xi;s)\Big|_{s=-1}\frac{d\Omega_\xi}{(2\pi)^3} = \frac{i}{2\pi}\int_\Gamma\frac{d\lambda}{\lambda}\log\lambda\int_{|\xi|=1}b_{-3}(x,\xi;\lambda)\frac{d\Omega_\xi}{(2\pi)^3}, \tag{33}$$

where Seeley's coefficient b_{-3} is given by[8]

$$b_{-3}(x,\xi;\lambda) = \frac{i}{(\lambda^2-|\xi|^2)^3}\{2\not{\xi}\xi_\mu\,\partial_\mu\not{A}\not{\xi} - 2\lambda\not{\xi}\xi_\mu\,\partial_\mu\not{A} - 2\lambda\xi_\mu\,\partial_\mu\not{A}\not{\xi} + 2\lambda^2\xi_\mu\,\partial_\mu\not{A}$$

$$+(\lambda^2-|\xi|^2)\partial\!\!\!/ A\!\!\!/ \xi\!\!\!/ - \lambda(\lambda^2-|\xi|^2)\partial\!\!\!/ A\!\!\!/ \tag{34}$$
$$- i\xi\!\!\!/ A\!\!\!/ \xi\!\!\!/ A\!\!\!/ \xi\!\!\!/ + i\lambda(\xi\!\!\!/ A\!\!\!/ \xi\!\!\!/ A\!\!\!/ + \xi\!\!\!/ A\!\!\!/ A\!\!\!/ \xi\!\!\!/ + A\!\!\!/ \xi\!\!\!/ A\!\!\!/ \xi\!\!\!/)$$
$$- i\lambda^2(\xi\!\!\!/ A\!\!\!/ A\!\!\!/ + A\!\!\!/ \xi\!\!\!/ A\!\!\!/ + A\!\!\!/ A\!\!\!/ \xi\!\!\!/) + i\lambda^3 A\!\!\!/ A\!\!\!/ \}.$$

One finally gets

$$\mathbb{C}_2 = -(1/8\pi)(\partial\!\!\!/ A\!\!\!/ - iA\!\!\!/ A\!\!\!/) + (1/16\pi)(\partial_\mu A_\mu - 2iA_\mu A_\mu). \tag{35}$$

Note that in the computation of $\mathbb{C}_2$, *one does not need the explicit form of* A_μ. That is the reason why our result will be valid for arbitrary A_μ. Concerning $h_0(x,z)$ it is interesting to stress that, as $h_0(x,tz) = (\operatorname{sgn} t)^n h_0(x,z)$, no regular contribution to it can arise whenever n is odd. In the present $n=3$ case it reads

$$h_0(x,z) = 2\pi^2[-\tfrac{1}{2}\partial_\alpha A_\mu(z_\alpha z_\mu/|z|^3)z\!\!\!/ + (1/4|z|)(z_\alpha\gamma_\mu + z_\mu\gamma_\alpha + z\!\!\!/\delta_{\alpha\mu})\partial_\alpha A_\mu - (z_\alpha/|z|)\partial_\alpha A\!\!\!/\,] + A_\nu A_\alpha i\pi^2 \times [z\!\!\!/(z_\nu z_\alpha/|z|^3) + i\epsilon_{\nu\alpha\beta}(z_\beta/|z|)] + \pi^2\,\partial\!\!\!/ A\!\!\!/\,(z\!\!\!//|z|). \tag{36}$$

Together with the terms involving C_0 and C_1 in the right-hand side (rhs) of (14), $h_0(x,z)$ cancels the nonregular part of $G(x,y)$ at the diagonal. Of course, in order to give the complete expression for J_μ one has to compute the contribution of the regular part of $G(x,x)$. This would in principle require the knowledge of $G(x,y)$ for arbitrary $A_\mu(x)$. However, no parity-violating term can arise from this contribution since regular terms, not induced by the regularization method, cannot break an invariance of the theory. Using this fact, it is easy to obtain (32) from Eqs. (33), (12), and (14).

As we stressed above, our result is valid for arbitrary non-Abelian $F_{\mu\nu}$. It confirms the results obtained in Refs. 3 and 4 for the constant and static $F_{\mu\nu}$ cases. One should note that in these last works, there is a sign ambiguity in J_μ as a spur of the regularization procedure. In our approach, this ambiguity can be traced back to the choice of the Γ curve in the λ integrations [see Eq. (15)]. These integrations are of the type

$$I_\Gamma = \int_\Gamma d\lambda \frac{\lambda^q}{(\lambda^2-1)^p}\log\lambda, \tag{37}$$

where q depends on the order of the operator and the space-time dimensions and Γ is taken, for example, in the upper half-plane. Had we chosen a curve Γ in the negative half-plane the result should read

$$I_{\bar\Gamma} = (-1)^{q+1}I_\Gamma. \tag{38}$$

For the $n=2$ case, $q=1$ and there is no ambiguity. On the other hand, for $n=3$ all integrals correspond to $q=0$ and hence there is a sign ambiguity due to the arbitrariness of the Γ-curve choice.

Throughout this work we have treated $D\!\!\!\!/$ as an invertible operator, an assumption which is not in general satisfied. When zero modes are present, one proceeds as follows: one constructs the invertible operator $D\!\!\!\!/ + im$, computes J_μ, and then makes m go to zero. Since Seeley's coefficients are polynomials in m, the (finite) limit coincides with the result obtained with $m=0$ from the beginning.[16]

III. CONCLUSIONS

In summary, we have developed a regularization method based on Seeley's technique for complex powers of pseudodifferential operators. In our approach, the generating functional Z (the fermion determinant) is regularized once and for all using the ζ-function method and this guarantees, as it has been rigorously proven using the ζ-function differentiability properties, that the current obtained from Z by functional differentiation are finite.

In contrast with the point-splitting technique, our prescription is, *per se*, gauge invariant and it does not need the addition of a phase factor *à la* Schwinger. One can also clearly understand the relation between axial anomalies and parity violations[3,4,15] in our approach: the ζ function provides a definition of the fermionic generating functional (that is, a definition of the path-integral fermionic measure). From this finite generating functional, ground-state currents can be computed; the preservation of gauge invariance then forces violation of chiral symmetry in even dimensions or parity in odd ones. These violations appear as topological terms in current conservation equations (n even) or currents (n odd).

Concerning explicit computations, the evaluation of J_μ (as well as $\partial_\mu J_\mu^5$ for n even[8]) reduces to calculating a few Seeley's coefficients. Although in some cases this can become tedious, it is important to notice that after calculations one can analyze fermion currents in arbitrary background fields.

ACKNOWLEDGMENTS

R. E. Gamboa Saraví and F. A. Schaposnik thank CIC, Buenos Aires, Argentina, for financial support. M. A. Muschietti and J. E. Solomin thank Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina, for financial support.

APPENDIX: PROOF OF THE PROPOSITION

Since D^s is a pseudodifferential operator of order ms, $K_s(x,y;D)$ can be written, for $\mathrm{Re}(-s)$ sufficiently large,[6] as

$$K_s(x,y;D) = \sum_{j=0}^{n-m}\int\frac{d\xi}{(2\pi)^n}C_j(x,\xi;s)e^{i(x-y)\xi} + K(x,y;R(s)), \tag{A1}$$

where the C_j are the terms of an asymptotic expansion of the symbol of D^s and $K(x,y;R(s))$ is the kernel of an operator $R(s)$, regularizing of order $ms-(n-m+1)$. Moreover, $R(s)$ is an analytic function of s in a region containing $s=-1$, taking values in the space of continuous linear operators from the Sobolev space $H^{-n/2-\epsilon}(M)$ into $H^{n/2+\epsilon}(M)$, $\epsilon>0$. From this, we can see that

$$\lim_{y\to x}\lim_{s\to -1} K(x,y;R(s)) = \lim_{s\to -1}\lim_{y\to x}K(x,y;R(s)). \tag{A2}$$

Note that if $R(s):H^{-n/2-\epsilon}(M)\to H^{n/2+\epsilon}(M)$, $K(x,y;R(s))$ is continuous even for $x=y$.

Now for $x\neq y$, $G(x,y) = \lim_{s\to -1} K_s(x,y;D)$. Then

$$K(x,y;R(-1)) = G(x,y) - \sum_{j=0}^{n-m}\int\frac{d\xi}{(2\pi)^n}C_j(x,\xi;-1)e^{i\xi(x-y)}, \tag{A3}$$

even for $x=y$. Now, from (A3) we have the following identity:

$$\lim_{y\to x}\left[G(x,y)-\sum_{j=0}^{n-m}\int\frac{d\xi}{(2\pi)^n}C_j(x,\xi;-1)e^{i\xi(x-y)}\right]$$
$$=\lim_{y\to x}K(x,y;R(-1))=\lim_{s\to -1}K(x,x;R(s))$$
$$=\lim_{s\to -1}\left[K_s(x,x;D)-\sum_{j=0}^{n-m}\int\frac{d\xi}{(2\pi)^n}C_j(x,\xi;s)\right]. \qquad \text{(A4)}$$

If we now split the ξ integration and use the fact that $C_j(x,\xi;s)$ is a homogeneous function of degree $ms-j$ in ξ for $|\xi|\geqslant 1$ to write the identity

$$\int_{|\xi|>1}C_j(x,\xi;s)\frac{d\xi}{(2\pi)^n}$$
$$=\frac{-1}{ms-j+n}\int_{|\xi|=1}C_j(x,\xi;s)\frac{d\xi}{(2\pi)^n}, \qquad \text{(A5)}$$

we have

$$\operatorname*{im}_{y\to x}\left[G(x,y)-\sum_{j=0}^{n-m}\int\frac{d\xi}{(2\pi)^n}C_j(x,\xi;-1)e^{i\xi(x-y)}\right]$$
$$=\lim_{s\to -1}\left\{K_s(x,x;D)-\left(\frac{-1/m}{(s+1)}\int_{|\xi|=1}C_{n-m}(x,\xi;s)\frac{d\xi}{(2\pi)^3}\right.\right.$$
$$\left.+\int_{|\xi|<1}C_{n-m}(x,\xi;s)d\xi\right)-\sum_{j=0}^{n-m-1}\frac{1}{(2\pi)^n}\left(\frac{-1}{ms+n-j}\right.$$
$$\left.\left.\times\int_{|\xi|=1}C_j(x,\xi;s)d\xi+\int_{|\xi|<1}C_j(x,\xi;s)d\xi\right)\right\}. \qquad \text{(A6)}$$

Then, calling

$$\widetilde{C}_j(x,\xi)=|\xi|^{-m-j}C_j(x,\xi/|\xi|;-1),$$

Eq. (A6) can be rewritten as

$$\lim_{y\to x}\left[G(x,y)-\sum_{j=0}^{n-m-1}\int\frac{d\xi}{(2\pi)^n}\widetilde{C}_j(x,\xi)e^{i\xi(x-y)}\right.$$
$$\left.-\int\frac{d\xi}{(2\pi)^n}C_{n-m}(x,\xi;-1)e^{i\xi(x-y)}\right]$$
$$=\lim_{s\to -1}\left[K_s(x,x;D)-\frac{-1/m}{s+1}\int_{|\xi|=1}C_{n-m}(x,\xi;s)\frac{d\xi}{(2\pi)^n}\right.$$
$$\left.-\int_{|\xi|<1}C_{n-m}(x,\xi;s)\frac{d\xi}{(2\pi)^n}\right]. \qquad \text{(A7)}$$

Taking the expression of $C_{n-m}(x,\xi,s)$ in powers of $(s+1)$ we can write (A7) in the form

$$\lim_{s\to -1}\left[K_s(x,x;D)-\frac{1}{s+1}\left(-\frac{1}{m}\right)\int_{|\xi|=1}C_{n-m}(x,\xi;-1)\frac{d\xi}{(2\pi)^n}\right]$$
$$=-\frac{1}{m}\int_{|\xi|=1}\frac{d\xi}{(2\pi)^n}\frac{d}{ds}C_{n-m}(x,\xi;s)\Big|_{s=-1}$$
$$+\int_{|\xi|<1}C_{n-m}(x,\xi;-1)\frac{d\xi}{(2\pi)^n}+\lim_{y\to x}\left[G(x,y)\right.$$
$$-\sum_{j=0}^{n-m-1}\int\frac{d\xi}{(2\pi)^n}C_j(x,\xi;-1)e^{i\xi(x-y)}-\int\frac{d\xi}{(2\pi)^n}$$
$$\left.\times C_{n-m}(x,\xi;-1)e^{i\xi(x-y)}\right]. \qquad \text{(A8)}$$

Then, using[17]

$$\int_{|\xi|>1}e^{i\xi z}C_{n-m}(x,\xi;-1)\frac{d\xi}{(2\pi)^n}$$
$$=\text{P.V.}\int\frac{d\xi}{(2\pi)^n}e^{i\xi z}|\xi|^{-n}\left[C_{n-m}\left(x,\frac{\xi}{|\xi|};-1\right)-M(x)\chi\right]$$
$$+M(x)[\log|z|+\mathbb{C}(z)+g(z)],$$

with $g(z)$ and $\mathbb{C}(z)$ regular functions and $g(0)=0$, we get Eq. (14).

[1]J. Schwinger, Phys. Rev. **82**, 664 (1951).
[2]J. Schwinger, Phys. Rev. **128**, 2425 (1962).
[3]A. N. Redlich, Phys. Rev. Lett. **52**, 18 (1984) and MIT preprint 1128, 1984.
[4]A. J. Niemi and G. W. Semenoff, Phys. Rev. Lett. **51**, 2077 (1983).
[5]Perturbative calculations can be found in S. Deser, R. Jackiw, and S. Templeton, Ann. Phys. (NY) **140**, 372 (1982).
[6]R. T. Seeley, Am. Math. Soc. Proc. Symp. Pure Math. **10**, 288 (1967).
[7]S. W. Hawking, Commun. Math. Phys. **55**, 133 (1977).
[8]R. E. Gamboa Saraví, M. A. Muschietti, F. A. Schaposnik, and J. E. Solomin, Ann. Phys. (NY) **157**, 360 (1984).
[9]R. E. Gamboa Saraví, M. A. Muschietti, and J. E. Solomin, Commun. Math. Phys. **89**, 363 (1983).
[10]For simplicity we take the manifold M as S^n.
[11]For a summary see, for example, R. Jackiw, in S. Treiman, R. Jackiw, and D. Gross, *Lectures on Current Algebra and its Applications* (Princeton U.P., Princeton, NJ, 1972).
[12]R. Roskies and F. A. Schaposnik, Phys. Rev. D **23**, 558 (1981).
[13]R. Jackiw and S. Templeton, Phys. Rev. D **23**, 2291 (1981).
[14]J. Schonfeld, Nucl. Phys. B **185**, 157 (1981).
[15]R. Jackiw, "comments on nuclear and particle physics," MIT preprint 1140, 1984 (to be published).
[16]R. E. Gamboa Saraví, M. A. Muschietti, and J. E. Solomin, Commun. Math. Phys. **93**, 407 (1984).
[17]A. P. Calderón, *Lecture Notes on Pseudo-differential Operators and Elliptic Boundary Value Problems* (I.A.M., Buenos Aires, 1976).

Scalar field propagators in anti-de Sitter space-time

C. Dullemond and E. van Beveren
Institute for Theoretical Physics, University of Nijmegen, 6525 ED Nijmegen, The Netherlands

(Received 27 December 1984; accepted for publication 5 April 1985)

Simple expressions are found for the retarded, advanced, and Feynman propagators, as well as several other auxiliary invariant functions, for scalar fields of arbitrary mass in anti-de Sitter space-time.

I. INTRODUCTION

One of the still-unsolved mysteries of present-day high-energy physics is the mechanism nature uses for confining quarks and gluons. The hope is that the rules of QCD are sufficient to prove that colored particles are always clustered into colorless objects like mesons and baryons. It is, however, practical not to wait for such a proof, but to make models in which this confinement has been built in. Bag models[1] have this property, but ordinary potential models in which an indefinitely rising potential has been used[2] also have this property.

Making use of ideas put forward by several authors,[3] we have studied a model in which confinement has been built in by geometrical means.[4] Quarks and gluons move in a microuniverse of anti-de Sitter symmetry, but this world manifests itself to outside observers as a spherical bag. Inside this bag, and in zeroth-order approximation, all colored particles carry out harmonic oscillations with a universal frequency ω equal to c/R, where c is the light velocity and R is the bag radius. This universal frequency can be read off from the average quarkonium level spacing ΔE through $\omega = \Delta E/\hbar$ and gives rise to a universal bag radius for mesons equal to $R = c\hbar/\Delta E$. This turns out to be about 1 fm. For baryons this radius becomes somewhat larger. Absolute confinement within a sphere is now related to the impossibility to surpass light velocity.

The anti-de Sitter space comes about because the QCD Lagrangian for massless quarks and gluons is impervious to a conformal change of metric.[5] By introducing a Higgs-like scalar field, the conformal symmetry of SO(4,2) type becomes manifest. The field can now be used to induce "minimal" spontaneous symmetry breaking towards SO(3,2) or anti-de Sitter symmetry. In the mean time, quarks acquire a nonzero mass through coupling with the scalar field.

If one wants to go beyond zeroth order in studying the behavior of quarks and gluons inside the "harmonic oscillator bag," due account must be given to gluon exchange between quarks, quark–antiquark production and annihilation by gluons, gluon exchange between gluons, etc. Of course these activities have their limitations because of the too-static character of the bag, and good judgement should be the guiding principle. As in flat space, propagators may be useful in practical calculations.

Much work on propagators in maximally symmetric spaces has been done in the past. Early work includes that of Adler.[6] For the anti-de Sitter space homogeneous propagators have been found by Fronsdal for scalar fields[7] and by Fronsdal and Haugen[8] for spinor fields for the massive as well as the massless case. The massless case is of special interest and contains many surprises. It has been studied for arbitrary spin by Fronsdal[9] and Fang and Fronsdal.[10] Conformally invariant propagators for QED in anti-de Sitter geometry have been studied by Binegar, Fronsdal, and Heidenreich.[11]

The purpose of the present paper is to derive simple expressions for SO(3,2) symmetric massive scalar propagators of the homogeneous as well as the inhomogeneous kind using solely configuration space methods.[12] The simplicity of these methods is worth exposing. The inhomogeneous propagators include Feynman propagators which can be interpreted as the vacuum expectation values of time-ordered field products. Since "time" is here a many-valued noninvariant function in five space, a discussion of the invariance properties of these propagators is necessary.

Two further remarks must be made. In the first place, the meaning of "spacelike" and "timelike" separation is not *a priori* clear and must be discussed in connection with the non-simply-connected character of the SO(3,2) invariant manifold. This makes the introduction of a covering space necessary. A good discussion of this is given by Fronsdal.[7] Second, implicit boundary conditions at infinity have to be imposed in order to make the propagators unique. This will in general be possible by requiring the propagators to approach zero "sufficiently" fast when a certain invariant quantity approaches minus infinity.[13] For the massless case, extra conditions are needed to prevent "information leaks" at infinity.[16]

A demonstration of how the configuration space method works in Minkowski space-time is given in Sec. II, while in Sec. III the method is applied to anti-de Sitter space-time. As distinct from the Minkowski case, to write down propagators for an arbitrary reference point is not completely trivial and needs some attention. Section IV is devoted to this point. Then, in order to exhibit the bag structure and to make the harmonic oscillations manifest, in Sec. V a transformation to special coordinates is carried out. These coordinates are obtained by a process of central projection[4] and act as natural coordinates for outside observers who can look inside the bag by means of electromagnetic waves interacting with the fractionally charged quarks. "Infinity" has thus been projected on the bag surface and the boundary conditions at infinity have been transformed into boundary conditions on that surface.

II. PROPAGATORS IN MINKOWSKI SPACE-TIME

In this section a review is given of the space-time method leading to the well-known invariant solutions of the Klein–Gordon equation with and without a δ-source term.

 0022-2488/85/082050-09$02.50

In many textbooks the results are only mentioned, completely or incompletely, either in the main text or in an appendix.[17] Most derivations make use of momentum space.

Let us start with the massless homogeneous equation in x^μ space with $\eta_{\mu\nu} = \mathrm{diag}\,(-1, -1, -1, 1)$

$$\Box\, \Psi(\lambda) = 0, \tag{2.1}$$

where

$$\Box = \partial_\mu\, \partial^\mu = -\Delta + \frac{\partial^2}{\partial x^{4^2}} \tag{2.2}$$

and

$$\lambda = x_\mu x^\mu. \tag{2.3}$$

The solution which is of interest for our purposes is the one which drops off to zero at infinity

$$\Psi(\lambda) = 1/\lambda, \quad \lambda \neq 0. \tag{2.4}$$

The singularity at $\lambda = 0$ can be circumvented by making the following replacement:

$$x^4 \to x^4 \pm i\epsilon/2, \quad \epsilon > 0. \tag{2.5}$$

Under this replacement, where ϵ is not necessarily infinitesimally small, the operator $\Box$ remains unchanged, but λ undergoes a change

$$\lambda \to \lambda \pm i\epsilon x^4 - \epsilon^2/4. \tag{2.6}$$

Thus we conclude that, for ϵ not necessarily infinitesimal,

$$\Psi_\pm(\lambda) = 1/(\lambda \pm i\epsilon x^4 - \epsilon^2/4) \tag{2.7}$$

satisfies the equation

$$\Box\, \Psi_\pm(\lambda) = 0. \tag{2.8}$$

In the limit of small ϵ, by taking the real and imaginary parts of (2.8), we obtain

$$\Box\, P(1/\lambda) = 0 \tag{2.9}$$

and

$$\Box\, \epsilon(x^4)\delta(\lambda) = 0, \tag{2.10}$$

where

$$\epsilon(x^4) = \pm 1, \quad x^4 \gtrless 0.$$

So, while $\Psi_\pm(\lambda)$ is not invariant under Lorentz transformations around the origin $x^\mu = 0$, we have now obtained invariant solutions $P(1/\lambda)$ and $\epsilon(x^4)\delta(\lambda)$, the first being even under time reversal, the second odd. The latter can be used to obtain the general solution of the homogeneous equation when the function and its time derivative are given on a suitable spacelike surface, for example, $x^4 = 0$. Then a general reference point has to be chosen; x^μ must be replaced by $x^\mu - x_0^\mu$. In order to carry out this program the time derivative of $\Psi_\pm(\lambda)$ as given by (2.7) must be determined.

We have for the space integral over the time derivative

$$\int \left[\frac{\partial}{\partial x^4} \frac{1}{\lambda \pm i\epsilon x^4 - \epsilon^2/4}\right]\bigg|_{x^4=0} d^3x$$
$$= \mp i\epsilon \int_0^\infty \frac{4\pi r^2\, dr}{(r^2 + \epsilon^2/4)^2} \underset{\epsilon\to 0}{\to} \mp 2\pi^2 i. \tag{2.11}$$

Since $P(1/\lambda)$ is an even function and $\epsilon(x^4)\delta(\lambda) = 0$ when $x^4 = 0$ and $\mathbf{x} \neq 0$ we have, therefore, using Gauss' theorem

$$\left[\frac{\partial}{\partial x^4} \frac{1}{\lambda \pm i\epsilon x^4 - \epsilon^2/4}\right]\bigg|_{x^4=0} \underset{\epsilon\to 0}{\to} \mp 2\pi^2 i \delta^{(3)}(\mathbf{x}). \tag{2.12}$$

Thus, if we define

$$G(x^\mu) = (1/2\pi)\epsilon(x^4)\delta(\lambda) \tag{2.13}$$

and

$$\overline{G}(x^\mu) = (1/2\pi^2)P(1/\lambda), \tag{2.14}$$

we have

$$\Box\, G = \Box\, \overline{G} = 0 \tag{2.15}$$

and

$$\left[\frac{\partial}{\partial x^4} G\right]\bigg|_{x^4=0} = \delta^{(3)}(\mathbf{x}), \tag{2.16}$$

together with

$$\left[\frac{\partial}{\partial x^4} \overline{G}\right]\bigg|_{x^4=0} = 0. \tag{2.17}$$

One sometimes encounters "positive frequency" and "negative frequency" parts. These are defined by

$$G_\pm = \frac{1}{2}(\overline{G} \pm iG) = \frac{1}{4\pi^2} \frac{1}{\lambda \mp i\epsilon x^4 - \epsilon^2/4}, \tag{2.18}$$

in the limit of small ϵ.

Next we consider the inhomogeneous equation. Since $\epsilon(x^4)\,\delta(\lambda)$ is a solution of the homogeneous equation (2.1) it follows that $\delta(\lambda)$ is a solution as long as $x^\mu \neq 0$. We find therefore that

$$\lim_{\epsilon\to 0} [1/(\lambda \pm i\epsilon)] \tag{2.19}$$

is a solution of the homogeneous wave equation outside the origin. If we now integrate time derivatives over the two four-planes $x^4 = \pm a$ $(a > 0)$ we find for small ϵ and small a

$$\int \left[\frac{\partial}{\partial x^4} \frac{1}{\lambda - i\epsilon}\right]\bigg|_{x^4=\pm a} d^3x = \pm 2\pi^2 i. \tag{2.20}$$

With the help of Gauss' theorem we now find

$$\Box\, [1/(\lambda \pm i\epsilon)] = \mp 4\pi^2 i \delta^{(4)}(x^\rho). \tag{2.21}$$

From (2.18) and (2.21) it follows that

$$\Box\, [\theta(x^4)/2\pi]\delta(\lambda) = \delta^{(4)}(x^\rho). \tag{2.22}$$

This leads to the retarded and advanced Green's functions

$$G_{\text{ret}} = (1/2\pi)\theta(x^4)\delta(\lambda) \tag{2.23}$$

and

$$G_{\text{adv}} = (1/2\pi)\theta(-x^4)\delta(\lambda), \tag{2.24}$$

which satisfy the inhomogeneous equations

$$\Box\, G_{\substack{\text{ret}\\ \text{adv}}} = \delta^{(4)}(x^\rho). \tag{2.25}$$

The Feynman propagator is defined by

$$G_F = (1/4\pi^2 i)/(\lambda - i\epsilon), \tag{2.26}$$

in the limit of small ϵ. It is an even function of x^4, an analytic function of λ, and it satisfies the equation

$$\Box\, G_F = \delta^{(4)}(x^\rho). \tag{2.27}$$

This concludes the massless case.

If the mass is nonzero we have to find invariant solutions of the Klein–Gordon equation

$$(\Box + m^2)\Psi(\lambda) = 0, \quad m \neq 0, \tag{2.28}$$

where it is now appropriate to define λ by

$$\lambda = m^2 x_\mu x^\mu. \tag{2.29}$$

Let us first consider the cases $\lambda > 0$ and $\lambda < 0$ separately. When $\lambda > 0$ we can introduce the variable $\eta = \sqrt{\lambda}$. Then Eq. (2.28) can be written in the form

$$\left(\frac{d^2}{d\eta^2} + \frac{3}{\eta}\frac{d}{d\eta} + 1\right)\Psi(\eta^2) = 0. \tag{2.30}$$

From the identity

$$\left(\frac{d^2}{d\eta^2} + \frac{3}{\eta}\frac{d}{d\eta} + 1\right)\frac{1}{\eta} = \frac{1}{\eta}\left[\frac{d^2}{d\eta^2} + \frac{1}{\eta}\frac{d}{d\eta} + \left(1 - \frac{1}{\eta^2}\right)\right], \tag{2.31}$$

we then find that if $U_1(\eta)$ is a solution of the Bessel equation with index 1, then $U_1(\eta)/\eta$ is a solution of Eq. (2.30) and, written in terms of λ, of Eq. (2.28).

When $\lambda < 0$ we introduce $\zeta = \sqrt{-\lambda}$. Then we find that if $V_1(\zeta)$ is a solution of the modified Bessel equation with index 1, then $V_1(\zeta)/\zeta$, written in terms of λ, is again a solution of Eq. (2.28). In the latter case we shall be interested in the solution which drops off to zero when $\zeta \to \infty$. This is $1/\zeta$ times the modified Hankel function $K_1(\zeta)$ defined by[18]

$$\begin{aligned} K_1(\zeta) &= \tfrac{1}{2}\pi i e^{(1/2)\pi i} H_1^{(1)}(\zeta e^{(1/2)\pi i}) \\ &= -\tfrac{1}{2}\pi i e^{-(1/2)\pi i} H_1^{(2)}(\zeta e^{-(1/2)\pi i}), \end{aligned} \tag{2.32}$$

with $H_1^{(1)}$ and $H_1^{(2)}$ being the first and second Hankel functions with index 1.

It is clear that the point $\lambda = 0$ causes trouble. Just like in the massless case, we can avoid this by making the replacement (2.5) which we modify as follows:

$$mx^4 \to mx^4 \pm i\epsilon/2m, \quad \epsilon > 0. \tag{2.33}$$

This does not affect the Klein–Gordon operator and we find that if $\Psi(\lambda)$ is a solution of (2.28), then so is $\Psi_\pm(\lambda) = \Psi(\lambda \pm i\epsilon x^4 - \epsilon^2/4m^2)$, with ϵ not necessarily infinitesimal

$$(\Box + m^2)\Psi_\pm(\lambda) = (\Box + m^2)\Psi(\lambda_{\pm\epsilon}) = 0, \tag{2.34}$$

where we have used the notation $\lambda_{\pm\epsilon} = \lambda \pm i\epsilon x^4 - \epsilon^2/4m^2$.

For real x^μ the function $\Psi(\lambda_{\pm\epsilon})$ is regular everywhere. The solution of interest is

$$K_1(\zeta_{\pm\epsilon})/\zeta_{\pm\epsilon}, \tag{2.35}$$

where

$$\zeta_{\pm\epsilon} = (-\lambda_{\pm\epsilon})^{1/2}, \quad \mathrm{Re}\,\zeta_{\pm\epsilon} > 0. \tag{2.36}$$

Now, for $x^4 > 0$, let

$$\eta_{\pm\epsilon} = \zeta_{\pm\epsilon} e^{\pm(1/2)\pi i}. \tag{2.37}$$

Then, if we move λ from negative to positive values along the real axis, we have

$$\lim_{\epsilon \to 0} \eta_{\pm\epsilon} = \eta > 0. \tag{2.38}$$

By continuous change of x^μ, a change of sign of x^4 can only occur when $\lambda < 0$. This does not affect the condition $\mathrm{Re}\,\zeta_{\pm\epsilon} > 0$. If for $x^4 < 0$ we now move λ from negative to positive values along the real axis, we have

$$\lim_{\epsilon \to 0} \eta_{\pm\epsilon} = -\eta < 0. \tag{2.39}$$

Thus, from (2.32) we find that for positive λ, independent of the sign of x^4, $\Psi_\pm(\lambda)$ has the following form:

$$\Psi_\pm(\lambda) = \Psi(\lambda_{\pm\epsilon}) = \mp \tfrac{1}{2}\pi i H_1^\pm(\eta_{\pm\epsilon})/\eta_{\pm\epsilon}, \tag{2.40}$$

where $H_1^+ = H_1^{(1)}$ and $H_1^- = H_1^{(2)}$.

The solutions (2.40) are, explicitly, in the limit $\epsilon \to 0$

$$\psi_\pm(\lambda) = \frac{J_1(\eta)}{2\eta}[\mp \pi i + \ln(\lambda_{\pm\epsilon})] - \frac{1}{\lambda_{\pm\epsilon}} + \chi(\lambda), \tag{2.41}$$

where J_1 is the cylindrical Bessel function with index 1 and $\chi(\lambda)$ is a real analytic function, regular in the entire complex plane. The exact form of $\chi(\lambda)$ is given in any book on Bessel functions.[18,19]

Since $\Psi_\pm(\lambda)$ is an exact solution of the Klein–Gordon equation, the real and imaginary parts are exact solutions themselves. We have in the limit of small ϵ

$$\mathrm{Re}\,\Psi_\pm(\lambda) = [J_1(\eta)/2\eta]\ln|\lambda| - P(1/\lambda) + \chi(\lambda) \tag{2.42}$$

and

$$\mathrm{Im}\,\Psi_\pm(\lambda) = \mp\{[J_1(\eta)/2\eta][\pi \mp \arg(\lambda_{\pm\epsilon})] - \pi\epsilon(x^4)\delta(\lambda)\}. \tag{2.43}$$

The latter solution is of most importance. When $\lambda > 0$ and $x^4 > 0$ we find for the expression between square brackets the value π, since for positive η the argument is supposed to be zero. If now λ is moved along the real axis to negative values, the argument becomes $\pm\pi$ and the expression becomes zero. Now we can move x^4 to negative values. If we move λ back to positive values the change of argument is again $\pm\pi$, so we find

$$\mathrm{Im}\,\Psi_\pm(\lambda) = \mp\{\pi\epsilon(x^4)\theta(\lambda)[J_1(\eta)/2\eta] - \pi\epsilon(x^4)\delta(\lambda)\}. \tag{2.44}$$

We now look for the proper normalization. From the results obtained for $m = 0$ and by considering the pole term we find

$$\left[\frac{\partial}{\partial x^4}\Psi_\pm(\lambda)\right]\bigg|_{x^4=0} = \pm\frac{2\pi^2 i}{m^2}\delta^{(3)}(\mathbf{x}). \tag{2.45}$$

This enables us to define $G^m(x^\mu)$ and $\overline{G}^m(x^\mu)$ together with the positive and negative frequency parts

$$G^m(x^\mu) = (m^2/2\pi)\epsilon(x^4)[\delta(\lambda) - \theta(\lambda)[J_1(\eta)/2\eta]], \tag{2.46}$$

$$\overline{G}^m(x^\mu) = -\frac{m^2}{2\pi^2}\left[\frac{J_1(\eta)}{2\eta}\ln|\lambda| - P\frac{1}{\lambda} + \chi(\lambda)\right], \tag{2.47}$$

and

$$G_\pm^m(x^\mu) = \tfrac{1}{2}(\overline{G}^m \pm iG^m) = (-m^2/4\pi^2)\Psi_\mp(\lambda), \quad \epsilon \to 0. \tag{2.48}$$

Finally we have the inhomogeneous case. Using the same arguments as for $m = 0$ we find that the functions $\Psi(\lambda \pm i\epsilon)$ are solutions of Eq. (2.28) as long as $x^\mu \neq 0$. Using (2.20) we obtain for small ϵ

$$\int\left[\frac{\partial}{\partial x^4}\Psi(\lambda - i\epsilon)\right]\bigg|_{x^4=\pm a} d^3x = \mp\frac{2\pi^2 i}{m^2}, \tag{2.49}$$

which gives

$$(\Box + m^2)\Psi(\lambda - i\epsilon) = (-4\pi^2 i/m^2)\delta^{(4)}(x^\mu), \tag{2.50}$$

and similarly

$$(\Box + m^2)\Psi(\lambda + i\epsilon) = (4\pi^2 i/m^2)\delta^{(4)}(x^\mu). \tag{2.50'}$$

Then, since

$$\mathrm{Im}\,\Psi(\lambda \pm i\epsilon) = \pm\pi[\delta(\lambda) - \theta(\lambda)[J_1(\eta)/2\eta]], \tag{2.51}$$

we find

$$(\Box + m^2)(m^2/4\pi)[\delta(\lambda) - \theta(\lambda)[J_1(\eta)/2\eta]] = \delta^{(4)}(x^\mu). \tag{2.52}$$

If the retarded and advanced Green's functions are defined by

$$G^m_{\substack{\text{ret}\\ \text{adv}}}(x^\mu) = \pm\theta(\pm x^4)G^m(x^\mu) = (m^2/2\pi)\theta(\pm x^4)[\delta(\lambda) - \theta(\lambda)[J_1(\eta)/2\eta]], \tag{2.53}$$

then these invariant functions satisfy the inhomogeneous equation

$$(\Box + m^2)G^m_{\substack{\text{ret}\\ \text{adv}}}(x^\mu) = \delta^{(4)}(x^\mu). \tag{2.54}$$

The Feynman propagator G^m_F satisfies the same inhomogeneous equation, is an analytic function of λ, and an even function of x^4

$$G^m_F(x^\mu) = -\lim_{\epsilon\to 0}(m^2/4\pi^2 i)\Psi(\lambda - i\epsilon), \tag{2.55}$$

with

$$\Psi(\lambda) = K_1(\zeta)/\zeta, \quad \zeta = (-\lambda)^{1/2},$$

$$\operatorname{Re}\zeta \geqslant 0 \text{ and } \eta = \lambda^{1/2}.$$

Although the results of this section are of course well known, the method of derivation is not. The reason for exposing it here is that a generalization to SO(3,2) propagators is straightforward and hardly more complicated.

III. PROPAGATORS IN ANTI-DE SITTER SPACE-TIME

Consider a five-dimensional flat space with coordinates ξ^M $(M = 1,...,5)$ and metric

$$\eta_{MN} = \operatorname{diag}(-1, -1, -1, +1, +1). \tag{3.1}$$

We are specially interested in points lying on the hyperboloid

$$\xi_M\xi^M = -\boldsymbol{\xi}^2 + \xi^{4^2} + \xi^{5^2} = R^2 = \text{const} > 0. \tag{3.2}$$

The linear transformations leaving this hyperboloid invariant and which are continuously connected with the identity form the group of restricted O(3,2) or "orthochronous" SO(3,2) transformations.

The hyperboloid (3.2) is not simply connected. The functions with which we have to deal are in general many valued and for its specification a winding number must be defined. However, such a winding number can be avoided by defining a many-valued reference function which turns out to play the role of time variable. The quality which has been called orthochronous derives its name from it.

Consider the equation

$$\Box_5\Psi = \frac{\partial^2}{\partial\xi^M\,\partial\xi_M}\Psi = \left(\Box_4 + \frac{\partial^2}{\partial\xi^{5^2}}\right)\Psi = 0, \tag{3.3}$$

where

$$\Box_4 = \frac{\partial^2}{\partial\xi^\mu\,\partial\xi_\mu} = \Box = -\Delta + \frac{\partial^2}{\partial\xi^{4^2}}, \quad \mu = 1,...,4. \tag{3.4}$$

For the moment we consider $R > 0$ as defined by (3.2) as a radial variable.

We are interested in those functions which are invariant under the restricted Lorentz transformations leaving the reference point

$$\xi^M_0 = (0, 0, 0, 0, R) \tag{3.5}$$

invariant. These transformations form a stability subgroup of the restricted O(3,2) group, leaving the "angular variable" λ defined by

$$\lambda = \xi_\rho\xi^\rho/R^2 = 1 - \xi^{5^2}/R^2 \tag{3.6}$$

invariant.

The functions of interest to us have the following form:

$$\Psi(R, \lambda, n) = R^m\Phi_n(\lambda, m), \tag{3.7}$$

where n is a winding number. The function Ψ satisfies Eq. (3.3) in the non-simply-connected domain $R > 0$ of five-space. In order to study its properties we introduce the generalized angular momentum operator M_{MN}

$$M_{MN} = i\left(\xi_M\frac{\partial}{\partial\xi^N} - \xi_N\frac{\partial}{\partial\xi^M}\right). \tag{3.8}$$

Then, if the total angular momentum operator M^2 is defined as

$$M^2 = \tfrac{1}{2}M_{MN}M^{MN}, \tag{3.9}$$

we find that Ψ satisfies the equation

$$\Box_5\Psi = \left(\frac{\partial^2}{\partial R^2} + \frac{4}{R}\frac{\partial}{\partial R} - \frac{M^2}{R^2}\right)\Psi = 0, \tag{3.10}$$

as an eigenstate of the M^2 operator. From (3.7) we obtain

$$\Box_5\Phi_n(\lambda, m) = -[m(m+3)/R^2]\Phi_n(\lambda, m) = -(M^2/R^2)\Phi_n(\lambda, m), \tag{3.11}$$

so that

$$[M^2 - m(m+3)]\Phi_n(\lambda, m) = 0. \tag{3.12}$$

In the following we shall suppress the variable m in Φ. The invariant functions in which we are interested have, just like in the Minkowski case, in general a singularity when $\lambda = 0$. In order to avoid this we look for a complex coordinate transformation which not only leaves the operator $\Box_5$, but also R^2 and M^2, invariant. The transformation which does the job is a complex rotation in (ξ^4, ξ^5) space. Let

$$\xi^4 = \sqrt{R^2 + \boldsymbol{\xi}^2}\sin(t/R) \tag{3.13}$$

and

$$\xi^5 = \sqrt{R^2 + \boldsymbol{\xi}^2}\cos(t/R),$$

where $-\infty < t < \infty$.

The parameter t is many valued in ξ^M space and serves as a time variable. We make the replacement

$$t \to t \pm i\epsilon R/2, \quad \epsilon > 0. \tag{3.14}$$

This results into the following replacements for ξ^4 and ξ^5:

$$\xi^4 \to \xi^4\cosh(\epsilon/2) \pm i\xi^5\sinh(\epsilon/2),$$
$$\xi^5 \to \mp i\xi^4\sinh(\epsilon/2) + \xi^5\cosh(\epsilon/2). \tag{3.15}$$

The transformation amounts to a complex rotation generated by M_{45} which commutes with $\Box_5$ and M^2. The quantity λ transforms as follows:

$$\lambda \to \lambda_{\pm\epsilon} = 1 - \frac{1}{R^2}\Big(\xi^5 \cosh\frac{\epsilon}{2} \mp i\xi^4 \sinh\frac{\epsilon}{2}\Big)^2 \quad (3.16)$$

$$\approx \lambda \pm i\epsilon\xi^4\xi^5/R^2 - [(\xi^{5^2} - \xi^{4^2})/4R^2]\,\epsilon^2. \quad (3.17)$$

When for $\xi^2 = 0$ the time variable t is changed from $-\infty$ to $+\infty$ along the real axis, we find that λ describes an ellipse in the complex plane, with the points 0 and 1 as foci. The upper sign in (3.14) corresponds to clockwise and the lower sign to anticlockwise motion (see Fig. 1).

One revolution corresponds to $\Delta t = \pi R$. When $\xi^2 > 0$, but kept fixed, the contour is an ellipse which lies around the ellipse mentioned above. The points 0 and 1 are always enclosed.

Let us next work out Eq. (3.11). We find with suppression of the index n

$$[\Box_5 + m(m+3)/R^2]\Phi(\lambda)$$
$$= (4/R^2)[\lambda(1-\lambda)\Phi''(\lambda) + (2 - \tfrac{5}{2}\lambda)\Phi'(\lambda)$$
$$+ [m(m+3)/4]\Phi(\lambda)] = 0, \quad (3.18)$$

where a prime means differentiation with respect to λ. This equation is of the hypergeometric type[19]

$$\lambda(1-\lambda)\Phi''(\lambda) + [c - (a+b+1)\lambda]\Phi'(\lambda) - ab\Phi(\lambda) = 0, \quad (3.19)$$

so that we can identify

$$a = -m/2, \quad b = (m+3)/2, \quad \text{and} \quad c = 2. \quad (3.20)$$

The regular solution is the hypergeometric function

$$F(a,b;c;\lambda) = F(-m/2, (m+3)/2; 2; \lambda). \quad (3.21)$$

However, of special interest to us is the solution which converges to zero when λ goes to $-\infty$. There are two candidates

$$(-\lambda)^{-a}F(a+1-c, a; a+1-b; \lambda^{-1})$$
$$= (-\lambda)^{m/2}F(-(m+2)/2, -m/2; -m-\tfrac{1}{2}; \lambda^{-1}), \quad (3.22)$$

which satisfies the condition when $m < 0$, $m = -\frac{1}{2}$ excepted, and

$$(-\lambda)^{-b}F(b+1-c, b; b+1-a; \lambda^{-1})$$
$$= (-\lambda)^{-(m+3)/2}F((m+1)/2,$$
$$(m+3)/2; m+\tfrac{5}{2}; \lambda^{-1}), \quad (3.23)$$

which satisfies the condition when $m > -3$, $m = -\frac{5}{2}$ excepted. In order to discriminate, we demand that the function approaches zero faster than $(-\lambda)^{-3/4}$ when $\lambda \to -\infty$.[15] The exceptional points $m = -\frac{1}{2}, -\frac{3}{2}$, and $-\frac{5}{2}$ will then be left out of consideration.

Because m and $-(m+3)$ are interchangeable, we can limit ourselves to $m > -\frac{3}{2}$, which includes the important special case $m = -1$. We shall therefore limit ourselves to Eq. (3.23).

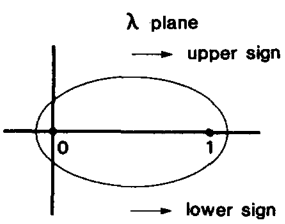


FIG. 1. Contour in complex λ plane corresponding to real t.

The condition on the way functions must approach zero when $\lambda \to -\infty$ amounts to the specification of boundary conditions at infinity in configuration space ξ. This is less trivial than in Minkowski space. In a later stage of development we come back to this point.

The function as defined by the analytic continuation of

$$\Phi(\lambda) = (-\lambda)^{-(m+3)/2}F\Big(\frac{m+1}{2}, \frac{m+3}{2}; m + \frac{5}{2}; \lambda^{-1}\Big),$$
$$|\lambda| > 1, \quad (3.24)$$

is singular in $\lambda = 0$ and $\lambda = 1$, so in order to avoid these singularities we replace λ by $\lambda_{\pm\epsilon}$ as given by (3.16). When t is moved from t_0 to $t_0 + \pi R$ with fixed ξ^2, then $\lambda_{\pm\epsilon}$ makes one complete turn along its ellipse and returns to the same value. The function $\Phi(\lambda)$ then undergoes a change of phase which is independent of the contour as long as this is continuously deformed without passing the points 0 and 1. Together with the point at infinity these are the only singularities. When the contour is deformed to a large circle, the change in phase can be read off from Eq. (3.24). In the limit of large $|\lambda|$ we have

$$\Phi(\lambda) \approx (-\lambda)^{-(m+3)/2}. \quad (3.25)$$

For

$$\pi(n - \tfrac{1}{2}) < t/R < \pi(n + \tfrac{1}{2}), \quad n = 0, \pm 1, \ldots, \quad (3.26)$$

we can now replace λ in (3.24) and (3.25) by $\lambda e^{\mp 2\pi i n}$ and indicate the corresponding analytic continuation of $\Phi(\lambda)$ by $\Phi_n(\lambda)$. Then we have

$$\Phi_n(\lambda) \approx (e^{\mp 2\pi i n})^{-(m+3)/2}(-\lambda)^{-(m+3)/2},$$

so that

$$\Phi_n(\lambda) = e^{\pm \pi i(m+1)n}\Phi_0(\lambda). \quad (3.27)$$

Next let us find out how the function $\Phi_0(\lambda)$ looks when it is analytically continued to $|\lambda| < 1$. The appropriate expressions can be found using Ref. 20, page 63, formulas 18 and 19:

$$\Phi_0(\lambda) = C\Big[-\frac{1}{\lambda}$$
$$+ \frac{(m+1)(m+2)}{4}F\Big(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\Big)$$
$$\times \ln(-\lambda) + \chi(m,\lambda)\Big], \quad (3.28)$$

where

$$C = \Gamma(m + \tfrac{5}{2})/\Gamma((m+3)/2)\Gamma((m+4)/2), \quad (3.29)$$

and $\chi(m,\lambda)$ is a real analytic function of λ, regular in the domain $|\lambda| < 1$, which vanishes for $m = -1$.

The following restriction must be observed:

$$|\arg(-\lambda)| < \pi. \quad (3.30)$$

Note the similarity of this with the Minkowski case.

The expression (3.28) shows a simple pole and logarithmic singularity at $\lambda = 0$. These are avoided by considering $\Phi_0(\lambda_{\pm\epsilon})$ and its analytic continuations which are exact solutions of Eq. (3.12). The real and imaginary parts of $\Phi(\lambda_{\pm\epsilon})$ are exact solutions of (3.12) themselves. Since

$$\lim_{\epsilon\to 0}\left[\frac{1}{\lambda \pm i\epsilon\xi^4\xi^5/R^2}\right] = P(1/\lambda) \mp i\pi\epsilon(\xi^4\xi^5)\delta(\lambda), \tag{3.31}$$

we obtain with the help of (3.17) in the limit of small ϵ:

$$\mathrm{Re}\,\Phi_0(\lambda_{\pm\epsilon}) = C\left[-P\frac{1}{\lambda} + \frac{(m+1)(m+2)}{4} \times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right) \times \ln|\lambda| + \chi(m,\lambda)\right] \tag{3.32}$$

and

$$\mathrm{Im}\,\Phi_0(\lambda_{\pm\epsilon}) = \pm C\pi\epsilon(\xi^4\xi^5)\left[\delta(\lambda) - \frac{(m+1)(m+2)}{4} \times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right)\theta(\lambda)\right]. \tag{3.33}$$

The latter solution is zero for points which are spacelike separated from the origin, which in this context means that $\lambda < 0$ and $n = 0$ (Ref. 7).

The condition $|\lambda| < 1$ prevents ξ^5 from changing sign. Moreover, for the region $n = 0$, ξ^5 is positive and we could replace $\epsilon(\xi^4\xi^5)$ by $\epsilon(\xi^4)$. However, for odd n, we must insert negative values for ξ^5 in $\Phi_0(\lambda)$ as it is used in expression (3.28) so it is better to leave $\epsilon(\xi^4\xi^5)$ as it stands.

Since $\mathrm{Re}\,\Phi_0(\lambda_{\pm\epsilon})$ is an even function of ξ^4 and $\mathrm{Im}\,\Phi_0(\lambda_{\pm\epsilon})$ vanishes for $\boldsymbol{\xi}\neq 0$ and $\xi^4 = 0$ in the limit $\epsilon\to 0$, we find by direct calculation of the space integral over a ξ^4 derivative at $\xi^4 = 0$

$$\left[\frac{\partial}{\partial\xi^4}\Phi_0(\lambda_{\pm\epsilon})\right]\Bigg|_{\xi^4=0} = \pm 2\pi^2 iCR^2\delta^{(3)}(\boldsymbol{\xi}). \tag{3.34}$$

This enables us to define the propagators $G^m_{(0)}$ and $\overline{G}^m_{(0)}$ on the principal sheet ($n = 0$) by

$$G^m_{(0)}(\xi^M) = \frac{1}{2\pi R^2}\epsilon(\xi^4\xi^5)\left[\delta(\lambda) - \frac{(m+1)(m+2)}{4} \times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right)\theta(\lambda)\right] \tag{3.35}$$

and

$$\overline{G}^m_{(0)}(\xi^M) = \frac{1}{2\pi^2 R^2}\left[P\frac{1}{\lambda} - \frac{(m+1)(m+2)}{4} \times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right)\ln|\lambda| - \chi(m,\lambda)\right]. \tag{3.36}$$

With (3.27) we find the proper continuations of these functions

$$G^m_{(n)}(\xi^M) = \cos[\pi(m+1)n]G^m_{(0)}(\xi^M) - \sin[\pi(m+1)n]\overline{G}^m_{(0)}(\xi^M) \tag{3.37}$$

and

$$\overline{G}^m_{(n)}(\xi^M) = \sin[\pi(m+1)n]G^m_{(0)}(\xi^M) + \cos[\pi(m+1)n]\overline{G}^m_{(0)}(\xi^M). \tag{3.38}$$

All this is valid for $|\lambda| < 1$. For $\lambda < -1$ we must use

$$\Phi_n(\lambda) = e^{\pm\pi i(m+1)n}|\lambda|^{-(m+3)/2} \times F\left(\frac{m+1}{2}, \frac{m+3}{2}; m+\frac{5}{2}; \lambda^{-1}\right), \tag{3.39}$$

giving

$$G^m_{(0)}(\xi^M) = 0 \tag{3.40}$$

and

$$\overline{G}^m_{(0)}(\xi^M) = \frac{-1}{2\pi^2 CR^2}|\lambda|^{-(m+3)/2} \times F\left(\frac{m+1}{2}, \frac{m+3}{2}; m+\frac{5}{2}; \lambda^{-1}\right), \tag{3.41}$$

while for $G^m_{(n)}$ and $\overline{G}^m_{(n)}$ the relations (3.37) and (3.38) are still valid.

Apparently no mixing takes place when m is an integer. When m is even, no sign change occurs, but when m is odd there is a change of sign due to $\epsilon(\xi^5)$.

As already remarked, the case $m = -1$ is special. The expressions become particularly simple. We then have

$$G^{-1}_{(0)}(\xi^M) = (1/2\pi R^2)\epsilon(\xi^4\xi^5)\delta(\lambda) \tag{3.42}$$

and

$$\overline{G}^{-1}_{(0)}(\xi^M) = (1/2\pi^2 R^2)P(1/\lambda). \tag{3.43}$$

Here we must come back to the question of boundary conditions at infinity. Since $G^{-1}_{(0)}$ never mixes with $\overline{G}^{-1}_{(0)}$, the expression $G^{-1}_{(0)}$ as given by (3.42) is valid for all n. However,

$$\tilde{G}^{-1} = (1/2\pi R^2)\epsilon(\xi^4)\delta(\lambda) \tag{3.44}$$

is also a solution of the homogeneous equation, and there are many more.[16] It is therefore essential in the formulation of the boundary conditions at infinity that G^{-1} be obtained from G^m by taking the limit $m\to -1$. Then only expression (3.42) qualifies.

The full analytic continuations G^m and $\overline{G}^m$ satisfy the homogeneous equations

$$[M^2 - m(m+3)]G = 0. \tag{3.45}$$

One can define the positive and negative frequency parts by

$$G^m_{\pm} = \tfrac{1}{2}(\overline{G}^m \pm iG^m). \tag{3.46}$$

Then

$$G^m_{\pm(n)} = e^{\mp\pi i(m+1)n}G^m_{\pm(0)}, \tag{3.47}$$

where, for $|\lambda| < 1$, in the limit $\epsilon\to 0$

$$G^m_{\pm(0)} = \frac{1}{4\pi^2 R^2}\left[\frac{1}{\lambda_{\mp\epsilon}} - \frac{(m+1)(m+2)}{4} \times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right) \times \ln(-\lambda_{\mp\epsilon}) - \chi(m,\lambda)\right]. \tag{3.48}$$

For $m = -1$ we obtain

$$G^{-1}_{\pm(0)} = G^{-1}_{\pm} = 1/4\pi^2 R^2\lambda_{\mp\epsilon}. \tag{3.49}$$

for all m we have, from (3.34) and (3.13),

$$\left[\frac{\partial}{\partial t}G^m\right]\Bigg|_{t=0} = \delta^{(3)}(\boldsymbol{\xi}). \tag{3.50}$$

Next we consider the inhomogeneous equation. For small $a \neq 0$ we can calculate the space integral over the ξ^4 derivative of $\theta(\xi^4)\, G^m_{(0)}$ at the surface $\xi^4 = a$ and we find

$$\int d^3\xi \left\{ \frac{\partial}{\partial \xi^4} [\theta(\xi^4) G^m_{(0)}] \right\} \bigg|_{\xi^4 = a} = \theta(a). \tag{3.51}$$

Then, by choosing the surfaces $\xi^4 = \pm |a|$ for the application of Gauss' theorem and by making use of the properties of $G^m_{(0)}$ we find with the help of Eq. (3.11)

$$[M^2 - m(m+3)]\theta(\pm t) G^m = \mp R^2 \delta^{(4)}(\xi^\mu)\delta_{n0}. \tag{3.52}$$

Thus we can now define the retarded and advanced Green's functions by

$$G^m_{\substack{\text{ret}\\ \text{adv}}} = \pm \theta(\pm t) G^m, \tag{3.53}$$

which satisfy the equation

$$[M^2 - m(m+3)] G = -R^2 \delta^{(4)}(\xi^\mu)\delta_{n0}. \tag{3.54}$$

Finally, the Feynman propagator is defined as

$$G^m_{\mathrm{F}} = (1/2i)[\overline{G}^m + i\epsilon(t) G^m], \tag{3.55}$$

and satisfies the same equation (3.54). It is an even function of t and analytic in the variable λ. For $m = -1$ the Feynman propagator assumes the simple form

$$G_{\mathrm{F}}^{-1} = (1/4\pi^2 i R^2)/(\lambda - i\epsilon t \xi^4 \xi^5), \quad \epsilon \to 0. \tag{3.56}$$

Before closing this section we have to consider two exceptional points. The functions discussed so far are not defined by the given expressions when $|\lambda| = 1$. The case $\lambda = 1$ occurs when $\xi^5 = 0$ and is a singularity, while the case $\lambda = -1$ lies in the range $-\infty < \lambda < 0$ and is a regular point. It is useful to have expressions at hand which are valid in regions containing these points.

For $-\pi < t/R < \pi$ and $\lambda > 0$ (i.e., $|\xi^5| < R$) we have

$$\begin{aligned} G^m(\xi^M) = \frac{\epsilon(\xi^4)}{2\pi R^2} \Bigg\{ & \frac{\sqrt{\pi}}{\Gamma(-(m+1)/2)\Gamma((m+2)/2)} \\ & \times F\left(-\frac{m}{2}, \frac{m+3}{2}; \frac{1}{2}; \frac{\xi^{5^2}}{R^2}\right) \\ & - \frac{2\sqrt{\pi}}{\Gamma(-(m+2)/2)\Gamma((m+1)/2)} \\ & \times \frac{\xi^5}{R} F\left(-\frac{m+2}{2}, \frac{m+1}{2}; \frac{3}{2}; \frac{\xi^{5^2}}{R^2}\right) \Bigg\} \end{aligned} \tag{3.57}$$

and

$$\begin{aligned} \overline{G}^m(\xi^M) = \frac{1}{2\pi R^2} \Bigg\{ & \frac{\sqrt{\pi} \cot[\pi(m+1)/2]}{\Gamma(-(m+1)/2)\Gamma((m+2)/2)} \\ & \times F\left(-\frac{m}{2}, \frac{m+3}{2}; \frac{1}{2}; \frac{\xi^{5^2}}{R^2}\right) \\ & + \frac{2\sqrt{\pi} \tan[\pi(m+1)/2]}{\Gamma(-(m+2)/2)\Gamma((m+1)/2)} \\ & \times \frac{\xi^5}{R} F\left(-\frac{m+2}{2}, \frac{m+1}{2}; \frac{3}{2}; \frac{\xi^{5^2}}{R^2}\right) \Bigg\}. \end{aligned} \tag{3.58}$$

The relations (3.37) and (3.38) can now be used for finding G^m and $\overline{G}^m$ for other t values.

The expressions valid in the whole region $-\infty < \lambda < 0$ are particularly simple. For $-\pi/2 < t/R < \pi/2$ we have[20]

$$G^m(\lambda) = 0 \tag{3.59}$$

and

$$\begin{aligned} \overline{G}^m(\lambda) = \frac{-1}{2\pi^2 R^2} & \frac{\Gamma((m+3)/2)\Gamma((m+4)/2)}{\Gamma(m+\frac{5}{2})} \\ & \times (1-\lambda)^{-(m+3)/2} \\ & \times F\left(\frac{m+3}{2}, \frac{m+4}{2}; m+\frac{5}{2}; \frac{1}{1-\lambda}\right). \end{aligned} \tag{3.60}$$

Again, Eq. (3.39) and (3.40) can be used for finding G_m and $\overline{G}_m$ for other t values.

IV. THE FORM OF THE ANTI-DE SITTER PROPAGATORS FOR AN ARBITRARY REFERENCE POINT

So far all propagators have been written down for the case that the reference point ξ_0^M has been specially chosen: $\xi_0^M = (0, 0, 0, 0, R)$. We now choose the reference point arbitrarily on the hypersurface $\xi_0^M \xi_{0M} = R^2$ and denote the propagators and Green's functions by

$$G(\xi^M; \xi_0^M). \tag{4.1}$$

The quantities λ, $\epsilon(\xi^4\xi^5)$ and the winding number n must be written in invariant form. The following quantities are manifestly invariant under general O(3, 2):

$$\xi^M \xi_M = \xi_0^M \xi_{0M} = R^2 \quad \text{and} \quad \xi^M \xi_{0M} = \gamma. \tag{4.2}$$

If S is defined by

$$S = \xi^4 \xi_0^5 - \xi^5 \xi_0^4, \tag{4.3}$$

then

$$\operatorname{sgn} S \tag{4.4}$$

is an invariant under orthochronous O(3, 2) and so are the functions $\theta(S)$ and $\epsilon(S)$.

Now, when ξ_{0M} has the special form (0, 0, 0, 0, R) we find

$$\xi^5 = \gamma/R \quad \text{and} \quad \xi^4 = S/R. \tag{4.5}$$

So we find the following invariant forms:

$$\lambda = 1 - [\gamma/R^2]^2 \tag{4.6}$$

and

$$\epsilon(\gamma S). \tag{4.7}$$

The definition of winding number n can be given as follows.

(i) $n = 0$ when ξ^M can be obtained from ξ_0^M by continuous displacement within the allowed domain, without changing the sign of γ.

(ii) $\Delta n = \pm 1$ whenever γ changes sign and $\Delta t \gtrless 0$ with t given by (3.13).

With λ defined by (4.6) we now have, for $|\lambda| < 1$

$$\begin{aligned} G^m_{(0)}(\xi^M; \xi_0^M) = \frac{1}{2\pi R^2} \epsilon(\gamma S) \Bigg[\delta(\lambda) - & \frac{(m+1)(m+2)}{4} \\ & \times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right) \theta(\lambda) \Bigg], \end{aligned} \tag{4.8}$$

while $\overline{G}^{\,m}_{(0)}(\xi^M;\xi^M_0)$ is given by (3.36). Moreover, we have

$$G^{\,m}_{(n)}(\xi^M;\xi^M_0)=\cos[\pi(m+1)n]G^{\,m}_{(0)}(\xi^M;\xi^M_0)$$
$$-\sin[\pi(m+1)n]\overline{G}^{\,m}_{(0)}(\xi^M;\xi^M_0) \qquad (4.9)$$

and

$$\overline{G}^{\,m}_{(n)}(\xi^M;\xi^M_0)=\sin[\pi(m+1)n]G^{\,m}_{(0)}(\xi^M;\xi^M_0)$$
$$+\cos[\pi(m+1)n]\overline{G}^{\,m}_{(0)}(\xi^M;\xi^M_0). \qquad (4.10)$$

When $\lambda<-1$ the functions $G^{\,m}_{(0)}(\xi^M;\xi^M_0)$ and $\overline{G}^{\,m}_{(0)}(\xi^M;\xi^M_0)$ have the forms (3.40) and (3.41), respectively, while (4.9) and (4.10) are maintained. A similar discussion can be given for the expression (3.57)–(3.60). Similar replacements can be made for the positive and negative frequency parts. The retarded, advanced, and Feynman propagators need some extra attention. Let t be defined by (3.13) and t_0 similarly by

$$\xi^4_0=\sqrt{R^2+\boldsymbol{\xi}_0^2}\,\sin(t_0/R)$$

and

$$\xi^5_0=\sqrt{R^2+\boldsymbol{\xi}_0^2}\,\cos(t_0/R). \qquad (4.11)$$

Then $t-t_0$ can only change sign when $\lambda<0$ and $n=0$. In that case either $G^m(\xi^M;\xi^M_0)=0$ or $\xi^M=\xi^M_0$, which makes $\theta(t-t_0)$ and $\epsilon(t-t_0)$ effectively invariant functions under orthochronous transformations when they are used in combination with G^m. Thus we find

$$G^{\,m}_{\substack{\rm ret\\ \rm adv}}(\xi^M;\xi^M_0)=\pm\theta[\pm(t-t_0)]G^m(\xi^M;\xi^M_0) \qquad (4.12)$$

and

$$G^m_{\rm F}(\xi^M;\xi^M_0)=(1/2i)[\overline{G}^m(\xi^M;\xi^M_0)+i\epsilon(t-t_0)G^m(\xi^M;\xi^M_0)]. \qquad (4.13)$$

These functions satisfy the inhomogeneous equation

$$[M^2-m(m+3)]G(\xi^M;\xi^M_0)=-R\xi^5_0\delta^{(4)}(\xi^\mu-\xi^\mu_0)\delta_{n0}, \qquad (4.14)$$

where the right-hand side is an invariant function.

To avoid the explicit use of n we may define for all propagators

$$G(\boldsymbol{\xi},t;\boldsymbol{\xi}_0,t_0)=G_{(n)}(\xi^M;\xi^M_0). \qquad (4.15)$$

It is perhaps expedient to remark that n is not just a function of $t-t_0$, so that a generalization of (3.26) is not possible.

For G as defined by (4.15) we have the relation

$$G(\boldsymbol{\xi},t;\boldsymbol{\xi}_0,t_0)=G(\boldsymbol{\xi},t-t_0;\boldsymbol{\xi}_0,0). \qquad (4.16)$$

The generalization of Eq. (3.50) is

$$\left[\frac{\partial}{\partial t}G^m(\boldsymbol{\xi},t;\boldsymbol{\xi}_0,t_0)\right]\bigg|_{t=t_0}=\frac{R^2+\boldsymbol{\xi}_0^2}{R^2}\delta^{(3)}(\boldsymbol{\xi}-\boldsymbol{\xi}_0). \qquad (4.17)$$

Due to the fact that $G^m(\boldsymbol{\xi},t;\boldsymbol{\xi}_0,t_0)$ is an odd function and its time derivative is an even function of $t-t_0$, one can now make use of expression (4.17) to find the general solution of the equation

$$[M^2-m(m+3)]\chi(\boldsymbol{\xi},t)=0, \qquad (4.18)$$

when χ and its time derivative are given on a suitable spacelike surface (for example, $t=0$) and the implicit boundary conditions at infinity are satisfied.

In terms of G as defined by (4.15) we can rewrite (4.14) as follows:

$$[M^2-m(m+3)]G(\boldsymbol{\xi},t;\boldsymbol{\xi}_0,t_0)$$
$$=-R^2\delta^{(3)}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)\delta(t-t_0). \qquad (4.19)$$

This enables one to use the advanced or retarded Green's functions for solving the inhomogeneous equation

$$[M^2-m(m+3)]\chi(\boldsymbol{\xi},t)=\rho(\boldsymbol{\xi},t). \qquad (4.20)$$

V. CENTRAL PROJECTION COORDINATES

Let us consider the following cylinder in ξ^M space:

$$\xi^{4^2}+\xi^{5^2}=R^2. \qquad (5.1)$$

The intrinsic curvature of this non-simply-connected manifold is zero and the signature is that of ordinary Minkowski space. The covering space of this cylinder is simply connected and could be identified with ordinary Minkowski spacetime. The reason for introducing it is that by central projection one can represent every point on the hyperboloid $\xi_M\xi^M=R^2$ by a point on the cylinder in such a way as to preserve the winding number. The time as it has been introduced in earlier sections now becomes a natural time variable in flat space. Infinity on the hyperboloid is represented by $\boldsymbol{\xi}^2=R^2$ on the cylinder and any other point is represented by $\boldsymbol{\xi}^2<R^2$. thus all points of the original hyperboloid are now represented by points inside a sphere of radius R. Moreover, a timelike trajectory on the hyperboloid is represented by a timelike trajectory on the cylinder, confined to the sphere. A timelike geodesic on the hyperboloid is represented by a timelike, confined oscillatory trajectory on the cylinder. Note that points on the cylinder outside the sphere are not represented on the hyperboloid, so there is no one-to-one correspondence.

Let x^M represent a point on the cylinder. Then we have, when x^M is obtained from ξ^M on the hyperboloid by central projection

$$x^M=(R/\sqrt{R^2+\boldsymbol{\xi}^2})\xi^M. \qquad (5.2)$$

The "natural coordinates" $\mathbf{x}$ and t are then obtained by

$$\mathbf{x}=(R/\sqrt{R^2+\boldsymbol{\xi}^2})\,\boldsymbol{\xi} \qquad (5.3)$$

and

$$t=R\arctan(\xi^4/\xi^5)+n\pi \quad (n=0,\pm 1,\ldots).$$

Let t be x_0. When the SO(3,2) metric η_{MN} as it is present on the hyperboloid is projected on the cylinder then an effective metric $g_{\mu\nu}(\mu,\nu=0,1,2,3)$ is generated which has the following analytic form:

$$g_{kl}=\frac{-R^2}{R^2-r^2}\left(\delta_{kl}+\frac{x^kx^l}{R^2-r^2}\right) \quad (k=1,2,3),$$
$$g_{k0}=g_{0k}=0,\quad g_{00}=R^2/(R^2-r^2), \qquad (5.4)$$

where $r^2=x_1^2+x_2^2+x_3^2<R^2$. This metric is not defined for $r^2\geqslant R^2$.

For finding the propagators in central projection coordinates, the only thing one has to do is to write the variables λ, γ, etc., in terms of $\mathbf{x}$ and t. We have

$$\xi^M = (R/\sqrt{R^2 - r^2})x^M, \quad x^4 = R\sin(t/R),$$
$$x^5 = R\cos(t/R), \tag{5.5}$$

so that with (4.2) we obtain

$$\gamma = \frac{R^2[-\mathbf{x}\cdot\mathbf{x}_0 + R^2\cos((t-t_0)/R)]}{\sqrt{(R^2-r^2)(R^2-r_0^2)}}. \tag{5.6}$$

Furthermore, with (4.6),

$$\lambda = 1 - \frac{[-\mathbf{x}\cdot\mathbf{x}_0 + R^2\cos((t-t_0)/R)]^2}{(R^2-r^2)(R^2-r_0^2)}, \tag{5.7}$$

and finally

$$\mathrm{sgn}(\xi^4\xi_0^5 - \xi^5\xi_0^4) = \mathrm{sgn}[\sin(t-t_0)/R]. \tag{5.8}$$

With the help of (5.4)–(5.8), Eqs. (4.19) for the inhomogeneous propagators can be rewritten in the form

$$[M^2 - m(m+3)]G(\mathbf{x},t;\mathbf{x}_0,t_0)$$
$$= -[(R^2-r^2)^{3/2}/R]\delta^{(3)}(\mathbf{x}-\mathbf{x}_0)\delta(t-t_0), \tag{5.9}$$

where it may be of use to reexpress M^2 by means of the Laplace–Beltrami operator

$$\partial_\mu\sqrt{-g}\,g^{\mu\nu}\partial_\nu = -(1/R^2)\sqrt{-g}\,M^2, \tag{5.10}$$

with $g^{\mu\nu}$ the inverse of $g_{\mu\nu}$ given by (5.4) and g the determinant of $g_{\mu\nu}$.

ACKNOWLEDGMENTS

We wish to thank Dr. T. A. Rijken for useful discussions. One of the authors (E.v.B.) wishes to thank Dr. W. van Neerven for many stimulating discussions and the Dutch Institute for Nuclear and High Energy Physics (NIKHEF) for the kind hospitality during his stay in Amsterdam. Also we have benefited from discussions with Professor Dr. L. S. Frank and Mr. H. Janssen (Nijmegen).

[1]For the MIT bag model see A. Chodos, R. L. Jaffe, K. Johnson, C. B. Thorn, and V. Weisskopf, Phys. Rev. D **9**, 3471 (1974); T. A. DeGrand, R. L. Jaffe, K. Johnson, and J. J. Kiskis, *ibid.* **12**, 2060 (1975); P. J. Mulders, A. T. Aerts, and J. J. de Swart, Phys. Rev. D **19**, 2635 (1979); **21**, 1370, 2653 (1980).

[2]T. Appelqvist, A. de Rùjula, S. L. Glashow, and H. D. Politzer, Phys. Rev. Lett. **34**, 365 (1975); R. Barbieri, R. Gatto, R. Kögerler, and Z. Kunst, Phys. Lett. B **57**, 445 (1975); E. Eichten, K. Gottfried, K. D. Lane, T. Kinoshita, and T. M. Yan, Phys. Rev. D **17**, 3090 (1978); **21**, 313 (1980); **21**, 203 (1980); J. S. Kang and H. J. Schnitzer, Phys. Rev. D **12**, 841 (1975); J. Pumplin, W. Repko, and A. Sato, Phys. Rev. Lett. **35**, 1538 (1975); J. L. Richardson, Phys. Lett. B **82**, 272 (1979); C. Quigg and J. L. Rosner, Phys. Rep. **56**, 167 (1979); A. Martin, Phys. Lett. B **100**, 511 (1981); W. Buchmüller and S-H. H. Tye, Phys. Rev. D **24**, 132 (1981).

[3]C. Isham, A. Salam, and J. Strathdee, Phys. Rev. D **3**, 867 (1971); S. Fubini, Nuovo Cimento A **34**, 521 (1976); B. Zumino, Nucl. Phys. B **127**, 189 (1977); A. Salam and J. Strathdee, Phys. Rev. D **18**, 4596 (1978); Phys. Lett. B **67**, 429 (1977); Z. Haba, Phys. Rev. D **18**, 4610 (1978); Phys. Lett. B **78**, 421 (1978); see also C. Sivaram and K. P. Sinha, Phys. Rep. **51**, 111 (1979).

[4]C. Dullemond and E. van Beveren, Phys. Rev. D **28**, 1028 (1983); C. Dullemond, T. A. Rijken, and E. van Beveren, Nuovo Cimento A **80**, 401 (1984); E. van Beveren, C. Dullemond, and T. A. Rijken, Phys. Rev. D **30**, 1103 (1984); C. Dullemond, J. Math. Phys. **25**, 2638 (1984); E. van Beveren, T. A. Rijken, and C. Dullemond, THEF-NYM-84.02, which appeared in *Proceedings of the XIX Rencontre de Moriond*, edited by J. Tran Thanh Van (Editions Frontières, Gif sur Yvette, 1984), p. 847; E. van Beveren, T. A. Rijken, C. Dullemond, and G. Rupp, in *Proceedings of the Bielefeld Workshop on Resonances*, edited by S. Albeverio, L. S. Ferreira, and L. Streit (Springer, Berlin, 1984), p. 331.

[5]For classical electrodynamics, see H. Weyl, Ann. Phys. **59**, 101 (1919).

[6]S. L. Adler, Phys. Rev. D **6**, 3445 (1972).

[7]C. Fronsdal, Phys. Rev. D **10**, 589 (1974).

[8]C. Fronsdal and R. B. Haugen, Phys. Rev. D **12**, 3810 (1975). See also H. Janssen and C. Dullemond, University of Nijmegen preprint THEF-NYM-85.01.

[9]C. Fronsdal, Phys. Rev. D **20**, 848 (1979).

[10]J. Fang and C. Fronsdal, Phys. Rev. D **22**, 1361 (1980).

[11]B. Binegar, C. Fronsdal, and W. Heidenreich, Ann. Phys. (NY) **149**, 254 (1983); J. Math. Phys. **24**, 2828 (1983).

[12]Configuration space methods are widely used in curved spaces. See *Relativity, Groups and Topology*, edited by B. S. DeWitt and C. DeWitt (Gordon and Breach, New York, 1963). See also P. Candelas and D. J. Raine, Phys. Rev. D **12**, 965 (1975); E. Mottola, ITP-84-123, Institute for Theoretical Physics, Santa Barbara. More complete references can be found in N. D. Birrell and P. C. W. Davies, *Quantum fields in Curved Space* (Cambridge U.P., London, 1982).

[13]This requirement may be too stringent for certain physically interesting systems, as has been shown by Breitenlohner and Freedman.[14] See also Ref. 15.

[14]P. Breitenlohner and D. Z. Freedman, Ann. Phys. (NY) **144**, 249 (1982). See also L. Mezincescu and P. K. Townsend, University of Texas preprint UTTG-8-84 and C. P. Burgess and C. A. Lutken, University of Texas preprint UTTG-29-84.

[15]As proven by Breitenlohner and Freedman,[14] only in this case can the regular definition of the energy momentum stress tensor for field fluctuations lead to meaningful total energies. There exists however, an improved energy momentum stress tensor which remains meaningful for solutions which fall off faster than $(-\lambda)^{-1/4}$ but slower than $(-\lambda)^{-3/4}$. In that case there is a choice of boundary conditions.

[16]C. Fronsdal, Phys. Rev. D **12**, 3819 (1975); see also S. J. Avis, C. J. Isham, and D. Storey, Phys. Rev. D **18**, 3565 (1978).

[17]See, e.g., N. N. Bogoliubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields* (Interscience, New York, 1959); N. N. Bogoliubov and D. V. Shirkov, *Quantum Fields* (Benjamin/Cummings, New York, 1983); J. D. Bjørken and S. D. Drell, *Relativistic Quantum Fields* (McGraw–Hill, New York, 1965; C. Itzykson and J-B. Zuber, *Quantum Field Theory* (McGraw–Hill, New York, 1980)).

[18]Bateman Manuscript Project, *Higher Transcendental Functions*, Vol. 2, edited by A. Erdélyi. (McGraw–Hill, New York, 1953).

[19]See, e.g., G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge U.P., London, 1966).

[20]Bateman Manuscript Project, *Higher Transcendental functions*, Vol. 1, edited by A. Erdélyi. (McGraw–Hill, New York, 1953).

A prediction of the Cabibbo angle in the vector model for electroweak interactions

Frank Reifler and Randall Morris
RCA/Government & Commercial Systems, Moorestown, New Jersey 08057

(Received 31 October 1984; accepted for publication 1 March 1985)

In a recent paper we presented a vector model for the electroweak interactions which is similar to the Weinberg–Salam model but differs in the following features. (1) In the vector model all fermion wave functions are bispinors or equivalently isotropic Yang–Mills triplets (as opposed to a state vector composed of a spinor and bispinors in the Weinberg–Salam model). Particles are distinguished by their Higgs fields. (2) The vector model predicts that $\sin^2\theta_W = \frac{1}{4}$, where θ_W is the Weinberg angle. (3) The vector model accounts for conservation of lepton number, electric charge, and baryon number. (4) In the vector model an antiparticle is characterized by opposite lepton number, electric charge, and baryon number; yet both particles and antiparticles propagate forward in time with positive energies. In this paper we extend the vector theory to include interactions between fermions and the gauge bosons mediating the electroweak force. We model the bosons as Yang–Mills fields with their own Higgs fields. We further propose a specific configuration of Higgs fields for the u,d,s, and c quarks. With these features, the model accounts for electroweak transitions of quarks and leptons and predicts that $\cos\theta_C = 0.9744$, where θ_C is the Cabibbo angle. We further show that the vector model accounts for the intrinsic parity of particles and antiparticles, and parity violations and *CPT* invariance for electroweak interactions.

I. A SUMMARY OF THE VECTOR MODEL FOR ELECTROWEAK INTERACTIONS

In a recent paper,[1] we presented a vector model for the electroweak interactions. We showed that the Cartan map gives an isomorphism between bispinors and an isotropic class of Yang–Mills vector fields. A bispinor $\tilde{\psi} = (\xi, \eta^*)$ consists of a spinor ξ and a conjugated[2] spinor η^*. The Cartan map takes each bispinor $\tilde{\psi}$ to a triplet of Yang–Mills vector fields $(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)$. These Yang–Mills vector fields $(\mathbf{F}_k = \mathbf{E}_k + i\mathbf{H}_k$, for $k = 1,2,3)$ satisfy the isotropic condition that the matrix of scalar invariants, $(\mathbf{F}_j \cdot \mathbf{F}_k)$, be a scalar multiple of the identity matrix. That is, by definition,

$$\mathbf{F}_j \cdot \mathbf{F}_k = \rho^2 \delta_{jk}, \tag{1.1}$$

with j, $k = 1,2,3$, and ρ is a complex scalar field[3] determined by the $\mathbf{F}_k$.

We showed that the Cartan map is locally one-to-one from C^4 onto the manifold of isotropic Yang–Mills vector fields. We also showed that the Cartan map commutes with all bispinor observables, and Lorentz and gauge transformations.[4]

The study of isotropic Yang–Mills vector fields reveals that bispinors have SL(2, C) gauge symmetry. The gauge group SL(2, C) acts on Yang–Mills triplets $(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)$ via the complex orthogonal matrices. The equivalent SL(2, C) action on the bispinors cannot be represented by complex matrices. Instead the equivalent SL(2, C) action on bispinors is noncomplex linear which obfuscates the SL(2, C) structure. (See Appendix A.)

In fact, isotropic Yang-Mills vector fields $\mathbf{F}_k$ transform under the even bigger gauge symmetry group SL(2, C)$\times$U(1). The subgroup U(1) consists of the neutral (chiral) gauge transformations which map $\mathbf{F}_k$ to $\mathbf{F}_k e^{i\chi}$, where χ is a phase.

Associated with each isotropic triplet of Yang–Mills vector fields $(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)$, besides the unique complex scalar ρ, there is a unique quadruplet of orthogonal real Lorentz four-vectors $(j_0^\alpha, j_1^\alpha, j_2^\alpha, j_3^\alpha)$. By the Cartan map these are all SL(2, C) invariants associated with each bispinor $\tilde{\psi}$. For the wave equation to be SL(2, C)$\times$U(1) gauge invariant, a quadruplet of *real* Higgs scalars $(\phi_0, \phi_1, \phi_2, \phi_3)$ as well as a singlet complex Higgs scalar ϕ_5 are also required.[5] That is, the ϕ_K for $K = 0,1,2,3$, and 5 are scalars for Lorentz transformations. The ϕ_α for $\alpha = 0,1,2,3$ transform as a quadruplet under SL(2, C) gauge transformations, but as scalars under the neutral U(1) gauge transformations. Similarly, the singlet ϕ_5 is a scalar for SL(2, C) gauge transformations, but undergoes a phase change $e^{i\chi}$ for the U(1) gauge transformations.

Both the Dirac equation for bispinors and the electric current are derived from a Lagrangian. In our previous paper,[6] we extended this Lagrangian to incorporate these Higgs scalars and to derive seven conserved Noether currents [one for each generator of SL(2, C)$\times$U(1)], denoted $(J_1^\alpha, J_2^\alpha, J_3^\alpha)$ and J_5^α, and given by the formulas

$$\begin{aligned} \operatorname{Re} J_k^\alpha &= \phi_0 j_k^\alpha - \phi_k j_0^\alpha, \\ \operatorname{Im} J_k^\alpha &= -\epsilon_{kmn}\phi_m j_n^\alpha, \quad J_5^\alpha = \phi^\beta j_\beta^\alpha, \end{aligned} \tag{1.2}$$

with $k, m, n = 1,2,3$ and $\alpha, \beta = 0,1,2,3$. The gauge-invariant vector equivalent of the Dirac equation derived from this extended Lagrangian is given for isotropic Yang–Mills triplets $\mathbf{F}_k$ by[7]

$$iD_\alpha S^\alpha \mathbf{F}_k + (\mathbf{D}\mathbf{F}_m)\cdot \mathbf{F}_n/\rho = M\phi_5 \mathbf{J}_k, \tag{1.3}$$

with subscripts $(k\ m\ n)$ taken in cyclic order, and where the D_α are the Yang-Mills covariant derivatives, the S^α are the Proca spin-one matrices, and M is the mass.

Because the mass M is given explicitly in Eq. (1.3), the

 0022-2488/85/082059-08$02.50

Higgs scalars must be normalized as follows:

$$\phi^\beta \phi_\beta = -1, \quad |\phi_5| = 1. \tag{1.4}$$

With J_1, J_2, J_3 and J_5 as the electroweak currents, and with $(\phi_0, \phi_1, \phi_2, \phi_3)$ and ϕ_5 as the Higgs scalars, we have a vector model for the electroweak interactions different from the Weinberg–Salam model. In our previous paper, we showed that the electric current is given by Re(J_3) and the neutral current is J_5. Equating these currents to those in the Weinberg–Salam model leads to the prediction that

$$\sin^2 \theta_{\rm W} = \tfrac{1}{4},$$

where $\theta_{\rm W}$ is the Weinberg angle.[8]

Another prediction of the vector model is that the quadruplet of Higgs scalars $\phi_\beta = (\phi_0, \phi_1, \phi_2, \phi_3)$ is *additive in all electroweak interactions*. This additivity was shown to be a consequence of conserving the Noether currents, given in formulas (1.2). Thus, the vector model predicts the existence of four additive "quantum numbers" consisting of the Higgs scalars ϕ_β.

Actual assignment of Higgs scalars depends partly on historical convention, and partly on the condition of additivity. Conventionally we must represent the electron by the Higgs scalars,

$$\phi_\alpha(e) = (0, 0, 0, 1), \quad \phi_5(e) = 1.$$

This makes Eq. (1.3) equivalent to the usual Dirac equation for electrons. Via the Cartan map the electromagnetic gauge action becomes the formal gauge rotation about $\phi_\alpha(e)$, i.e., the "three-axis." [This implies that the electric current is Re(J_3).] Since currents must be reversed for antiparticles, we must have

$$\phi_\alpha(\bar{a}) = -\phi_\alpha(a),$$

where a denotes a particle and $\bar{a}$ denotes its antiparticle. We define ϕ_5 as the intrinsic parity of the particle, which is also reversed for antifermions. Substituting $\phi_\alpha(\bar{a})$ by $-\phi_\alpha(a)$ and $\phi_5(\bar{a})$ by $-\phi_5(a)$ in the plane wave solutions of Eq. (1.3), shows that antiparticles propagate forward in time with positive energies.

With the electron assigned, the Higgs scalars ϕ_β for the particles ν, u, and d are almost uniquely determined.[9] Since the electric current Re(J_3) of the neutrino vanishes, we must have from (2) that $\phi_3(\nu) = 0$. Thus $\phi_\beta(\nu)$ is any *unit* vector orthogonal to $\phi_\beta(e)$, and we may as well designate it as $\phi_\beta(\nu) = (0, 1, 0, 0)$, without any loss of generality. Then the *unit* vectors $\phi_\beta(u)$ and $\phi_\beta(d)$ are constrained by the electric charges $\phi_3(d) = \frac{1}{3}$ and $\phi_3(u) = -\frac{2}{3}$, and also by the formula for beta decay:

$$\phi_\beta(d) + \phi_\beta(\nu) = \phi_\beta(u) + \phi_\beta(e). \tag{1.5}$$

As a consequence of formula (1.5), $\phi_\beta(u)$ and $\phi_\beta(d)$ are orthogonal unit vectors, as are $\phi_\beta(e)$ and $\phi_\beta(\nu)$.

It is evident from formula (1.2) that ϕ_0 must vanish to make the vector part of the electric current vanish in the particle's rest frame (in which by definition $\mathbf{j}_0$ vanishes but the $\mathbf{j}_k$ do not). For this reason ϕ_0 will be omitted from further discussion (only the Higgs scalars ϕ_k, ϕ_5 with $k = 1,2,3$ will be discussed). Also, we restrict the theory to the gauge transformations in the subgroup SU(2)$\times$U(1), which leaves ϕ_0 invariant. Formulas (1.2) for the conserved Noether currents then reduce to

$$J^\alpha_k = -\phi_k j^\alpha_0, \quad J^\alpha_5 = \phi^k j^\alpha_k, \tag{1.6}$$

i.e., four real conserved currents, one for each generator of SU(2)$\times$U(1). Now the vector part of the electric current $\mathbf{J}_3 = -\phi_3 \mathbf{j}_0$ vanishes in the particle's rest frame (i.e., when $\mathbf{j}_0 = 0$).

With these restrictions there are two possible assignments for $\phi_k(u)$ and two corresponding assignments for $\phi_k(d)$, which are shown in Table I. We chose the first of these in our previous paper.[10] This choice of $\phi_k(u)$ and $\phi_k(d)$ along with the assignment of the electron and neutrino leads to the following definition of electric charge Q, lepton number L, and baryon number B:

$$Q = -\phi_3, \quad L = \phi_1 - \tfrac{1}{2}\phi_2 + \phi_3, \quad B = -\tfrac{1}{2}\phi_2. \tag{1.7}$$

Since B, L, and Q are linear functions of the ϕ_k they must be additive also. Thus the vector model predicts the conservation of baryon number and lepton number, as well as the electric charge.

In this paper we extend the previous Lagrangian to include the W bosons. Although we have not yet modeled the Higgs field dynamics, the following assumptions allow analysis of particle transitions.

(1) The total Noether currents for fermions and the W bosons are conserved for an interaction of the type $a \to b + W$, where a and b are fermions and W is a gauge boson. This implies that the sum of the W boson and fermion Higgs scalars are additive in the interaction; i.e., $\phi_k(a) = \phi_k(b) + \mu_k(W)$, where the μ_k denote the boson Higgs scalars.

(2) The transition currents, denoted by $j^\alpha(a \to b)$ have vector and axial components which are both SU(2)$\times$U(1) gauge invariant.

Using these assumptions, we argue in Sec. II that the transition matrix T for the four fermion interaction $(a \to b)$, $(a' \to b')$ has the following form:

$$T = 2e^2 j^\alpha(a \to b) j_\alpha(a' \to b') \hat{\phi}_k \hat{\phi}'_k / (m_w^2 - q^2), \tag{1.8}$$

where

$(\hat{\phi}_k = \frac{1}{2}(\phi_k(a) + \phi_k(b)), m_{\rm W}$ is the W boson mass,

$q^2 = q^\alpha q_\alpha$, and q^α is the momentum transfer.

In Sec. III we show the gauge-invariant transition currents have the form

$$j^\alpha(a \to b) = \cos(\theta_{ab}/2) \tilde{j}^\alpha(a \to b), \tag{1.9}$$

where θ_{ab} is the angle between Higgs formal vectors, $\phi_k(a)$

TABLE I. Higgs scalars for quarks and leptons.

Quark flavors[a]	ϕ_α
u, c, t	$(0, \frac{1}{3}, \pm\frac{2}{3}, -\frac{2}{3})$
d, s, b	$(0, -\frac{2}{3}, \pm\frac{2}{3}, \frac{1}{3})$
Lepton flavors[a]	ϕ_α
e, μ, τ	$(0, 0, 0, 1)$
ν_e, ν_μ, ν_τ	$(0, 1, 0, 0)$

[a] Antiparticles: replace ϕ_α with $-\phi_\alpha$.

and $\phi_k(b)$, and where $\tilde{j}^\alpha(a\to b)$ is the linear combination of the usual bispinor vector and axial currents.

By application of this transition current to Eq. (1.8) we further show in Sec. III that the mass of the W boson is the same as that predicted by the Weinberg–Salam model.

In Sec. IV we extend our assignment of fermion Higgs scalars to the charmed and strange quarks. A consequence of this assignment is the nonconservation of the fermion Higgs scalar ϕ_2 and the production of a more massive boson W' during strangeness changing interactions. A straightforward extension of formula (1.8) leads to a prediction of the Cabibbo angle.

Contrary to observation, in the standard bispinor model the charge conjugation, parity, and time reversal operations, denoted C, P, and T, should commute with weak interactions. Translation of the C, P, T bispinor operations into operators on isotropic triplets ($\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$) or equivalently spinor pairs[11] reveals that these operators do not commute with general $SU(2)\times U(1)$ gauge transformations and consequently are not conserved across electroweak interactions in the vector model. However, the product, CPT, does commute with all the gauge transformations and hence with the transition matrices for electroweak interactions. Thus the vector model predicts CPT invariance for electroweak interactions. These results are stated formally in Appendix B and are summarized in Table II.

We are still investigating a dynamical model for the Higgs fields. Such a model would include a Lagrangian for the Higgs fields which would support the conservation of fermion and boson Noether currents which we assume here. We will address the origin of the axial current in a forthcoming paper. The main achievements of this paper are the following.

(1) The establishment of the form of the transition matrix derived from conservation of Noether currents.

(2) The derivation of formulas for gauge-invariant vector and axial transition currents.

(3) The extension of the vector model to strange and charmed quarks and a prediction of the Cabibbo angle.

(4) The prediction of CPT invariance and parity violation for electroweak interactions.

TABLE II. C, P, T operations. $R_2 = 180°$ gauge rotation about the two-axis. $R_5 = 180°$ neutral gauge rotation. $F^* = (\mathbf{E} + i\mathbf{H})^* = \mathbf{E} - i\mathbf{H}$. $x^\alpha = (t, \mathbf{x})$ = space-time coordinates.

Charge conjugation C:	$\mathbf{F}_k \to R_5 R_2 \mathbf{F}_k$
	$\phi_\beta \to -\phi_\beta$
	$\phi_5 \to -\phi_5$
Parity P:	$\mathbf{F}_k \to \mathbf{F}_k^*$
	$\mathbf{x} \to -\mathbf{x}$
Time reversal T:	$\mathbf{F}_k \to R_2 \mathbf{F}_k^*$
	$t \to -t$
CPT:	$\mathbf{F}_k \to R_5 \mathbf{F}_k$
	$\phi_\beta \to -\phi_\beta$
	$\phi_5 \to -\phi_5$
	$x^\alpha \to -x^\alpha$

II. ELECTROWEAK TRANSITIONS

To describe the electroweak interactions, we must start from the interaction Lagrangian, which may be defined as

$$L_I = J_k^\alpha W_{k\alpha} + J_5^\alpha Z_\alpha, \tag{2.1}$$

where the W_k^α are the electroweak potentials, the Z^α are the neutral potentials, and J_k^α and J_5^α are the Noether currents.

We will not discuss the neutral potentials, denoted Z^α, in this paper. The electroweak potentials, denoted W_k^α with $k = 1,2,3$, satisfy the linearized Yang–Mills equations which are written in the following form with the fermion Noether currents J_k^α as sources[12]:

$$D^\alpha D_\alpha W_k^\beta + M_{kn} W_n^\beta = J_k^\beta, \tag{2.2}$$

where the D_α are the Yang–Mills derivatives and M_{kn} is the mass matrix. In the vector model the mass matrix M_{kn} will be composed of two parts:

$$M_{kn} = M_{kn}^{(W)} + M_{kn}^{(0)}. \tag{2.3}$$

The "weak" mass matrix component $M_{kn}^{(W)}$ is given in terms of the Higgs scalars of W, which will be denoted as μ_k as follows:

$$M_{kn}^{(W)} = M_W^2(\|\mu\|^2\delta_{kn} - \mu_k\mu_n) \tag{2.4}$$

with M_W = const. (We will discuss the μ_k presently.) The other mass matrix component $M_{kn}^{(0)}$ is a *constant* matrix given by

$$M_{kn}^{(0)} = \begin{bmatrix} M_0 & 0 & 0 \\ 0 & M_0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

with M_0 = const. The matrix $M_{kn}^{(0)}$ is used to break the SU(2) symmetry by inhibiting the propagation of W_1^α and W_2^α at long range. To do this M_0 must be massive enough to allow *only* the long-range propagation of the electromagnetic waves W_3^α. However, we assume that M_0^2 is still much smaller than m_w^2 (where m_w is the mass of W) and hence may be neglected when $M_{kn}^{(W)} \neq 0$.

These restrictions on M_0 are easily satisfied in the vector model because of the huge mass of the W particle which is given by $\|\mu\| M_W$ in formula (2.4). We show in Sec. IV that $m_w = 75$ GeV. This prediction for m_w agrees with that predicted by the Weinberg–Salam model.

Unlike the constant smaller mass matrix $M_{kn}^{(0)}$, the "weak" mass matrix $M_{kn}^{(W)}$ defined by (2.4), depends on the Higgs scalars μ_k of W, which depend on the interaction. Like the fermion currents J_k^α, the W boson Noether currents, denoted I_k^α, depend on the Higgs scalars μ_k. For normalized plane waves, we have $I_k^\alpha = -\mu_k v^\alpha$ (where v^α denotes a four-velocity), whereas $J_k^\alpha = -\phi_k v^\alpha$. By the well-known method of integrating $I_k^0 + J_k^0$ over a large volume of space and then applying the four-divergence and Gauss's theorem, we see that $\mu_k + \phi_k$ are additive for interactions that conserve $I_k^\alpha + J_k^\alpha$. For the decay, $a \to b + W$, where a and b denote two fermions, it follows that

$$\phi_k(a) = \phi_k(b) + \mu_k(W), \tag{2.5}$$

where $\phi_k(a)$ and $\phi_k(b)$ are Higgs scalars of the particles a and b and $\mu_k(W)$ are Higgs scalars of the gauge boson W. Note that if a and b are the *same* particles, then from (2.5),

$\mu_k(W) = (0, 0, 0)$, so that, in particular, a photon is assigned zero electric charge, zero lepton number, and zero baryon number. Moreover, from (2.4) we see that $\mu_k = 0$ implies $M^{(W)}_{kn} = 0$, so that the mass matrix M_{kn} reduces to $M^{(0)}_{kn}$.

We see that formula (2.5) defines the W-boson Higgs scalars μ_k in terms of the fermion Higgs scalars $\phi_k(a)$ and $\phi_k(b)$. We assume that the conservation of Noether currents expressed by (2.5) for the transition, $a \to b + W$, is generally valid, regardless of whether the transition is real or vitual. Thus, always

$$\mu_k = \phi_k(a) - \phi_k(b),$$

for the transition $a \to b + W$.

Next, consider the transition Noether currents J^α_k in formula (2.1). In analogy with Eq. (1.6) we propose that these currents have the form

$$J^\alpha_k = -e\hat{\phi}_k j^\alpha(a \to b), \tag{2.6}$$

where $j^\alpha(a \to b)$ denotes the SU(2)×U(1) gauge-invariant transition current for changing particle a into particle b, and where the $\hat{\phi}_\alpha$ denotes the *transition* Higgs scalars, and e denotes the magnitude of the electron charge. We discuss the gauge-invariant current $j^\alpha(a \to b)$ in Sec. III. To restrict the interaction to a short range, it is necessary that J_k be orthogonal to the boson Higgs scalars μ_k. This implies from (2.6) that the $\hat{\phi}_k$ must be orthogonal to the μ_k. Thus, from (2.5) we may set

$$\hat{\phi}_k(a \to b) = \tfrac{1}{2}[\phi_k(a) + \phi_k(b)], \quad \mu_k = \phi_k(a) - \phi_k b), \tag{2.7}$$

which then satisfy all the requirements for leptons.

However, if the particles a and b are hadrons and hence are *composite* of quarks, we define the transition Higgs scalars as

$$\hat{\phi}_k(a \to b) = \frac{1}{2} \sum \phi_k \text{ (quarks in } a \text{ and } b\text{)}. \tag{2.8}$$

Using (8) for hadrons removes a factor $\frac{1}{3}$ which otherwise constantly appears in the following transition matrix (2.11) for quarks. The $\hat{\phi}_k$ and the μ_k are still orthogonal.

Also, for four particle interactions, $a \to b$ and $a' \to b'$, twice as many Higgs scalars must be defined; e.g., for leptons,

$$\hat{\phi}_k(a \to b) = \tfrac{1}{2}[\phi_k(a) + \phi_k(b)],$$
$$\hat{\phi}_k(a' \to b') = \tfrac{1}{2}[\phi_k(a') + \phi_k(b')],$$

and also

$$\mu_k = \phi_k(a) - \phi_k(b), \quad \mu'_k = \phi_k(a') - \phi_k(b'). \tag{2.9}$$

For the interactions which *conserve* the Noether currents

$$\phi_k(a) + \phi_k(a') = \phi_k(b) + \phi_k(b'),$$

and thus from (9), $\mu'_k = -\mu_k$. For this case, the mass matrix M given by (3) and (4) is the *same* for both μ_k and μ'_k and is therefore well defined. In Sec. IV we discuss the mass matrix for strangeness changing interactions which do not conserve ϕ_2.

For four particle interactions, $a \to b$ and $a' \to b'$, the transition matrix T is given by[13]

$$T = cJ^\alpha_k(a \to b)(M - q^2 I)^{-1}_{kn} J_{n\alpha}(a' \leftarrow b'), \tag{2.10}$$

where q_α is the momentum transfer, M is the mass matrix, and c is a normalization factor which equals 2 when *charged* W bosons are exchanged, and equals 1 otherwise. The factor c, which also occurs in the Weinberg–Salam transition matrix,[14] normalizes the phasor representation of charged bosons $W^\alpha_\pm$. The derivation of the transition matrix T given by (10), exactly parallels the same derivation of the Weinberg–Salam transition matrix. The only difference is the form of the mass matrix M and the currents J^α_k.

Substituting (2.4) and (2.6) into (2.10) and using the facts that J^α_k and J'^α_k are orthogonal to μ_k and μ'_k gives

$$T = e^2 c j^\alpha(a \to b) j_\alpha(a' \to b')\hat{\phi}_k \hat{\phi}'_k/(m^2_w - q^2), \tag{2.11}$$

where

$$m_w = M_W \|\mu\| = W\text{-boson mass.}$$

In the familiar case of elastic electromagnetic scattering of two fermions by exchange of photons, for which $\mu_k = \mu'_k = 0$, and for which the electric charges are given by $e\phi_3$ and $e\phi'_3$, formula (2.11) reduces to the following well-known expression for the electromagnetic interaction:

$$T = e^2 \phi_3 \phi'_3 j^\alpha j'_\alpha / -q^2. \tag{2.12}$$

In the next section we derive the gauge-invariant vector and axial transition currents j^α, and compare the resulting transition matrix with the Weinberg–Salam model.

III. TRANSITION CURRENTS

In this section we derive the SU(2)×U(1) gauge-invariant vector and axial transition currents used in the transition matrix T in Sec. II. As discussed in the Introduction, the SU(2)×U(1) gauge action on bispinors is noncomplex linear. However, it may be represented linearly on isotropic Yang–Mills triplets ($\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$) or equivalently on spinor pairs,[15]

$$\Psi = (\xi, \mu).$$

Since spinor pairs can be mapped bijectively to the bispinors,

$$\tilde{\Psi} = (\xi, \mu^*),$$

we may express the *usual* nongauge invariant vector and axial currents in terms of *spinor pairs*, rather than bispinors. We shall generalize these currents to a form which is SU(2)×U(1) gauge invariant.

Consider the usual bispinor vector and axial currents,

$$j^\alpha_v = (\overline{\tilde{\Psi}' \gamma^0})\gamma^\alpha \tilde{\Psi}, \quad j^\alpha_A = (\overline{\tilde{\Psi}' \gamma^0})\gamma_5 \gamma^\alpha \tilde{\Psi}, \tag{3.1}$$

where the $\gamma^0, \ldots, \gamma_5$ are Dirac matrices acting on bispinors $\tilde{\Psi}$ and $\tilde{\Psi}'$. We see from (1) that j^0_v has the form

$$j^0_v = \overline{\tilde{\Psi}'}\tilde{\Psi} = \overline{\xi'}\xi + \overline{\mu'^* \mu^*}, \tag{3.2}$$

where $\tilde{\Psi} = (\xi, \mu^*)$ and $\tilde{\Psi}' = (\xi', \mu'^*)$. By the map $\tilde{\Psi} \to \Psi$, Eq. (3.2) becomes

$$j^0_v = \overline{\xi}'\xi + \overline{\overline{\mu}'\mu} = (\xi', \xi) + \overline{(\mu', \mu)}, \tag{3.3}$$

where we have introduced the inner product notation (,).

Introducing the electron projection operators P_e and Q_e,

$$P_e = \tfrac{1}{2}(I + \tau_3), \quad Q_e = \tfrac{1}{2}(I - \tau_3), \tag{3.4}$$

where

$$\tau_3 = \begin{bmatrix} I_2 & 0 \\ 0 & -I_2 \end{bmatrix} = \text{third gauge matrix},$$

j_v^0 becomes for the electron

$$j_v^0 = (P_e\Psi', P_e\,\Psi) + \overline{(Q_e\Psi', Q_e\Psi)}. \tag{3.5}$$

To generalize (3.5), we write as general projections

$$P_a = \tfrac12 (I + \phi_k(a)\tau_k), \quad Q_a = \tfrac12 (I - \phi_k(a)\tau_k), \tag{3.6}$$

where a denotes an arbitrary particle with Higgs scalars ϕ_k(a). then, we generalize the inner product for spinor pairs as follows:

$$\langle \Psi'_b, \Psi_a \rangle = (P_b\Psi'_b, P_a\,\Psi_a) + \overline{(Q_b\Psi'_b, Q_a\Psi_a)}, \tag{3.7}$$

where Ψ_a and Ψ'_b are spinor pairs associated with two particles, a and b, whose Higgs scalars are $\phi_k(a)$ and $\phi_k(b)$.

We show in Appendix C that the inner product (3.7) is SU(2)$\times$U(1) gauge invariant. The gauge invariant vector and axial transition currents for spinor pairs can then be defined as follows:

$$j_v^\alpha(a\to b) = \langle \Psi'_b, \sigma^\alpha\Psi_a\rangle, \quad j_A^\alpha(a\to b) = \langle \Psi'_b, i\sigma^\alpha\Psi_a\rangle, \tag{3.8}$$

where

$$\sigma^\alpha = (I, \boldsymbol{\sigma}) = \text{Pauli matrices acting on spinor pairs}.$$

The following theorem, which is a straightforward application of definitions (3.6) and (3.7), translates formulas (3.8) into bispinor notation.

Theorem: Let Ψ_a and Ψ'_b be spinor pairs associated with two particles a and b, whose Higgs scalars are $\phi_k(a)$ and $\phi_k(b)$. Set

$$\phi_k(a) = R_3(\delta_a)R_2(\theta_a)\phi_k(e),$$
$$\phi_k(b) = R_3(\delta_b)R_2(\theta_b)\phi_k(e),$$

where $R_2(\theta)$ is the formal gauge rotation through the angle θ about the two axis, $R_3(\delta)$ is the formal gauge rotation through the angle δ about the three-axis, and $\phi_k(e) = (0, 0, 1)$ are the electron Higgs scalars. Define

$$\Psi = R_2(-\theta_a)R_3(-\delta_a)\Psi_a,$$
$$\Psi' = R_2(-\theta_b)R_3(-\delta_b)\Psi'_b.$$

Then, the SU(2)$\times$U(1) gauge-invariant vector and axial currents (3.8) are equal to

$$\begin{aligned} j_v^\alpha(a\to b) &= \cos(\theta_{ab}/2)e^{i\delta/2}(\hat{\tilde\Psi}{}'_b\gamma^\alpha\tilde\Psi_a) \\ j_A^\alpha(a\to b) &= \cos(\theta_{ab}/2)e^{i\delta/2}(\hat{\tilde\Psi}{}'_b\gamma_5\gamma^\alpha\tilde\Psi_a), \end{aligned} \tag{3.9}$$

where $\tilde\Psi_a$ and $\tilde\Psi'_b$ are the bispinors associated with the spinor pairs Ψ and Ψ', $\hat{\tilde\Psi}{}'_b = \overline{\tilde\Psi'_b}\,\gamma^0$, $\delta = \delta_a - \delta_b$, and θ_{ab} is the angle between $\phi_k(a)$ and $\phi_k(b)$.

We see from (9) that

$$j_v^\alpha(e\to e) = \hat{\tilde\Psi}{}'_e\gamma^\alpha\tilde\Psi_e, \quad j_A^\alpha(e\to e) = \hat{\tilde\Psi}{}'_e\gamma_5\gamma^\alpha\tilde\Psi_e$$

reduce to the usual vector and axial transition currents for electrons. Also we see that

$$j_v^\alpha(a\to\bar a) = j_A^\alpha(a\to\bar a) = 0,$$

where $\bar a$ is the antiparticle of a, since $\cos(\theta_{a\bar a}/2) = \cos 90^\circ = 0$. Thus, there are *no* transitions allowed between a particle a and its own antiparticle $\bar a$.

Note that the phasor $e^{i\delta}$ in formulas (3.9) which represents a formal gauge rotation about the three-axis, is an inessential factor, and may be omitted without affecting the transition matrix T. Thus we may define

$$j^\alpha(a\to b) = \cos(\theta_{ab}/2)\tilde j^\alpha(a\to b), \tag{3.10}$$

where $\tilde j^\alpha(a\to b)$ is the linear combination of the *usual* bispinor vector and axial currents $\hat{\tilde\Psi}_b\gamma^\alpha\tilde\Psi_a$ and $\hat{\tilde\Psi}_b\gamma_5\gamma^\alpha\tilde\Psi_a$, then $j^\alpha(a\to b)$ is the SU(2)$\times$U(1) gauge-invariant transition current describing electroweak interactions, and, using Eq. (2.6), the SU(2) transition Noether currents are

$$J_k^\alpha = e\sum_{a,b}\frac{\phi_k(a) + \phi_k(b)}{2}\cos\left(\frac{\theta_{ab}}{2}\right)\tilde j^\alpha(a\to b). \tag{3.11}$$

In the Weinberg–Salam model, the transition matrix for the electromagnetic interactions is the same as the following vector model transition matrix:

$$T = e^2(\tilde j^\alpha\tilde j'_\alpha / -q^2)\phi_3\phi'_3, \tag{3.12}$$

where $\tilde j_\alpha = \hat{\tilde\Psi}_b\gamma_\alpha\tilde\Psi_a$, q_α is the momentum transfer, e is the magnitude of the electron charge, and ϕ_3 is the third Higgs scalar. On the other hand, when charged W bosons are exchanged, the Weinberg–Salam transition matrix is given by[16]

$$T = (g^2/8)[\tilde j^\alpha\tilde j'_\alpha/(m_w^2 - q^2)], \tag{3.13}$$

where $g^2 = e^2/\sin^2\theta_w = 4e^2$, θ_w is the Weinberg angle, $\tilde j_\alpha = \hat{\tilde\Psi}_b(1+\gamma_5)\gamma_\alpha\tilde\Psi_a$, and m_w = W-boson mass.

In the vector model, for ordinary beta decay which is represented by the transitions $d\to u$ and $\nu\to e$, $\cos(\theta_{du}/2) = \cos(\theta_{\nu e}/2) = 1/\sqrt2$, $\hat\phi_k\hat\phi'_k = \frac12$, and so, in the vector model,

$$T = (e^2/2)[\tilde j^\alpha\tilde j'_\alpha/(m_w^2 - q^2)], \tag{3.14}$$

which agrees with (3.13) for the Weinberg–Salam model.

From (3.14) we obtain Fermi's constant G_F as follows:

$$G_F/\sqrt2 = e^2/2m_w^2$$

so that the W-boson mass is given by

$$m_w = e/2^{\frac14}G_F^{\frac12} = 75\ \text{GeV}.$$

IV. PREDICTION OF THE CABIBBO ANGLE

As previously mentioned in Sec. I, there are *two equivalent* ways for representing the Higgs scalars for the up and down quarks. Namely,

$$\phi_k(u) = (\tfrac13 - \tfrac23, -\tfrac23), \quad \phi_k(d) = (-\tfrac23, -\tfrac23, \tfrac13), \tag{4.1}$$

and also,

$$\phi_k(u) = (\tfrac13, \tfrac23, -\tfrac23), \quad \phi_k(d) = (-\tfrac23, \tfrac23, \tfrac13). \tag{4.2}$$

To extend the vector model to the charmed and strange quarks, we arbitrarily chose one set of Higgs scalars for the up and down quarks, and the other set for the charmed and strange quarks.

However, this assignment of Higgs scalars to the strange and charmed quarks leads to the nonconservation of the Higgs scalar ϕ_2 during transitions between quark families. A look at the strangeness changing beta decay, $s\to u + e + \bar\nu$, shows that

$$\phi_2(s) + \phi_2(\nu) \neq \phi_2(u) + \phi_2(e). \tag{4.3}$$

Thus the Higgs scalar ϕ_2 is *not* conserved during strangeness changing interactions (whereas ϕ_1 and ϕ_3 are still conserved) and the Lagrangian describing the strangeness changing interactions, is *not* invariant under formal gauge rotations about the two-axis. As a consequence the strangeness changing beta decay $s \to u + e + \bar{\nu}$ requires a mixed state of W and W' bosons having different Higgs scalars $\mu_k \neq \mu'_k$. (See Fig. 1.) For transitions within the *same* family (e.g., $d \to u$), $\mu_k = \mu'_k$ and the mass matrix M for the W bosons is unique [see formula (2.4)]. However we propose that in the mixed state, the masses of W and W' are *mixed* in the *usual* way for bosons[17] which implies that the mass matrix equals the average of the mass matrices, i.e., $\frac{1}{2}(M + M')$. Using this mass matrix, leads to the following transition matrix [see formula (2.1)]:

$$T = 2e^2 j^\alpha(s \to u) j_\alpha(\nu \to e) \hat{\phi}_k \hat{\phi}'_k / [\tfrac{1}{2}(m_W^2 + m_{W'}^2) - q^2], \tag{4.4}$$

with

$$m_W = M_W \|\mu\|.$$

Thus, the squared mass m_W^2 has become the average squared mass of W and W', which agrees with Feynman's rule for mixing boson masses.

In formula (4), let us replace the particles s, u, ν, and e with arbitrary fermion particles a, b, a', and b', and consider the four fermion interactions $a \to b$ and $a' \to b'$. The Higgs scalars of $a,\dots,b'$ are denoted by $\phi_k(a),\dots,\phi_k(b')$ as before. Formula (4.4) can be put in a form that explicitly shows the Higgs scalars. Since from formula (2.9),

$$\mu_k = \phi_k(a) - \phi_k(b),$$
$$\|\phi(a)\| = \|\phi(b)\| = 1,$$

we derive that

$$\|\mu\| = 2\sin(\theta/2),$$

where θ is the angle between the formal vectors $\phi_k(a)$ and $\phi_k b)$. Similarly, we have

$$\|\mu'\| = 2\sin(\theta'/2),$$

where θ' is the angle between the formal vectors $\phi_k(a')$ and $\phi_k(b')$. Using the gauge-invariant currents from Sec. III,

$$j_\alpha = \cos(\theta/2)\tilde{j}_\alpha, \quad j'_\alpha = \cos(\theta'/2)\tilde{j}'_\alpha,$$

we get the result

$$T = 4e^2 \frac{\cos(\theta/2)\cos(\theta'/2)\hat{\phi}_k \hat{\phi}'_k \tilde{j}^\alpha \tilde{j}'_\alpha}{m_W^2[\sin^2(\theta/2) + \sin^2(\theta'/2)] - q^2}. \tag{4.5}$$

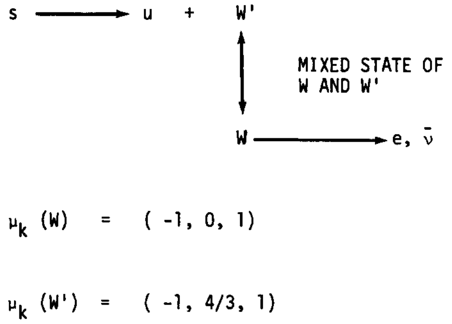


FIG. 1. Strangeness changing beta decay.

Neglecting the momentum transfer q_α gives

$$\mathrm{T} = \frac{\cos(\theta/2)\cos(\theta'/2)\hat{\phi}_k \hat{\phi}'_k}{\sin^2(\theta/2) + \sin^2(\theta'/2)}$$

$\times$ factors that do not depend on Higgs scalars. (4.6)

The *ratio* of the transition matrices T for the strange and nonstrange beta decay is defined to be $\tan\theta_c$; i.e., the tangent of the Cabibbo angle.

We now show how formula (6) predicts the Cabibbo angle. Consider first the ordinary beta decay $d \to u + e + \bar{\nu}$. From Table I, we derive $\hat{\phi}_k \hat{\phi}'_k = -\frac{1}{6}$, $\cos(\theta/2) = \cos(\theta'/2) = 1\sqrt{2}$. Consequently, from (4.6),

$$T \approx -\tfrac{1}{12}. \tag{4.7}$$

Similarly, for the strange beta decay $s \to u + e + \bar{\nu}$, we get

$$\hat{\phi}_k \hat{\phi}'_k = -\tfrac{1}{6}, \quad \cos(\theta/2) = 1/\sqrt{18}, \quad \cos(\theta'/2) = 1/\sqrt{2},$$

and consequently, from (4.6),

$$T \approx -\tfrac{1}{12} \cdot \tfrac{3}{13}. \tag{4.8}$$

The ratio of (4.8) and (4.7) shows that the Cabibbo angle θ_c satisifies

$$\tan\theta_c = \tfrac{3}{13},$$

or equivalently, $\cos\theta_c = 0.9744$ which is very close to the measured value[18] $\cos\theta_c = 0.9737$.

APPENDIX A: SL(2, C)×U(1) GAUGE INVARIANT LAGRANGIAN, DIRAC EQUATION, AND NOETHER CURRENTS

In Fig. 2 we depict the bijection between bispinors $\tilde{\Psi}$ and spinor pairs Ψ, as well as the Cartan map isomorphisms of $\tilde{\Psi}$ and Ψ with isotropic Yang–Mills vector triplets $(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)$. These three isomorphisms were discussed in our previous paper.[1] We use *spinor pairs* rather than the equivalent bispinors in writing the fermion Lagrangian, which is

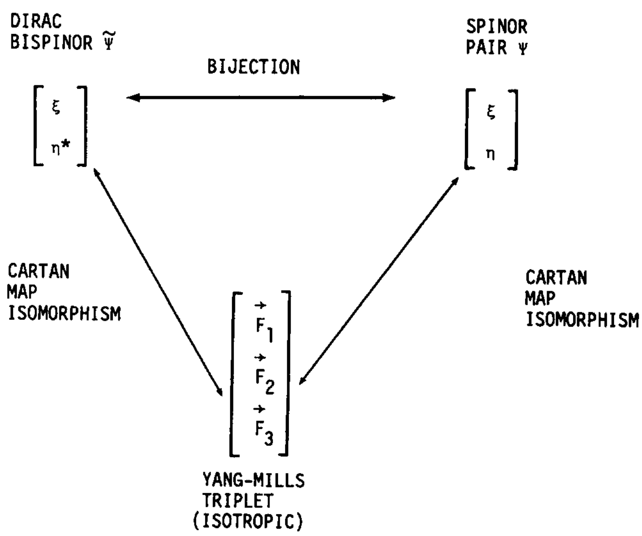


FIG. 2. Isomorphisms.

given in formula (A5) below. Whereas, SL(2, C)$\times$U(1) gauge transformations act on the spinor pairs (and also on the isotropic Yang–Mills triplets) via complex matrices, they are *not* complex linear on the bispinors, which obfuscates the SL(2, C)$\times$U(1) structure of the bispinors. Let

$$\xi = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \in C^2$$

denote a spinor. The conjugate spinor associated with ξ is defined to be

$$\xi^* = \begin{bmatrix} \bar{\xi}_2 \\ -\bar{\xi}_1 \end{bmatrix} \in C^2,$$

where the bar denotes complex conjugation. The map $\xi \to \xi^*$ is a bijection, since $\xi = -\xi^{**}$.

A bispinor $\widetilde{\Psi} = (\xi, \eta^*) \in C^4$ consists of a spinor $\xi \in C^2$ and a conjugated spinor $\eta^* \in C^2$. The generators of the electromagnetic and neutral gauge transformations acting on bispinors are, respectively, I and γ_5, where I is a 4×4 identity matrix and

$$\gamma_5 = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

Associated with each bispinor $\widetilde{\Psi} = (\xi, \eta^*)$ is a spinor pair $\Psi = (\xi, \eta)$, where now both ξ and η are spinors. The conjugate spinor pair is defined to be $\Psi^* = (\eta^*, -\xi^*)$. The maps $\widetilde{\Psi} \to \Psi$ and $\widetilde{\Psi} \to \Psi^*$ are *well-defined* bijections because, as noted above, the maps $\xi \to \xi^*$ and $\eta \to \eta^*$ are bijections.

Electromagnetic gauge transformations become part of a larger SL(2, C) gauge group as follows: Define 4×4 matrices $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$,

$$\tau_1 = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}, \quad \tau_2 = \begin{bmatrix} 0 & -iI \\ iI & 0 \end{bmatrix}, \quad \tau_3 = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix},$$

where τ_1, τ_2, and τ_3 are the generators of the gauge subgroup SU(2). By the map $\widetilde{\Psi} \to \Psi$, the electromagnetic gauge generator becomes τ_3, so that electromagnetic gauge transformations become the formal "rotations" about the three-axis. Whereas for bispinors, the gauge group is restricted to just U(1)$\times$U(1), the gauge group for spinor pairs is the much larger group SL(2, C)$\times$U(1).

The Dirac equation

$$D_\alpha \sigma^\alpha \Psi = -M\phi_5 \phi_\beta \hat{\tau}^\beta \Psi^* \tag{A1}$$

is invariant under *both* Lorentz and SL(2, C)$\times$U(1) gauge transformations, where the D_α are the Yang–Mills derivatives, $\sigma^\alpha = (I, \boldsymbol{\sigma})$ are the Pauli spin-half matrices, $\hat{\tau}^\beta = (I, -\boldsymbol{\tau})$ are the gauge matrices [note that $\hat{\tau}^\beta = (I, \boldsymbol{\tau})$], M is the mass, ϕ_α and ϕ_5 are the Higgs scalars, $\Psi = (\xi, \eta)$ is a spinor pair, $\Psi^* = (\eta^*, -\xi^*)$ is the conjugate of Ψ, and the Higgs scalars are normalized as follows:

$$\phi_\beta \phi^\beta = -1, \quad |\phi_5| = 1. \tag{A2}$$

Solutions of Eq. (A1) also satisfy the Klein–Gordon equation, which, in the case of free particles, is given by

$$D^\alpha D_\alpha \Psi = M^2 \Psi. \tag{A3}$$

Equation (A1) is equivalent to the usual Dirac equation when ϕ_β and ϕ_5 are chosen by convention to be

$$\phi_\beta = (0, 0, 0, 1), \quad \phi_5 = 1. \tag{A4}$$

Equation (A1) is the Euler–Lagrange equation for the following Lagrangian:

$$L = \mathrm{Re}\left[\phi_\beta\, \overline{\tau^\beta \Psi} \cdot D_\alpha \sigma^\alpha \Psi + \phi_5 M \overline{\Psi}^* \cdot \Psi\right]. \tag{A5}$$

The Lagrangian (A5) is a scalar invariant under *both* Lorentz and SL(2, C)$\times$U(1) gauge transformations. Furthermore (A5) reduces to the usual (non-gauge-invariant) fermion Lagrangian when the Higgs scalars ϕ_β and ϕ_5 assume the conventional values given in (A4). From (A5) one easily derives the seven real Noether currents [one for each generator of SL(2, C)$\times$U(1)], which are given as follows:

$$\begin{aligned} &J_5^\alpha = \phi^\beta j_\beta^\alpha, \quad \mathrm{Re}\, J_k^\alpha = \phi_0 j_k^\alpha - \phi_k j_0^\alpha, \\ &\mathrm{Im}\, J_k^\alpha = -\epsilon_{kmn} \phi_m j_n^\alpha, \end{aligned} \tag{A6}$$

with $\alpha, \beta = 0,1,2,3$ and $k, m, n = 1,2,3$, and

$$j_\beta^\alpha = \overline{\tau_\beta \Psi} \cdot \sigma^\alpha \Psi. \tag{A7}$$

APPENDIX B: *CPT* INVARIANCE

To define the operators C, P, and T for bispinors, we first define the 4×4 Dirac matrices,

$$\gamma_0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \gamma_k = \begin{bmatrix} 0 & -\sigma_k \\ \sigma_k & 0 \end{bmatrix},$$

for $k = 1,2,3$. Then for a bispinor field $\widetilde{\Psi}(\mathbf{x}, t)$, we define

$$\begin{aligned} &C\widetilde{\Psi}(\mathbf{x}, t) = \gamma_2 \overline{\widetilde{\Psi}(\mathbf{x}\ \ t)}, \\ &P\widetilde{\Psi}(\mathbf{x}, t) = \gamma_0 \widetilde{\Psi}(-\mathbf{x}, t), \\ &T\widetilde{\Psi}(\mathbf{x}, t) = \gamma_1 \gamma_3 \overline{\widetilde{\Psi}(\mathbf{x}, -t)}, \end{aligned} \tag{B1}$$

and

$$CPT\widetilde{\Psi}(\mathbf{x}, t) = i\gamma_5 \widetilde{\Psi}(-\mathbf{x}, -t). \tag{B2}$$

Other definitions can be given for C, P, and T, but they differ from (B1) by an electromagnetic gauge transformation; i.e., they differ by a formal "rotation" about the three-axis.

The role of C, P, T in the vector model is summarized in the following lemma and corollary. The proof of the lemma is a straightforward application of (B1) across the map $\widetilde{\Psi} \to \Psi$ (see Sec. III).

Lemma 1: Let Ψ be a spinor pair. The C, P, T operators acting on Ψ are given by

$$\begin{aligned} &C\Psi(\mathbf{x}, t) = U_C \Psi(\mathbf{x}, t), \\ &P\Psi(\mathbf{x}, t) = U_P \Psi^*(-\mathbf{x}, t), \\ &T\Psi(\mathbf{x}, t) = U_T \Psi^*(\mathbf{x}, -t), \end{aligned}$$

where

$$U_C = e^{-i\tau_2\pi/2} \cdot e^{-i\pi/2}, \quad U_P = I, \quad U_T = e^{i\tau_2\pi/2}, \tag{B3}$$

with $e^{-i\tau_2\pi/2}$, the formal "rotation" of 180° about the two-axis, and $e^{-i\pi/2}$, the 180° neutral gauge transformation.

Corollary: The *CPT* operator acting on a spinor pair Ψ is given by

$$CPT\Psi(\mathbf{x}, t) = e^{-i\pi/2} \Psi(-\mathbf{x}, -t). \tag{B4}$$

We see from (B3) that neither C nor T commute with all SL(2, C) gauge transformations. Neither P nor T commute with neutral gauge transformations. However, the operator *CPT* commutes with *all* SL(2, C)$\times$U(1) gauge transformations. Consequently, the theory is invariant under *CPT*, but not invariant under C, P, or T individually, or in the combinations CP, CT, or PT.

APPENDIX C: AN SU(2)×U(1) GAUGE-INVARIANT INNER PRODUCT

In Sec. III we defined an inner product, denoted $\langle\ ,\ \rangle$, as follows:

$$\langle \Psi'_b, \Psi_a \rangle = (P_b \Psi'_b, P_a \Psi_a) + \overline{(Q_b \Psi'_b, Q_a \Psi_a)}, \qquad \text{(C1)}$$

where Ψ_a and Ψ'_b are spinor pairs associated with two particles, a and b, whose Higgs scalars are $\phi_k(a)$ and $\phi_k(b)$. Also, P_a, P_b, Q_a, and Q_b denote the four projection operators

$$P_a = \tfrac{1}{2}(I + A), \quad P_b = \tfrac{1}{2}(I + B),$$
$$Q_a = \tfrac{1}{2}(I - A), \quad Q_b = \tfrac{1}{2}(I - B), \qquad \text{(C2)}$$

with

$$A = \phi_k(a)\tau_k, \quad B = \phi_k(b)\tau_k, \qquad \text{(C3)}$$

where τ_k are the three gauge matrices acing on spinor pairs, and I is the identity matrix, i.e.,

$$I = \begin{bmatrix} I_2 & 0 \\ 0 & I_2 \end{bmatrix},$$

$$\tau_1 = \begin{bmatrix} 0 & I_2 \\ I_2 & 0 \end{bmatrix}, \quad \tau_2 = i\begin{bmatrix} 0 & -I_2 \\ I_2 & 0 \end{bmatrix}, \quad \tau_3 = \begin{bmatrix} I_2 & 0 \\ 0 & -I_2 \end{bmatrix}.$$

Moreover, we denote by $(\ ,\ )$ the usual inner product,

$$(\Psi, \chi) = \overline{\Psi} \cdot \chi,$$

where Ψ, χ are spinor pairs and $\overline{\Psi}$ is the complex conjugate of Ψ.

In this Appendix, we show that the inner product $\langle\ ,\ \rangle$ given by formula (C1) is invariant (i.e., transforms as a scalar) under SU(2)×U(1) gauge transformations. The gauge invariant properties of the inner product (C1) will be summarized in the final lemma at the end of the Appendix.

The inner product (C1) is defined on two Hilbert spaces $\mathscr{H}_a$ and $\mathscr{H}_b$ of solutions (e.g., Ψ_a and Ψ_b) to a pair of Dirac equations. These are two *different* Hilbert spaces if the particles a and b have different Higgs scalars. The Hilbert space $\mathscr{H}_a$ contains the solutions of the Dirac equation

$$D^\alpha \sigma_\alpha \Psi = m\phi_5 \phi^k(a)\tau_k \Psi^*, \qquad \text{(C4)}$$

whereas the Hilbert space $\mathscr{H}_b$ contains the solutions of the Dirac equation

$$D^\alpha \sigma_\alpha \Psi = m\, \phi_5 \phi^k(b)\tau_k \Psi^*, \qquad \text{(C5)}$$

which are distinct from the solutions of (C4). An important difference between the Hilbert spaces $\mathscr{H}_a$ and $\mathscr{H}_b$ is that the solutions in $\mathscr{H}_a$ are transformed into other solutions in $\mathscr{H}_a$ by formal gauge rotations about the $\phi_k(a)$ axis; whereas the solutions in $\mathscr{H}_b$ are transformed by formal gauge rotations about the $\phi_k(b)$ axis. Thus, we may define two generators of SU(2), denoted A and B as in formulas (3): $A = \phi_k(a)\tau_k$, which is the generator of formal gauge rotations about the $\phi_k(a)$ axis; and $B = \phi_k(b)\tau_k$, which is the generator of formal gauge rotations about the $\phi_k(b)$ axis. The Hilbert space $\mathscr{H}_a$ is invariant under the formal gauge rotations [about the $\phi_k(a)$ axis] generated by the generator A; whereas $\mathscr{H}_b$ is invariant under B. Ordinary multiplication by complex scalars, however, is the neutral U(1) gauge action which does *not* leave $\mathscr{H}_a$ and $\mathscr{H}_b$ invariant. Therefore, complex scalar multiplication for the Hilbert spaces $\mathscr{H}_a$ and $\mathscr{H}_b$ must differ from ordinary complex scalar multiplication. Indeed, let $c = s + it$ be a complex number, then

$$c \cdot \Psi_a = (sI + itA)\Psi_a, \quad c \cdot \Psi_b = (sI + itB)\Psi_b \qquad \text{(C6)}$$

defines the scalar multiplications for the Hilbert spaces $\mathscr{H}_a$ and $\mathscr{H}_b$. For example, if the scalar $c = s$ is real, then $s \cdot \Psi_a = s\Psi$, i.e., $\mathscr{H}_a$ is a vector space over the reals in the usual way. However, if $c = e^{i\theta}$ is a phasor, then from (C6), the scalar multiplication

$$e^{i\theta} \cdot \Psi_a = e^{iA\theta}\Psi_a$$

acts by a formal gauge rotation about the $\phi_k(a)$ axis, which as we have seen leaves $\mathscr{H}_a$ invariant. Whereas, ordinary complex scalar multiplication by $e^{i\theta}$ will *not* map solutions in $\mathscr{H}_a$ to other solutions in $\mathscr{H}_a$.

One may show then that with the scalar multiplications on $\mathscr{H}_a$ and $\mathscr{H}_b$ defined by (C6), the inner product (C1) satisfies the usual axioms of invariant Hermitian bilinear forms from $\mathscr{H}_a \times \mathscr{H}_b$ into C.

Lemma: The inner product $\langle\ ,\ \rangle$ defined in (1) satisfies the following: (a) $\langle\ ,\ \rangle$ is a Hermitian bilinear form on $\mathscr{H}_a \times \mathscr{H}_b$; (b) if R is a gauge transformation in SU(2) (i.e., a formal gauge rotation), then

$$\langle R\, \Psi'_b, R\Psi_a \rangle = \langle \Psi'_b, \Psi_a \rangle,$$

(c) if $e^{i\chi}$ denotes a neutral U(1) gauge transformation, then

$$\langle e^{i\chi}\Psi'_b, e^{i\chi}\Psi_a \rangle = \langle \Psi'_b, \Psi_a \rangle.$$

[1]F. Reifler, "A vector model for electroweak interactions," J. Math. Phys. **26**, 542 (1985). A partial summary is given in Appendix A.

[2]If $\zeta = (\zeta_1, \zeta_2) \in C^2$ is a spinor, its conjugate is defined by $\zeta^* = (\bar{\zeta}_2, -\bar{\zeta}_1)$. Note that $\zeta = -\zeta^{**}$, so that the conjugation operation has an inverse.

[3]The scalar ρ is uniquely defined by setting $\rho = (\mathbf{F}_1 \times \mathbf{F}_2 \cdot \mathbf{F}_3)/\lambda$, where $\mathbf{F}_j \cdot \mathbf{F}_k = \lambda\delta_{jk}$.

[4]See Ref. 1. See also, F. Reifler, "A vector wave equation for Neutrinos," J. Math. Phys. **25**, 1088 (1984).

[5]ϕ_5, which multiplies the mass term, is an addition to the model previously presented to make both the wave equation and Lagrangian invariant under neutral gauge transformations. (See Appendix A.)

[6]Reference 1, Sec. I, Eq. (8). See also Appendix A.

[7]Reference 1, Sec. I, Eq. (3).

[8]K. Huang, *Quarks, Leptons, and Gauge Fields* (World Scientific, Singapore, 1982), p. 109.

[9]The symbols e, ν, u, and d denote the electron, neutrino, up, and down quarks. Their antiparticles are denoted by $\bar{e}$, $\bar{\nu}$, $\bar{u}$, and $\bar{d}$.

[10]See Ref. 1. See also (1) and (2) in Sec. IV.

[11]See Appendix A.

[12]See Ref. 8, p. 84.

[13]I. Aitchison and A. Hey, *Gauge Theories in Particle Physics* (Adam Hilger, Bristol, Great Britain, 1983), p. 115.

[14]See Ref. 13, p. 244.

[15]See Appendix A.

[16]See Ref. 13, p. 252.

[17]W. M. Gibson and B. R. Pollard, *Symmetry Principles in Elementary Particle Physics* (Cambridge U.P., New York, 1980), pp. 290–294.

[18]C. Quigg, *Gauge Theories of the Strong, Weak, and Electromagnetic Interactions* (Benjamin, Reading, MA, 1983), p. 150.

Boson basis realization for shell-model fermion problem

Y. K. Gambhir, R. S. Nikam, C. R. Sarma, and J. A. Sheikh
Department of Physics, Indian Institute of Technology, Powai, Bombay-400 076, India

(Received 7 September 1984; accepted for publication 25 January 1985)

The boson basis for the fermion shell-model problem has been realized over a representation space $V_{n(n-1)/2}$ of the unitary group $U(n(n-1)/2)$. A contraction of this space was found to yield the generator algebra of $U(n)$. Based on this result it is shown that the transformations induced by the Dyson-mapped boson Hamiltonian on the Dyson boson basis are identical to the corresponding transformations by the fermion Hamiltonian on the shell-model basis.

I. INTRODUCTION

Group theoretic approaches to physical problems depend mainly on the fact that the Hamiltonian of the system can be expressed as a polynomial in the generators of the invariance group of the Hamiltonian. This is exemplified, for example, by the many recent unitary group approaches (UGA) to single-particle shell-model studies in atoms,[1] molecules,[2,3] and nuclei.[4] In recent years many nuclear structure studies have been undertaken in which a bosonlike character has been attributed to paired fermion states[5–8] through a suitable mapping procedure. The need for such a mapping arises because of the fact that the paired fermion creation $c_i^+ c_j^+$ and destruction $c_i c_j$ operators do not exhibit a bosonlike character but are particle nonconserving operators of the rotation group SO($2n$). Both unitary[5,9] and nonunitary[6,9] mappings have been used. The unitary mappings are infinite boson expansions while nonunitary mappings are finite ones. The nonunitary mapping of bifermion operators is achieved through the generalized Dyson boson mapping (DBM).[6] Hence this mapping leads to a nonunitary basis and does not retain the Hermiticity properties of the original fermion problem. This then raises the question whether a breakdown of the unitary symmetry of the system has taken place during the mapping or whether this aspect is contained in the formulation based on the use of unitary group generators. We have attempted to analyze this problem in the present note, starting with a group $U(n(n-1)/2)$ defined over a basis set of $n(n-1)/2$ skew-symmetric second-rank boson operators. A contraction of the generators of this group is found to yield the Lie algebra of $U(n)$. Using these operators and the boson operator realization of the antisymmetrized physical states, it is shown that the effects of these generators on the states are identical to the ones obtained using DBM. It has also been shown that the DBM matrix elements of the boson-mapped Hamiltonian (H_D) are identical, to within a multiplicative factor, to the corresponding matrix elements of the fermion Hamiltonian (H_F). This multiplicative factor has been found[10,11] to depend solely on the prescription used for normalizing the basis states. The present studies along with the normalization prescription have been found identical to solving the original shell-model problem.

The general algebra of the method is developed in Sec. II and a brief discussion is presented in Sec. III.

II. BOSON OPERATOR METHOD AND THE UNITARY GROUP

Consider the ordered set of $n(n-1)/2$ second-rank tensors defined as

$$\{\phi_{ij};\phi_{ij}=-\phi_{ji}\,|\,i,j=1,\dots,n\},\tag{1}$$

spanning the space $V_{n(n-1)/2}$ and admitting the metric

$$(\phi_{ij}|\phi_{pq})=\Delta_{ip;jq},$$

where

$$\Delta_{ip;jq}=\delta_{ip}\delta_{jq}-\delta_{iq}\delta_{jp}.\tag{2}$$

We now realize this tensor basis as

$$|\phi_{ij})=B^+_{ij}|0),\tag{3}$$

where

$$\{B_{ij},B^+_{ij},B_{ij}=-B_{ji},\;B^+_{ij}=-B^+_{ji},\;|i,j=1,\dots,n\}\tag{4}$$

are simple boson operators satisfying

$$[B_{ij},B_{pq}]=[B^+_{ij},B^+_{pq}]=0,\tag{5}$$

$$[B_{ij},B^+_{pq}]=\Delta_{ip;jq},\tag{6}$$

subject to

$$B_{ij}|0)=(0|B^+_{ij}=0.\tag{7}$$

Using this realization it is now possible to define a set of shift operators on $V_{n(n-1)/2}$ as

$$E_{ij;pq}=B^+_{ij}B_{pq}.\tag{8}$$

Using the commutation relations of Eqs. (5) and (6) it readily follows that

$$[E_{ij;pq},B^+_{rs}]=\Delta_{pr;qs}B^+_{ij},\tag{9}$$

so that they act as shift operators on $V_{n(n-1)/2}$. It can also be seen that these operators define a Lie algebra of $U(n(n-1)/2)$ since

$$[E_{ij;pq},E_{km;rs}]=\Delta_{pk;qm}E_{ij;rs}-\Delta_{ir;js}E_{km;pq}\tag{10}$$

and

$$E^+_{ij;pq}=E_{pq;ij}.\tag{11}$$

Consider now the rth-rank tensor space $V_{n(n-1)/2}\otimes^r$ spanned by

 0022-2488/85/082067-05$02.50

$$\{\phi_{12}^{N_{12}}\phi_{\phi_{13}}^{N_{13}}\cdots\phi_{n-1\,n}^{N_{n-1\,n}} \equiv (B_{12}^{+})^{N_{12}}\cdots(B_{n-1\,n}^{+})^{N_{n-1\,n}}|0)\}, \tag{12}$$

where

$$\sum_{i<j=1}^{n} N_{ij} = r. \tag{13}$$

In view of Eq. (9) we find that only double excitations are permitted on this space of rth-rank products in which an index pair (pq) (for $N_{pq} \neq 0$) is replaced symmetrically by another pair (ij). Since configurations differing by single-index permutations such as $B_{ij}^{+}B_{pq}^{+} \leftrightarrow B_{ip}^{+}B_{jq}^{+}$ cannot be related by these generators, the only representation possible on this space is the totally symmetric one, which is of rank r in these boson operators. This representation is possible because the monomials defined by Eq. (12) are completely symmetric under all interchanges of paired indices. The above group space is too large and irreducible to be of any practical utility. It is thus necessary to use a subgroup restriction to U(n) in order to be able to reduce this space further. We define this restriction through the n^2 generators $\{E_{ij}|i,j=1,\ldots,n\}$ of U(n) defined using Eq. (8) as

$$E_{ij} = \sum_{p=1}^{n} E_{ip;jp} = \sum_{p=1}^{n} B_{ip}^{+}B_{jp}. \tag{14}$$

Combining the results of Eqs. (10) and (14) now leads to

$$[E_{ij}, E_{pq}] = \delta_{jp}E_{iq} - \delta_{iq}E_{pj}. \tag{15}$$

We also find that

$$E_{ij}^{+} = E_{ji}. \tag{16}$$

Thus E_{ij} of Eq. (14) define a Lie algebra of U(n). Their action on the space $V_{n(n-1)/2}$ follows readily on using the commutation relations of the boson operators as

$$[E_{ij}, B_{pq}^{+}] = \delta_{jp}B_{iq}^{+} + \delta_{jq}B_{ip}^{+}. \tag{17}$$

For all $n>3$, Eq. (14) defines a nontrivial restriction of U($n(n-1)/2$) to the subgroup U(n). A consequence of this brought out by Eq. (17) is that all tensor monomials of the set defined by Eq. (12) and differing by a single index interpair permutation can now be linked using the generators E_{ij}. Thus, for example, we have

$$E_{13}B_{12}^{+}B_{34}^{+}|0) = -E_{13}B_{14}^{+}B_{23}^{+}|0) = B_{12}^{+}B_{14}^{+}|0),$$

where the antisymmetry and the boson character of the B^{+}'s have been used to obtain the above result. Thus, if we consider a subset of distinct boson monomials of Eq. (12) differing from each other by single index interpair permutations of indices, they could also be linked through the action of polynomials of the generators E_{ij} of U(n). This distinct subset is generated by applying the permutations $P \in S_{2r}/S_r \otimes (S_2)^r$ of S_{2r} with respect to the subgroup of symmetric interpair interchanges (due to the boson character) and antisymmetric intrapair interchanges (due to skew symmetry) of the defining boson operators. This implies that linear combinations of such monomials have to be chosen spanning irreducible subspaces under U(n). A first step in this process of reduction is to define a reference monomial among the set related by single index interpair permutations. This is done by introducing an ordering on the monomials of Eq. (12) by assuming that, for any index pairs $\alpha=(ij)$, $\beta=(pq)$ such that $i<j$ and $p<q$, respectively, we have

$$\alpha<\beta, \quad \text{for all } i<p,$$

$$\alpha<\beta, \quad \text{for all } i=p, \text{ if } j<q.$$

Using the above ordering and neglecting explicit mention of all those index pairs α for which $N_\alpha=0$ we reexpress Eq. (12) as

$$V_{n(n-1)/2}{}^{\otimes r}: \Big\{ \,|\,\phi_{\alpha_1}^{N_{\alpha_1}}\phi_{\alpha_2}^{N_{\alpha_2}}\cdots\phi_{\alpha_r}^{N_{\alpha_r}}) \equiv (B_{\alpha_1}^{+})^{N_{\alpha_1}}(B_{\alpha_2}^{+})^{N_{\alpha_2}}\cdots(B_{\alpha_r}^{+})^{N_{\alpha_r}}|0)|,$$
$$\alpha_k \equiv (i_k\, j_k);\ i_k<j_k;\ (i_1 j_1), \ldots, (i_r j_r) \in (1,2,\ldots,n);\ \sum_k N_{\alpha_k} = r\Big\}. \tag{18}$$

For a given distribution of N_α the above sets are chosen as the linearly independent ones among those related by the set of coset permutations $P\in S_{2r}/S_r \otimes (S_2)^r$. A reduction of tensor spaces into irreducible subspaces $[\lambda]\equiv[\lambda_1\lambda_2\cdots\lambda_n]$ of U(n) using symmetrized Wigner operators of the permutation group is a well-studied problem.[12,13] We can adapt the same procedure to the monomials of Eq. (18) to obtain the basis spanning the irreducible representations of U(n). In the present context we restrict ourselves to a system of $2r$ identical fermions so that we need only generate a totally antisymmetric representation of the group. This can be done by applying the antisymmetrizer

$$\mathcal{A} = (1/\sqrt{(2r)!}) \sum_{P\in S_{2r}} (-)^{P} P \tag{19}$$

on all possible index locations in the monomials on the right of Eq. (18). In view of the antisymmetric nature of the boson operators we need only apply the normalized form

$$\mathcal{A} = \sqrt{2^r r!/(2r)!} \sum_{\{P\in S_{2r}/S_r\otimes(S_2)^r\}} (-)^{P} P \tag{20}$$

to the monomials on the right side of Eq. (18) in order to generate antisymmetric linear combinations. It is to be noted that if any $N_{\alpha_i}>1$ in Eq. (18), the antisymmetrizer annihilates the monomial. This follows since any such pair of indices implies symmetry while the antisymmetrizer is totally antisymmetric under all index location permutations. Thus, only $N_{\alpha_i}=1$ is allowed for all α_i and, in addition, no two distinct pairs $\alpha_k=(i_k,j_k)$ and $\alpha_m=(i_m,j_m)$ can have any common index. The possible linearly independent configurations as restricted above can be visualized using the example $n=5$, $r=2$, for which $(\alpha_1)=(12)\Rightarrow(\alpha_2)=(34)$, (35), (45); $(\alpha_1)=(13)\Rightarrow(\alpha_2)=(45)$; and $(\alpha_1)=(23)\Rightarrow(\alpha_2)=(45)$. Starting with these $\binom{n}{r}/r$ monomials, which are not related by any interpair or intrapair permutations, and applying $\mathcal{A}$ as defined in Eq. (20) to them we can readily generate a complete set of paired boson states. In applying $\mathcal{A}$ to those monomials, it is convenient to use a form for this operator expressed in terms of products of transpositions as

$$\mathcal{A} = \prod_{k=1}^{r-1} (2r - 2k + 1)^{-1/2}$$

$$\times \left\{ (j_k, j_k) - \sum_{m=k}^{r-1} [(j_k, i_{m+1}) + (j_k, j_{m+1})] \right\}, \tag{21}$$

where $(j_k, j_k) = e$ for all $k = 1, ..., r$. As an illustration of this form of the operator consider the case $r = 3$, for which Eq. (21) reduces to

$$\mathcal{A} = (1/\sqrt{5.3})[e - (j_1, i_2) - (j_1, i_3) - (j_1, j_2) - (j_1, j_3)][e - (j_2, i_3) - (j_2, j_3)].$$

This operator acting on $B^+_{i_1j_1} B^+_{i_2j_2} B^+_{i_3j_3}|0)$ can be readily shown to yield a 15-term normalized antisymmetrized linear combination as

$$\begin{aligned}
\mathcal{A} B^+_{i_1j_1} B^+_{i_2j_2} B^+_{i_3j_3}|0) = (1/\sqrt{15})\{ & B^+_{i_1j_1} B^+_{i_2j_2} B^+_{i_3j_3} - B^+_{i_1i_2} B^+_{j_1j_2} B^+_{i_3j_3} - B^+_{i_1j_2} B^+_{i_2j_1} B^+_{i_3j_3} \\
& - B^+_{i_1i_3} B^+_{i_2j_2} B^+_{j_1j_3} - B^+_{i_1j_3} B^+_{i_2j_2} B^+_{i_3j_1} - B^+_{i_1j_1} B^+_{i_2j_3} B^+_{i_3j_2} + B^+_{i_1i_2} B^+_{j_1j_3} B^+_{i_3j_2} \\
& + B^+_{i_1j_2} B^+_{i_2j_3} B^+_{i_3j_1} + B^+_{i_1i_3} B^+_{i_2j_3} B^+_{j_1j_2} + B^+_{i_1j_3} B^+_{i_2j_1} B^+_{i_3j_2} - B^+_{i_1j_1} B^+_{i_2i_3} B^+_{j_2j_3} \\
& + B^+_{i_1i_2} B^+_{j_1i_3} B^+_{j_2j_3} + B^+_{i_1j_1} B^+_{i_3i_2} B^+_{j_1j_3} + B^+_{i_1i_3} B^+_{i_2j_1} B^+_{j_2j_3} + B^+_{i_1j_3} B^+_{i_2j_2} B^+_{i_3j_1} \}|0).
\end{aligned}$$

We now consider the effect of E_{ij} as defined by Eq. (14) on the antisymmetrized states obtained from Eq. (18) by applying the normalized operator of Eq. (21). Such an operator only permutes the index locations of a given set i_k, j_k ($k = 1,...,r$) whereas E_{ij} replaces a $j = i_m$ or j_m by a corresponding $i = i_m$ or j_m at the same location and with the same sign. This follows from the result of Eq. (17). Hence E_{ij} is just a replacement operator which does not disturb the ordering in the set and commutes with $\mathcal{A}$. This can be readily illustrated using a simple example of $E_{14}\mathcal{A} B^+_{12} B^+_{34}|0)$. We have

$$\mathcal{A} B^+_{12} B^+_{34}|0) = (1/\sqrt{3})(B^+_{12} B^+_{34} - B^+_{13} B^+_{24} + B^+_{14} B^+_{23})|0)$$

so that

$$\begin{aligned}
E_{14}\mathcal{A} B^+_{12} B^+_{34}|0) &= (1/\sqrt{3})(B^+_{12} B^+_{31} - B^+_{13} B^+_{21} + B^+_{11} B^+_{23})|0) \\
&= (1/\sqrt{3})(-B^+_{12} B^+_{13} + B^+_{13} B^+_{12})|0) \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{A} E_{14} B^+_{12} B^+_{34}|0) &= -\mathcal{A} B^+_{12} B^+_{13}|0) \\
&= \mathcal{A} B^+_{11} B^+_{23}|0) = 0.
\end{aligned}$$

One of the important consequences of this commutativity of the unitary group generators with the antisymmetrizer is that these generators either map the physical space of antisymmetric tensors into itself or annihilate it.

It has been shown by earlier workers[6,8] that the antisymmetric combinations of boson operator states obtained as above are, in fact, the DBM states

$$\begin{aligned}
&\widetilde{B}^+_{i_1j_1} \widetilde{B}^+_{i_2j_2} \cdots \widetilde{B}^+_{i_rj_r}|0) \\
&= \mathcal{A} B^+_{i_1j_1} B^+_{i_2j_2} \cdots B^+_{i_rj_r}|0),
\end{aligned} \tag{22}$$

where a normalized form results on the right side of Eq. (22) if the $\mathcal{A}$ of Eq. (21) is used. The modified boson operators on the left side of Eq. (22) are defined as

$$\widetilde{B}^+_{ij} = B^+_{ij} - \sum_{p,q} B^+_{ip} B^+_{jp} B_{pq} \tag{23}$$

and

$$\widetilde{B}_{ij} = B_{ij}, \tag{24}$$

where the B^+ and B on the right sides of Eqs. (23) and (24) are ordinary boson operators satisfying Eqs. (4)–(7). The Dyson mapping from the bifermion to the boson space is generated using[6,8]

$$\begin{aligned}
&a_i^+ a_j^+ \rightarrow \widetilde{B}^+_{ij}, \quad a_i a_j \rightarrow B_{ij}, \\
&a_i^+ a_j \rightarrow \sum_p B^+_{ip} B_{jp},
\end{aligned} \tag{25}$$

where $\{a_i^+, a_i | i = 1,...,n\}$ are the usual shell-model fermion creation and destruction operators. The mapping of Eq. (25) implies that $a_i^+ a_j \rightarrow E_{ij}$, defined by Eq. (14) so that the one-body term of the Hamiltonian

$$H_{\rm I} = \sum_{i,j} f_{ij} a_i^+ a_j = \sum_{i,j} f_{ij} E_{ij} \tag{26}$$

is not affected so far as its Hermiticity, etc., are concerned by this mapping. We thus need to consider only the two-body term in the Hamiltonian which can be expressed in terms of shell-model fermion operators as

$$\begin{aligned}
H_{\rm II} &= \frac{-1}{4} \sum_{ijkl} V_{ij;\,kl} a_i^+ a_j^+ a_l a_k \\
&= \frac{1}{2} \sum_{i<j} \sum_{k<l} V_{(ij);\,(kl)} a_i^+ a_j^+ a_k a_l,
\end{aligned} \tag{27}$$

where $V_{(ij);(kl)}$ is defined over antisymmetrized pair orbitals $\phi_{ij} = 1/\sqrt{2}\,(\phi_i\phi_j - \phi_j\phi_i)$, etc. Using the anticommutation relations of the fermion operators we can readily establish the result[4]

$$H_{\rm II} = \frac{1}{2} \sum_{i<j} \sum_{k<l} V_{(ij);\,(kl)} (E_{ik} E_{jl} - \delta_{jk} E_{il}), \tag{28}$$

where the fermion operator realization of the generators

$$E_{ij} = a_i^+ a_j \tag{29}$$

of U(n) has been used. Since the elements of a Lie algebra are independent of their mode of realization, we use the definition of Eq. (14) to reexpress Eq. (28) as

$$H_{\rm II}=\frac{1}{2}\sum_{i<j}\sum_{k<l}V_{(ij);(kl)}\times\Big\{\sum_{p,q}B^+_{ip}B_{kp}B^+_{jq}B_{lq}-\delta_{jk}\sum_{p}B^+_{ip}B_{lp}\Big\}. \tag{30}$$

Using the commutation relations of Eq. (6), the right side of Eq. (30) can be transformed into

$$H_{\rm II}=\frac{1}{2}\sum_{i<j}\sum_{k<l}V_{(ij);(kl)}\times\Big(B^+_{ij}B_{kl}+\sum_{p,q}B^+_{ip}B^+_{jq}B_{kp}B_{lq}\Big)=H^1_{\rm II}+H^2_{\rm II}. \tag{31}$$

We now consider the effect of

$$H^2_{\rm II}=\frac{1}{2}\sum_{i<j}\sum_{k<l}V_{(ij);(kl)}\sum_{p,q}B^+_{ip}\,B^+_{jq}B_{kp}B_{lq} \tag{32}$$

on the Dyson mapped ket of Eq. (22). The presence of $B_{kp}B_{lq}$ on the right side of Eq. (32) and the result of Eq. (17) ensure that all terms in the multiple summations yield zero when acting on the state defined by Eq. (22), except those for which minimally $k,p,l,q\in\{i_k,j_k\,|\,k=1,...,r\}$. Neglecting those terms in the summations of Eq. (32) for which this does not happen, the antisymmetrizer ensures that one term exists in the linear combinations on the right of Eq. (22) with $B^+_{kp}B^+_{lq}$ occurring in it with the signature $(-)^P$. If this occurs, the antisymmetrizer ensures also that there is another term in the linear combination on the right side of Eq. (22) with $B^+_{kl}B^+_{pq}$ occurring in it with the signature $-(-)^P$. This change of sign occurs since $B^+_{kp}B^+_{lq}$ and $B^+_{kl}B^+_{pq}$ just differ by an index location transposition. Thus, insofar as the transformations induced by $H^2_{\rm II}$ of Eq. (32) acting on the states of Eq. (22) are concerned, we have the result

$$H^2_{\rm II}(\widetilde{B}^+_{i_1j_1}\widetilde{B}^+_{i_2j_2}\cdots\widetilde{B}^+_{i_rj_r})|0=\Big[-\frac{1}{2}\sum_{i<j}\sum_{k<l}V_{(ij);\,kl)}\sum_{p,q}B^+_{ip}B^+_{jq}B_{kl}B_{pq}\Big]\times\mathcal{A}(B^+_{i_1j_1}\cdots B^+_{i_rj_r}|0). \tag{33}$$

This result combined with that of Eq. (31) leads to

$$\begin{aligned}H_{\rm II}(\widetilde{B}^+_{i_1j_1}\cdots\widetilde{B}^+_{i_rj_r})|0)&=\frac{1}{2}\sum_{\substack{i<j\\k<l}}V_{(ij);(kl)}\Big[B^+_{ij}-\sum_{p,q}B^+_{ip}B^+_{jq}B_{pq}\Big]\times B_{kl}\mathcal{A}(B^+_{i_1j_1}\cdots B^+_{i_rj_r})|0)\\&=\frac{1}{2}\sum_{\substack{i<j\\k<l}}V_{(ij);(kl)}\,\widetilde{B}^+_{ij}B_{kl}\times(\widetilde{B}^+_{i_1j_1}\cdots\widetilde{B}^+_{i_rj_r})|0\\&=\widetilde{H}_{\rm II}(\widetilde{B}^+_{i_1j_1}\cdots\widetilde{B}^+_{i_rj_r})|0,\end{aligned} \tag{34}$$

where the definitions of Eqs. (22), (23), and (24) have been used to obtain the last equality.

Comparing the results of Eqs. (28) and (34) we thus find the equivalence

$$\begin{aligned}H_{\rm II}\mathcal{A}(B^+_{i_1j_1}\cdots B^+_{i_rj_r})|0)&=\widetilde{H}_{\rm II}(\widetilde{B}^+_{i_1j_1}\cdots\widetilde{B}^+_{i_rj_r})|0)\\&=\widetilde{H}_{\rm II}\mathcal{A}(B^+_{i_1j_1}\cdots B^+_{i_rj_r})|0,\end{aligned} \tag{35}$$

for the two-body part of H. The difference between the left and right sides of Eq. (35) is that $H^+_{II}=H_{\rm II}$, whereas $\widetilde{H}^+_{\rm II}\neq\widetilde{H}_{\rm II}$. Also, $H_{\rm II}$ being defined in terms of E_{ij} commutes with every permutation $P\in S_{2r}$, whereas $\widetilde{H}_{\rm II}$ does not. The need to use expanded forms of $\widetilde{B}^+$ operators could be a disadvantage when the number of pairs is large. This aspect of noncommutativity and the identity of Eq. (35) can be readily illustrated using a simple example of two boson states. Let

$$\widetilde{B}^+_{12}\widetilde{B}^+_{34}|0)=(1/\sqrt{3})(B^+_{12}B^+_{34}-B^+_{13}B^+_{24}+B^+_{14}B^+_{23})|0). \tag{36}$$

For a particular term of $H_{\rm II}$, say,

$$V_{56;(13)}\widetilde{B}^+_{56}\widetilde{B}_{13}=V_{(56);(13)}\Big(B^+_{56}-\sum_{p,q}B^+_{5p}B^+_{6q}B_{pq}\Big)B_{13}, \tag{37}$$

we have

$$\begin{aligned}V_{(56);(13)}\widetilde{B}^+_{56}\widetilde{B}_{13}(\widetilde{B}^+_{12}\widetilde{B}^+_{34})|0)&=\frac{1}{\sqrt{3}}V_{(56);(13)}\Big(B^+_{56}-\sum_{p,q}B^+_{5p}B^+_{6q}B_{pq}\Big)\times B_{13}(B^+_{12}B^+_{34}-B^+_{13}B^+_{24}+B^+_{14}B^+_{23})|0)\\&=-V_{(56);(13)}\widetilde{B}^+_{24}\widetilde{B}^+_{56}|0).\end{aligned} \tag{38}$$

On the other hand, the Hermitian form of Eq. (28) leads to

$$\begin{aligned}H_{\rm II}\widetilde{B}^+_{12}\widetilde{B}^+_{34}|0)&=V_{(56);(13)}E_{51}E_{63}\,\mathcal{A}B^+_{12}B^+_{34}|0)\\&=V_{(56);(13)}\mathcal{A}B^+_{52}B^+_{64}|0)\\&=-V_{(56);(13)}\,\mathcal{A}B^+_{24}B^+_{56}|0).\end{aligned} \tag{39}$$

III. DISCUSSION AND CONCLUSION

The main content of the analysis presented in Sec. II is the restriction of the group $U(n(n-1)/2)$ defined on the symmetric representation space of $V_{n(n-1)/2}{}^{\otimes r}$ to yield the representations $\langle\lambda\rangle$ of U(n). The representation $\langle 1^{2p}\,0^{n-2p}\rangle$ of U(n) was used to generate the antisymmetric physical boson states of rank p. The fermion Hamiltonian of Eq. (27) was then expressed in the conventional manner in terms of generators of U(n) as in Eq. (28). Reexpressing this Hamiltonian in terms of the boson operators, it was found that, insofar as its action on physical states was concerned, it was identical to the action of the DBM Hamiltonian. This result is of considerable importance since it provides correspondence between fermion and DBM Hamiltonian matrix elements with the corresponding boson states. This can be readily demonstrated as follows.

Consider the DBM biorthonormal basis states $|\tilde{i})$ and $(j|$ defined as

$$|\tilde{i})=N_i\overline{N}_i|\mathcal{A}\Psi_i(B^+)), \tag{40}$$

$$(i|=N_i(\Psi_i(B)|, \tag{41}$$

where $\Psi_i(B)$, $\Psi_i(B^+)$ are polynomials in B and B^+, respectively, and N_i and $\overline{N}_i$ are normalization constants such that

$$(\overline{N}_i)^2(A\Psi_i|A\Psi_i) = 1 \quad (42)$$

and

$$(i|\tilde{i}) = N_i{}^2\overline{N}_i(\Psi_i|A\Psi_i) = 1, \quad (43)$$

so that $N_i = (\overline{N}_i(\Psi_i|A\Psi_i))^{-1/2}$. Thus

$$|\tilde{i}) = (\overline{N}_i/(\Psi_i|A\Psi_i))^{1/2}|A\Psi_i) \quad (44)$$

and

$$(i| = (\overline{N}_i(\Psi_i|A\Psi_i))^{-1/2}(\Psi_i|. \quad (45)$$

These results lead to

$$\begin{aligned}(i|H_D|\tilde{j}) &= (\overline{N}_j/\overline{N}_i(\Psi_i|A\Psi_i)(\Psi_j|A\Psi_j))^{1/2} \\ &\quad\times(\Psi_i|H_D|A\Psi_j) \\ &= (\overline{N}_j/\overline{N}_i(A\Psi_i|A\Psi_i)(A\Psi_j|A\Psi_j))^{1/2} \\ &\quad\times(A\Psi_i|H_F|A\Psi_j), \end{aligned} \quad (46)$$

where we have replaced $A\Psi_i$ by $A^2\Psi_i$ using the essential idempotency of A and used the commutativity of H with all $P\in S_{2r}$. A similar procedure also yields

$$\begin{aligned}(19(j|H_D|\tilde{i}) &= (\overline{N}_i/\overline{N}_j(\Psi_i|A\Psi_i)(\Psi_j|A\Psi_j))^{1/2} \\ &\quad\times(\Psi_j|H_D|A\Psi_i) \\ &= (\overline{N}_i/\overline{N}_j(A\Psi_i|A\Psi_i)(A\Psi_j|A\Psi_j))^{1/2} \\ &\quad\times(A\Psi_i|H_F|A\Psi_j). \end{aligned} \quad (47)$$

It is to be noted that the normalization condition $(i|\tilde{i}) = 1$ is preserved even if one multiplies $|\tilde{i})$ by γ_i and $(i|$ by $1/\gamma_i$. This γ ambiguity in the normalization of the DBM basis states may yield different off-diagonal matrix elements of the Hamiltonian depending upon the choice of γ. This ambiguity can amicably be resolved by using the Hermitization procedure of Gambhir and Basavaraju (GB)[10] also advocated recently by other authors.[11] The GB procedure yields a Hermitian matrix h through

$$h_{ij} = ((i|H_D|\tilde{j})(j|H_D|\tilde{i}))^{1/2}. \quad (48)$$

The results of Eqs. (46) and (47) yield

$$\begin{aligned}h_{ij} &= ((A\Psi_i|A\Psi_i)(A\Psi_j|A\Psi_j))^{-1/2} \\ &\quad\times(A\Psi_i|H_D|A\Psi_j). \end{aligned} \quad (49)$$

For the present problem the matrix h as seen from Eq. (48) is identical to the corresponding Hamiltonian matrix of the original fermion problem. Therefore, DBM along with the GB hermitization procedure is identical to solving the original shell-model problem in fermion space. These observations are consistent with the earlier investigations,[6,7,9] though here it is made explicitly transparent and in addition gives a deeper insight into the understanding of this problem.

[1] W. G. Harter and C. W. Patterson, "A unitary calculas for electronic orbitals," in *Lecture Notes in Physics and Chemistry*, Vol. 49 (Springer, Berlin, 1976).
[2] J. Paldus, *Electrons in Finite and Infinite Structures*, edited by P. Phariseau and L. Schcire (Plenum, New York, 1977).
[3] C. R. Sarma and K. V. Dinesha, J. Math. Phys. **19**, 1662 (1978).
[4] M. Moshinsky, *Group Theory and Many-Body Problems* (Gordon and Breach, New York, 1978).
[5] S. T. Belyaev and V. G. Zelevinskii, Nucl. Phys. **39**, 582 (1962); T. Marumori, M. Yamamura, and A. Tokunaga, Prog. Theor. Phys. **31**, 1009 (1964).
[6] D. Janssen, F. Donau, S. Frauendorf, and R. V. Jolos, Nucl. Phys. A **172**, 145 (1971); F. Donau and D. Janssen, Nucl. Phys. A **209**, 109 (1973); F. J. Dyson, Phys. Rev. **102**, 1217 (1956).
[7] E. R. Marshalek, Nucl. Phys. A **224**, 221 (1974).
[8] P. Ring and P. Schuck, *The Nuclear Many-Body Problem* (Springer, New York, 1980), Chap. 9.
[9] T. Tamura, Phys. Rev. C **28**, 2840 (1983).
[10] Y. K. Gambhir and G. Basavaraju, Pramana **13**, 269 (1979).
[11] Y. K. Gambhir, P. Ring, and P. Schuck, Nucl. Phys. A **384**, 37 (1982); A **423**, 35 (1984); F. J. W. Hahne, Phys. Rev. C **23**, 2305 (1981); C. T. Li, Phys. Rev. C **29**, 2309 (1984).
[12] I. G. Kaplan, *Symmetry of Many-Electron Systems* (Academic, New York, 1975).
[13] C. R. Sarma and G. G. Sahasrabudhe, J. Math. Phys. **21**, 638 (1980).

Electrohydrodynamic stability of two cylindrical interfaces under the influence of a tangential periodic electric field

Nabil T. El Dabe, El Sayed F. El Shehawey, Galal M. Moatimid, and Abou El Magd A. Mohamed[a)]
Department of Mathematics, Faculty of Education, Ain Shams University, Heliopolis, Cairo, Egypt

(Received 31 October 1984; accepted for publication 1 February 1985)

The electrohydrodynamic stability of two cylindrical interfaces influenced by a periodic tangential field is studied. The model allows for general forms of deformations of the interfaces. Two simultaneous ordinary differential equations of the Mathieu type are obtained. The coupled equations are solved by the method of multiple scales and stability conditions are discussed. It is found that the constant tangential field has a stabilizing effect while the tangential periodic field has a stabilizing influence except at resonance points. Graphs are drawn to illustrate the resonance regions in a parameter space. It is also found that the thickness of the jet plays a role in the stability criterion. The frequency of the modulated field can be used to control the position of the resonance regions. The special cases of large modulation and small modulation are also examined. It is found that for large modulation the electric field exhibits an enhanced destabilizing influence.

I. INTRODUCTION

The stability of jets was the concern of many investigators in fluid mechanics. Early experiments by Bassat[1] and theoretical analysis by Rayleigh[2] showed that the jet breaks up for deformations having x less than unity ($x = ka$, k being the wave number and a the radius of the jet). If the fluid is bounded by two cylindrical interfaces with radii a,b ($a < b$), then the fluid is again stable if both ka,kb are greater than unity[3] and thus Rayleigh's criterion for a single interface is still true for two cylindrical interfaces.

It was shown by Nayyar and Murty[4] that the inclusion of a constant axial electric field improves the stability of a jet. For a suitable choice of the electric field, the insulating jet can be stable for $x < 1$.

The effect of a tangential periodic field was examined by Mohamed and Nayyar.[5] They introduced an axial periodic field to the jet and found that although the jet may be stable for $x < 1$, it may be unstable at some values of $x > 1$ when the system is brought to a state of resonance. The studies were later carried out to two concentric cylindrical surfaces bounding the fluid. It was found that the constant axial electric field leads to a stable state for some values of ka and kb less than unity.[6]

In this work, we extend our previous studies to examine the stability of two cylindrical interfaces separating dielectric fluids which are acted upon by a periodic field.

II. THE SYSTEM IN EQUILIBRIUM STATE

The system considered here consists of three infinite homogeneous dielectric inviscid fluids of densities ρ_i, $\rho_{\rm be}$, and ρ_o and dielectric constants ϵ^i, $\epsilon^{\rm be}$, and ϵ^o, where the subscripts or superscripts i, be, and o refer to quantities inside the inner cylinder, between the two cylinders, and outside the outer cylinder. The three fluids are separated by two coaxial cylindrical interfaces whose radii are $r = a$ and $r = b$. No volume charges are assumed to be present in the bulk of the fluids. Also, because of the continuity of the electric field, no surface charges are present at the interfaces in the equilibrium state and will therefore vanish during the perturbation.[7] In order to produce a parametric excitation in this system, we superimpose a modulated electric field ($E^* \cos \omega t$) in the z direction on the already existing constant field E_0. It is also assumed that the axis of the cylinder is the z axis [using the cylindrical polar coordinates (r,θ,z)]. We assume that the quasistatic approximation[8] is valid for the problem and hence the electric field can be determined from a scalar function ϕ.

In each of the fluids, the equations governing the motion are[8]

$$\nabla\cdot\epsilon\mathbf{E} = 0, \tag{1}$$

$$\nabla \wedge \mathbf{E} = 0, \tag{2}$$

$$\rho\left[\frac{\partial \mathbf{V}}{\partial t} + (\mathbf{V}\cdot\nabla)\mathbf{V}\right] = -\nabla\Pi, \tag{3}$$

and

$$\nabla\cdot\mathbf{V} = 0, \tag{4}$$

where $\Pi = P - (\epsilon/2)E^2$, P is the hydrostatic pressure, ϵ is the permittivity, $\mathbf{E}$ is the electric field, and $\mathbf{V}$ is the fluid velocity.

III. PERTURBATION EQUATIONS

Consider the effect of small wave disturbances to the interfaces $r = a$ and $r = b$, propagating in the positive z direction. The surfaces deflections are assumed to be of the form

$$\begin{aligned} r_a &= a + \gamma_1(t)\exp[i(kz + m\theta)], \\ r_b &= b + \gamma_2(t)\exp[i(kz + m\theta)], \end{aligned} \tag{5}$$

where $\gamma_1(t)$ and $\gamma_2(t)$ are arbitrary functions of time which determine the behavior of the amplitude of the disturbances of the interfaces and k is the wave number and is assumed to be positive.

[a)] Present address: Department of Mathematics, Faculty of Science, University of Qatar, POB 2713, Doha, Qatar.

 0022-2488/85/082072-10$02.50

For a small departure from the equilibrium state, the quantities $\mathbf{E}$, $\mathbf{V}$, and Π receive the increments $\mathbf{E}_1$, $\mathbf{V}_1$, and Π_1 to yield

$$\mathbf{E}=\mathbf{E}_0+\mathbf{E}_1,\quad \mathbf{V}=\mathbf{V}_0+\mathbf{V}_1,\quad \Pi=\Pi_0+\Pi_1.$$

The subscript "0" refers to quantities in the equilibrium state. The linearization of Eqs. (1)–(4) leads to the following:

$$\nabla^2\phi_1=0,\qquad \mathbf{E}_1=-\nabla\phi_1, \tag{6}$$

where

$$\nabla^2\Pi_1=0. \tag{7}$$

As a result of the perturbation ϕ_1 and Π_1, in view of the dependence given by Eq. (5) may have the form

$$\Phi_1^{(j)}=\hat{\phi}^{(j)}(r,t)\exp[i(kz+m\theta)], \tag{8}$$

$$\Pi_1^{(j)}=\hat{\Pi}^{(j)}(r,t)\exp[i(kz+m\theta)],\quad j=i,\text{be},o. \tag{9}$$

Substituting from Eqs. (8) and (9) into Eqs. (6) and (7), the solutions of the resulting differential equations are

$$\phi_1^i=AI_m(kr)\exp[i(kz+m\theta)],\quad r<a, \tag{10}$$

$$\phi_1^{\text{be}}=[CI_m(kr)+DK_m(kr)]\exp[i(kz+m\theta)],\quad a<r<b, \tag{11}$$

$$\phi_1^o=BK_m(kr)\exp[i(kz+m\theta)],\quad r>b, \tag{12}$$

and

$$\Pi_1^i=\dot{F}(t)I_m(kr)\exp[i(kz+m\theta)],\quad r<a, \tag{13}$$

$$\Pi_1^{\text{be}}=[\dot{G}(t)I_m(kr)+\dot{H}(t)K_m(kr)]\exp[i(kz+m\theta)],\quad a<r<b \tag{14}$$

$$\Pi_1^o=\dot{M}(t)I_m(kr)\exp[i(kz+m\theta)],\quad r>b, \tag{15}$$

where I_m and K_m are modified Bessel functions, A, B, C, D, F, G, H, and M are time-dependent constants which are to be evaluated by making use of the appropriate boundary conditions, and the overdot denotes the time derivative.

Substituting from Eqs. (13)–(15) into the rth component of the linearized equation of motion, and integrating with respect to time, we get

$$\rho_i V_1^i=-kF(t)I'_m(kr)\exp[i(kz+m\theta)],\quad r<a, \tag{16}$$

$$\rho_{\text{be}} V_1^{\text{be}}=-k[G(t)I'_m(kr)+H(t)K'_m(kr)]\times\exp[i(kz+m\theta)],\quad a<r<b, \tag{17}$$

$$\rho_o V_1^o=-kM(t)K'_m(kr)\exp[i(kz+m\theta)],\quad r>b, \tag{18}$$

where V_1 stands now for the radial velocity and the prime denotes differentiation with respect to the argument.

IV. BOUNDARY CONDITIONS

The determination of A, B, C, D, F, G, H, and M can be obtained by applying the following boundary conditions.

(i) The electric potential ϕ is continuous at the interfaces

$$r_a=a+\gamma_1(t)\exp[i(kz+m\theta)],$$

$$r_b=b+\gamma_2(t)\exp[i(kz+m\theta)].$$

(ii) The normal electric displacement is continuous across the interfaces.

(iii) The surface displacements at the perturbed interfaces r_a and r_b should be uniquely determined by the equations of motion.

(iv) The normal component of the stress tensor Π_{ij} should be discontinuous at the interfaces by the surface tension, where

$$\Pi_{ij}=-(\Pi+(\epsilon/2)E_k E_k)\delta_{ij}+\epsilon E_i E_j.$$

Applying the above boundary conditions and using the properties of the Bessel functions,[9] we find after some lengthy but straightforward calculations that γ_1 and γ_2 must be governed by the following coupled equations:

$$\ddot{\gamma}_1+[a_1+4b_1(E_o+E^*\cos\omega t)^2]\gamma_1+[a_2+4b_2(E_o+E^*\cos\omega t)^2]\gamma_2=0, \tag{19}$$

$$\ddot{\gamma}_2+[a_3+4b_3(E_o+E^*\cos\omega t)^2]\gamma_2+[a_4+4b_4(E_o+E^*\cos\omega t)^2]\gamma_1=0, \tag{20}$$

where the a_j's and b_j's are constants given in Appendix A.

The solutions of the simultaneous differential equations (19) and (20) determine the behavior of the amplitudes of the disturbances of the interfaces. Therefore the stability discussion will be based on the study of these two equations.

V. STABILITY ANALYSIS

The case where $E^*\to 0$ corresponds to the case of a constant axial electric field which has been extensively discussed elsewhere[3] and will not be discussed here.

Although the solutions and the properties of a single Mathieu equation are well known,[10] there is no general analytical available solution of coupled Mathieu equations (19) and (20). Therefore we shall discuss the stability of the system by using an asymptotic expansion treatment. The method of multiple scales has been successfully used to treat similar systems[11,12] and it will be adopted here to study the stabilities of the system. In order to apply the method of multiple scales, we shall make use of a smallness parameter δ. The definition of δ depends on which case of interest is required. For the general case when E_0 and E^* are of the same weight, we define δ such that

$$E_0=\sqrt{\delta}\,\hat{e}\quad\text{and}\quad E^*=\sqrt{\delta}\,e,$$

where $\hat{e}$, e are the magnitudes of the electric fields E_0, E^*, respectively, and δ is small dimensionless parameter. Thus

$$(E_0+E^*\cos\omega t)^2=\delta(\hat{e}+e\cos\omega t)^2,$$

and therefore Eqs. (19) and (20) become

$$\ddot{\gamma}_1+[a_1+4\delta b_1(\hat{e}+e\cos\omega t)^2]\gamma_1+[a_2+4\delta b_2(\hat{e}+e\cos\omega t)^2]\gamma_2=0, \tag{21}$$

$$\ddot{\gamma}_2+[a_3+4\delta b_3(\hat{e}+e\cos\omega t)^2]\gamma_2+[a_4+4\delta b_4(\hat{e}+e\cos\omega t)^2]\gamma_1=0. \tag{22}$$

According to the Floquet theory of linear differential equations with periodic coefficients[10] the ω-δ plane is divided into regions of stability and instability which are separated by transition curves along which $\gamma_j(t)$ is periodic with a period of either Π or 2Π (see Refs. 10 and 11).

VI. STABILITY OF THE GENERAL CASE

We seek a uniformly first-order expansion of the solution of Eqs. (21) and (22) for a small dimensionless parameter δ in the form

$$\gamma_j = \gamma_{j0}(T_0,T_1) + \delta\gamma_{j1}(T_0,T_1) + \cdots, \quad j = 1,2, \tag{23}$$

where T_0 and T_1 are time scales variables defined as

$$T_0 = t \quad \text{and} \quad T_1 = \delta t.$$

Substituting from Eqs. (23) into Eqs. (21) and (22), transforming the derivatives, and equating coefficients of like powers of δ, we obtain, for order δ^0,

$$D_0^2\,\gamma_{10} + a_1\,\gamma_{10} + a_2\,\gamma_{20} = 0, \tag{24}$$

$$D_j \equiv \frac{\partial}{\partial T_j},$$

$$D_0^2\,\gamma_{20} + a_3\,\gamma_{20} + a_4\,\gamma_{10} = 0, \tag{25}$$

and for order δ^1,

$$D_0^2\,\gamma_{11} + a_1\,\gamma_{11} + a_2\,\gamma_{21} = -2D_0D_1\,\gamma_{10} - 4(b_1\,\gamma_{10} + b_2\,\gamma_{20})(\hat{e} + e\cos\omega T_0)^2, \tag{26}$$

$$D_0^2\,\gamma_{21} + a_3\,\gamma_{21} + a_4\,\gamma_{11} = -2D_0D_1\,\gamma_{20} - 4(b_4\,\gamma_{10} + b_3\,\gamma_{20})(\hat{e} + e\cos\omega T_0)^2. \tag{27}$$

The solutions of Eqs. (24) and (25) may be represented in the form

$$\gamma_{10} = a_2A_j(T_1)\exp[i\omega_jT_0] + \text{c.c.}, \quad j = 1,2, \tag{28}$$

$$\gamma_{20} = (\omega_j^2 - a_1)A_j(T_1)\exp[i\omega_jT_0] + \text{c.c.}, \quad j = 1,2, \tag{29}$$

where c.c. represents the complex conjugate of the preceding terms, $A_1(T_1)$, $A_2(T_1)$ are unknown complex functions, and each of ω_1 and ω_2 satisfies the following equation:

$$\omega_0^4 - (a_1 + a_3)\omega_0^2 + (a_1a_3 - a_2a_4) = 0, \tag{30}$$

which is a quadratic equation in (ω_0^2) having the real distinct roots (see Appendix B) ω_1^2 and ω_2^2 $(\omega_1^2 > \omega_2^2)$ and they are given by

$$\omega_{1,2}^2 = \tfrac{1}{2}\{(a_1 + a_3) \pm [(a_1 - a_3)^2 + 4a_2a_4]^{1/2}\}. \tag{31}$$

It is easy to show that if ω_1^2 and ω_2^2 are both real and positive, then the system is statically stable in the absence of the electric field, while if ω_1^2 and ω_2^2 have different signs, this means that the system statically has two different modes, one stable and the other is unstable, i.e., the system is unstable in the absence of the electric field. Again, if ω_1^2 and ω_2^2 are both negative then the system will also be unstable in the absence of the electric field.

Substituting from Eqs. (28) and (29) into Eqs. (26) and (27), we obtain

$$\begin{aligned} D_0^2\,\gamma_{11} + a_1\,\gamma_{11} + a_2\,\gamma_{21} &= -2ia_2\omega_jD_1A_j\exp[i\omega_jT_0] - (b_1a_2 - b_2a_1 + b_2\omega_j^2) \\ &\quad \times\{(2e^2 + 4\hat{e}^2)A_j\exp[i\omega_jT_0] \\ &\quad + e^2A_j(\exp[i(\omega_j + 2\omega)T_0] \\ &\quad + \exp[i(\omega_j - 2\omega)T_0]) + 4e\hat{e}A_j(\exp[i(\omega_j + \omega)T_0] \\ &\quad + \exp[i(\omega_j - \omega)T_0])\} + \text{c.c.} \end{aligned} \tag{32}$$

and

$$\begin{aligned} D_0^2\,\gamma_{23} + a_3\,\gamma_{21} + a_4\,\gamma_{11} &= -2i\omega_j(\omega_j^2 - a_1)D_1A_j\exp[i\omega_jT_0] \\ &\quad - (b_4a_2 - b_3a_1 + b_3\omega_j^2)\{(2e^2 + 4\hat{e}^2)A_j\exp[i\omega_jT_0] \\ &\quad + e^2A_j(\exp[i(\omega_j + 2\omega)T_0] \\ &\quad + \exp[i(\omega_j - 2\omega)T_0]) + 4e\hat{e}A_j(\exp[i(\omega_j + \omega)T_0] \\ &\quad + \exp[i(\omega_j - \omega)T_0])\} + \text{c.c.} \end{aligned} \tag{33}$$

The analysis will be divided into two sections. In the first, the system is statically stable in the absence of the electric field. The second section is concerned with the case where the system is unstable in the absence of the electric field.

If ω_1^2 and ω_2^2 are both real and positive we are in need to distinguish between several possible combinations of ω, ω_1, and ω_2 in analyzing the particular solutions of Eqs. (33) and (34). These include the resonance cases and the nonresonance cases.

A. The general state

We have (1) ω away from ω_1, ω_2, $\frac{1}{2}(\omega_1 \pm \omega_2)$ and away from $2\omega_1$, $2\omega_2$, $(\omega_1 \pm \omega_2)$; (2) ω near ω_1 and near $2\omega_1$; (3) ω near ω_2 and near $2\omega_2$; (4) ω near $\frac{1}{2}(\omega_1 + \omega_2)$ and near $(\omega_1 + \omega_2)$; and (5) ω near $\frac{1}{2}(\omega_1 - \omega_2)$ and near $(\omega_1 - \omega_2)$.

B. Nonresonant case

Because Eqs. (33) and (32) are linear, we can obtain particular solutions for each of the terms on the right-hand sides independently. To determine the solvability conditions, we seek particular solutions corresponding to the terms containing the factor $\exp(i\,\omega_jT_0)$ in the form

$$\begin{aligned} \gamma_{11} &= P_1(T_1)\exp(i\omega_1T_0) + Q_1(T_1)\exp(i\omega_2T_0), \\ \gamma_{21} &= P_2(T_1)\exp(i\omega_1T_0) + Q_2(T_1)\exp(i\omega_2T). \end{aligned} \tag{34}$$

Substituting from Eqs. (34) into Eqs. (32) and (33) and equating the coefficients of $\exp(i\omega_jT_0)$ on both sides, using Eq. (30), we obtain

$$D_1A_j - i\nu_j(e^2 + 2\hat{e}^2)A_j = 0, \quad j = 1,2, \tag{35}$$

where

$$\nu_j = \frac{(a_2b_4 + a_4b_2) - (a_1b_3 + a_3b_1) + (b_1 + b_3)\omega_j^2}{\omega_j(2\omega_j^2 - a_1 - a_3)}. \tag{36}$$

The solution of Eq. (35) is given directly by

$$A_j = \alpha_j\exp[i\delta(e^2 + 2\hat{e}^2)\nu_jt], \tag{37}$$

where the α_j are the integration constants. Clearly the ν_j are real and hence the system, in the case of nonresonance, is stable.

Moreover, if the $\omega_j^2, j = 1,2$, are real and negative, then the ν_j are imaginary ($\nu_j = \pm i\nu_j^*$, ν_j^* real) and hence the two modes are possible, one of which has an amplitude of the form $\exp(-\nu_j^*\delta t)$, which is a damped mode, while the other is of amplitude $\exp(\nu_j^*\delta t)$, which generates a growing unstable state. The system is then unstable.

Again if ω_1^2 and ω_2^2 are real and have different signs (ω_1^2 is positive, say, and ω_2^2 is negative), then ν_1 is real while ν_2 is imaginary, and hence ν_1 results in a stable mode while ν_2 yields an unstable mode and the system is therefore unstable.

The above discussion shows that if the system is statically unstable in the absence of the electric field, it cannot be stabilized by a periodic field. A result proved valid for a single interface.[13] Note that the constant tangential field plays a stabilizing role.

C. Internal resonance

(a) The case of ω near $2\omega_1$. In this case we introduce a detuning parameter σ defined by

$$\omega = 2\omega_1 + \delta\sigma,$$

then we write

$$(\omega - \omega_1)T_0 = (\omega_2 T_0 + \sigma T_1). \tag{38}$$

In this case we seek particular solutions which correspond to the terms containing the factor $\exp(i\omega_j T_0)$ and have the form of Eqs. (34). Instead of Eq. (35) we get

$$D_1 A_1 - i\nu_1[(e^2 + 2\hat{e}^2)A_1 + 2e\hat{e}A_1 \exp(i\sigma T_1)] = 0. \tag{39}$$

Let the solution of Eq. (39) be

$$A_1 = (U + iV)\exp(i\sigma T_1), \tag{40}$$

where U,V are two real functions of T_1. Substituting from Eq. (40) into Eq. (39) and separating the real and imaginary parts we obtain

$$U = \alpha \exp(nT_1), \quad V = \beta \exp(nT_1),$$

where α and β are constants, and n is given by

$$n^2 = \left[\nu_1(e^2 - 2e\hat{e} + 2\hat{e}^2) - \frac{\sigma}{2}\right]\left[\frac{\sigma}{2} - \nu_1(e^2 + 2e\hat{e} + 2\hat{e}^2)\right]. \tag{41}$$

It follows that A_1 is bounded and hence the motion is bounded if, and only if, $n^2 \leqslant 0$.

Consequently, the transition from stability to instability corresponds to

$$\nu_1(e^2 - 2e\hat{e} + 2\hat{e}^2) - \sigma/2 = 0,$$

$$\sigma/2 - \nu_1(e^2 + 2e\hat{e} + 2\hat{e}^2) = 0.$$

The transition curves are given by

$$\begin{aligned} \omega &= 2\omega_1 + \delta 2\nu_1(e^2 - 2e\hat{e} + 2\hat{e}^2) + O(\delta^2), \\ \omega &= 2\omega_1 + \delta 2\nu_1(e^2 + 2e\hat{e} + 2\hat{e}^2) + O(\delta^2). \end{aligned} \tag{42}$$

(b) The case of ω near ω_2. Similar arguments may be used when ω near $2\omega_2$, ω near ω_1 and ω near ω_2 by simply replacing ω_1 by ω_2.

(c) The case of ω near $(\omega_1 + \omega_2)$. We express the nearness of ω to $(\omega_1 + \omega_2)$ by introducing the detuning parameter σ^*, that is defined by

$$\omega = (\omega_1 + \omega_2) + \delta\sigma^*,$$

and hence we can write

$$(\omega - \omega_1)T_0 = \omega_2 T_0 + \sigma^* T_1, \quad (\omega - \omega_2)T_0 = \omega_1 T_0 + \sigma^* T_1. \tag{43}$$

We continue to seek particular solutions of Eqs. (32) and (33) in a pattern similar to that applied to the case of ω near $2\omega_1$. Therefore the details of the analysis will be eliminated. We finally obtain

$$A_j = \sum_{s=1}^{4} \beta_{js} \exp i\left(\frac{\sigma}{2} + \lambda_{0s}\right)T_1, \quad j = 1,2, \tag{44}$$

where the β_{js} are complex constants, and each of λ_{0s} satisfies the following equation:

$$\lambda_0^4 - f_1\lambda_0^2 + f_2^2 = 0, \tag{45}$$

where f_1 and f_2 are given in Appendix A.

The restriction for stability implies that $f_1 > 0$, provided that the discriminant of Eq. (45) is positive.

The transition curves separating stability from instability correspond to $f_1 = 0$; which gives a quadratic equation in σ^* having the roots

$$\begin{aligned} \sigma_{1,2}^{**} = \tfrac{1}{2}\{ &(e^2 + 2\hat{e}^2)(\nu_1 + \nu_2) \\ &\pm [4e^4\mu_1\mu_2 - (e^2 + 2\hat{e}^2)(\nu_1 - \nu_2)^2]^{1/2}\}. \end{aligned} \tag{46}$$

The transition curves are

$$\begin{aligned} \omega &= (\omega_1 + \omega_2) + \delta\sigma_1^* + O(\delta^2), \\ \omega &= (\omega_1 + \omega_2) + \delta\sigma_2^* + O(\delta^2). \end{aligned} \tag{47}$$

(d) The case of ω near $\omega_1 - \omega_2$. Similar results can also be obtained for the case of δ near $(\omega_1 - \omega_2)$ by changing the sign of ω_2, and also for the case of ω near $\frac{1}{2}(\omega_1 \pm \omega_2)$.

From the numerical calculations it is found that $\sigma_{1,2}^{**}$ are complex conjugates and this contradicts Eqs. (47) since ω, ω_1, and ω_2 are real. Therefore, no transition curves can be obtained in the δ-ω plane, for the case of $\omega \cong \frac{1}{2}(\omega_1 \pm \omega_2)$. No resonance modes are therefore predicted for the present case.

It is well known that the system,[3] in the absence of the electric field, is stable for all deformations having $(x = ka) > 1$, where k is the wave number and a is the radius of the jet. In what follows we discuss the stability of the system under a periodic electric field by drawing the transition curves in the δ-k plane.

The δ-k diagram is partitioned into two regions R and Q separated by the dotted line at $x = 1$. The region R is characterized by $x < 1$ while Q represents the region $x > 1$. The letter (S) stands for stable regions and (U) for unstable regions in the δ-k plane.

Figure 1 represents the δ-k plane for a system having $\rho_i = 0.001\,293$ g/cm^3, $\rho_{be} = 0.879$ g/cm^3, $\rho_o = 0.998\,23$ g/cm^3, $\epsilon^i = 1.000\,59$, $\epsilon^{be} = 80.37$, $\epsilon^o = 2.284$, $T_1 = 72.75$ dyn/cm, $T_2 = 35.0$ dyn/cm, $a = 12.0$ cm, $b = 15.0$ cm, $\omega = 3.0$ Hz, $\hat{e} = 3.0$ V/cm, and $e = 2.0$ V/cm.

The system is statically stable in the absence of the electric field and ω_1^2, ω_2^2 are both greater than zero for every k. The graph indicates stable Q regions except at ω near ω_1, ω near ω_2, ω near $2\omega_1$, and ω near $2\omega_2$ where the resonance modes exist. Thus the effect of the periodicity is to produce resonance (unstable) regions. The δ_{ω_1} curves start from $k = 0.485$ on the k axis, the δ_{ω_2} start from $k = 0.795$, the $\delta_{2\omega_1}$ curves start from $k = 0.3$, and the $\delta_{2\omega_2}$ curves start from $k = 0.515$. The δ_{ω_i} corresponds to resonance regions for ω near ω_i.

Near $\omega \cong 2\omega_1$, there is a narrow unstable region. The region gets narrower at $\omega \cong 2\omega_2$, but wider for $\omega \cong \omega_1$. The last two regions intersect to form a bigger region of instability. The combined region resulting from the intersection represents the most dangerous modes of instability over a rela-

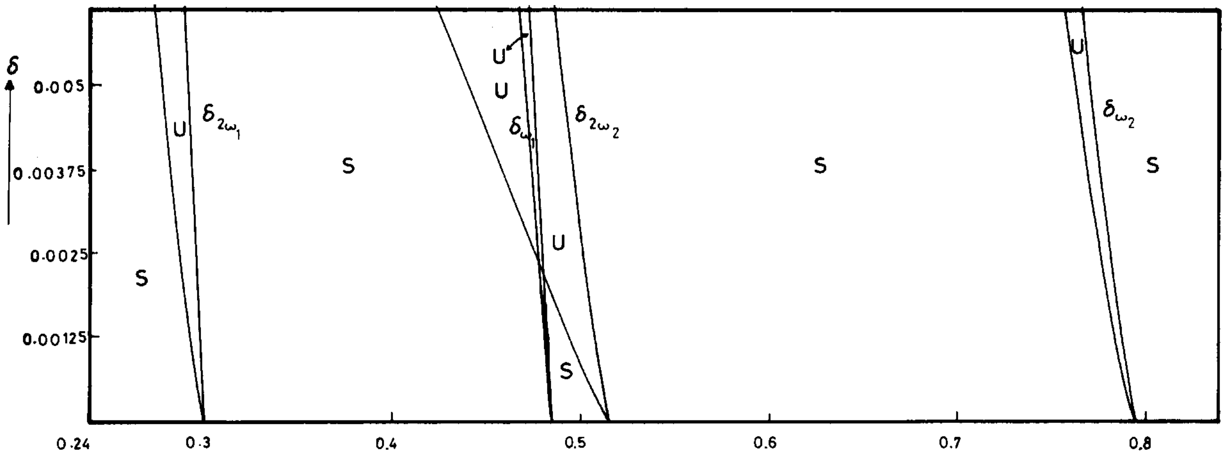


FIG. 1. For a system having $\rho_i = 0.001\,293$ g/cm^3, $\rho_{be} = 0.879$ g/cm^3, $\rho_o = 0.996\,23$ g/cm^3, $\epsilon^i = 1.000\,59$, $\epsilon^{be} = 80.37$, $\epsilon^o = 2.284$, $T_1 = 72.75$ dyn/cm, $T_2 = 35.0$ dyn/cm, $a = 12.0$ cm, $b = 15.0$ cm, $\omega = 3.0$ Hz, $\hat{e} = 3.0$ V/cm, and $e = 2.0$ V/cm. (U) denotes unstable regions and (S) refer to stable regions.

tively larger band of values of k for a given δ. The region corresponding to $\omega \cong \omega_2$ represents a thin region of instability.

Figure 2 represents the same system as in Fig. 1, but $\omega = 2.0$ Hz. We remark that the increase of ω makes the transition curves shift to the right. This can be used to control a system at a given wave number by suitable choice of ω.

In Fig. 3 we plotted the transition curves δ_{ω_1}, δ_{ω_2}, $\delta_{2\omega_1}$, and $\delta_{2\omega_2}$ for the same system considered in Fig. 2, but here $a = 2.0$ cm and $b = 3.0$ cm. The δ-k plane is partitioned into two regions R and Q as defined earlier.

The transition curves δ_{ω_1}, δ_{ω_2}, and $\delta_{2\omega_2}$ are drawn. The resonance regions are unstable. These regions may lie in the region R or Q according to the value of ω. For example, if $\omega = 2.0$ Hz the transition curves $\delta_{2\omega_1}$ lie in the R region (Fig. 4). This saves an unstable region from being in the Q region. If $\omega = 3.0$ Hz, the curve $\delta\, 2\omega_1$ lies in both regions R and Q (Fig. 5).

On the other hand, if we reduce the magnitude of the constant field with respect to $\hat{e}$[$e = 2.0$ V/cm, $\hat{e} = 1.5$ V/cm in Fig. 4, $e = 2.0$ V/cm and $\hat{e} = 6.0$ V/cm in Fig. 6], keeping the other parameters as in Fig. 1, we observe that the transition curves separate unstable regions larger than that in Fig. 1 which ensure again the stabilizing role of the constant axial field.

VII. STABILITY OF THE LARGE MODULATION

For the special case $E^* \gg E_0$, we set $E^* = \sqrt{\delta}\, e$. Thus, $(E_0 + E^* \cos \omega t)^2$ becomes $\delta e^2 \cos^2 \omega t$, the corresponding coupled equations are

$$\ddot{\gamma}_1 + [a_1 + 4\delta b_1(e \cos \omega t)^2]\gamma_1 + [a_2 + 4\delta b_2(e \cos \omega t)^2]\gamma_2 = 0, \tag{48}$$

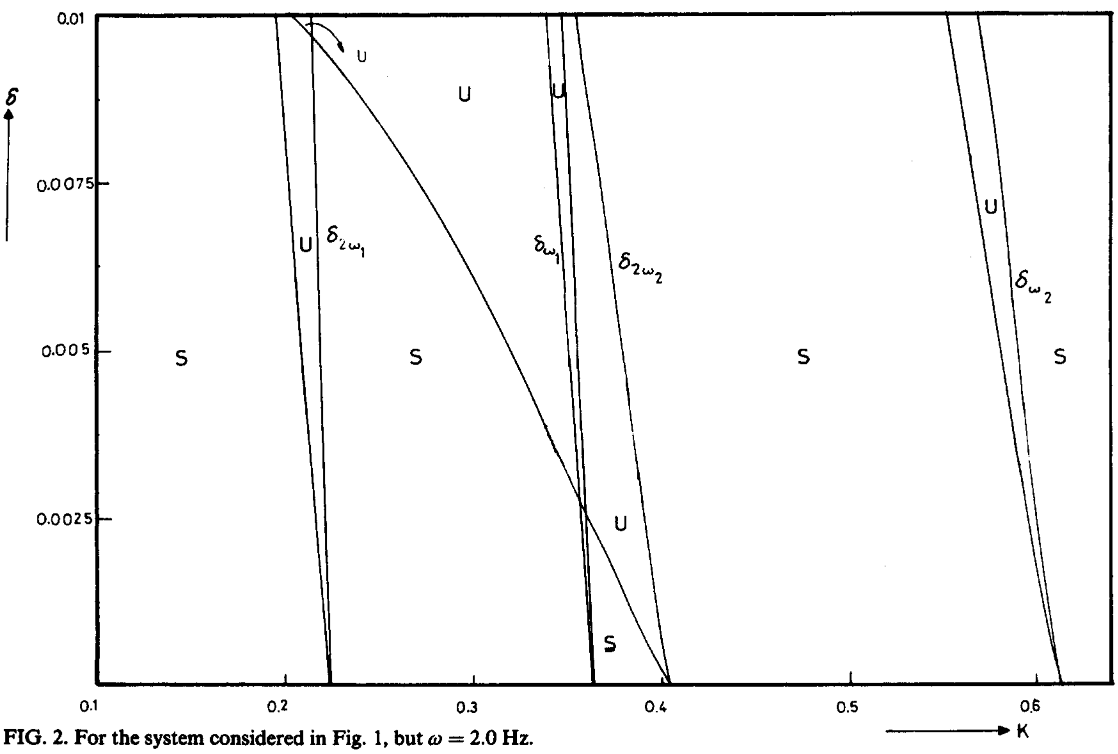


FIG. 2. For the system considered in Fig. 1, but $\omega = 2.0$ Hz.

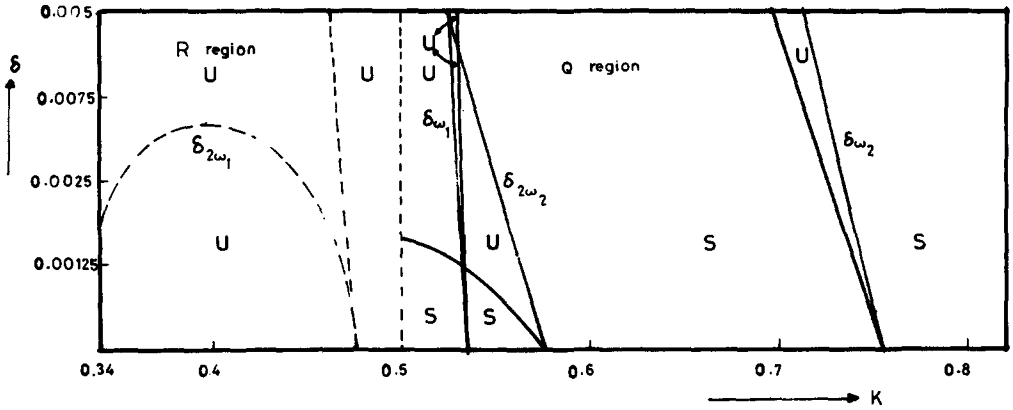


FIG. 3. A representation of the same system as in Fig. 1, but $a = 2.0$ cm, $b = 3.0$ cm, and $\omega = 2.0$ Hz.

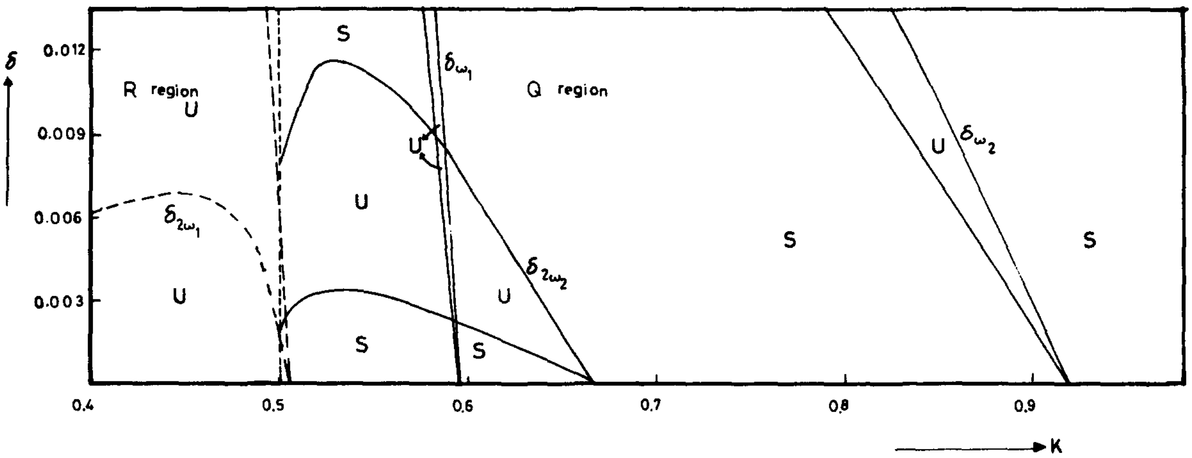


FIG. 4. A representation of the same system as in Fig. 2, but $a = 2.0$ cm, $b = 3.0$ cm, and $\hat{e} = 1.5$ V/cm.

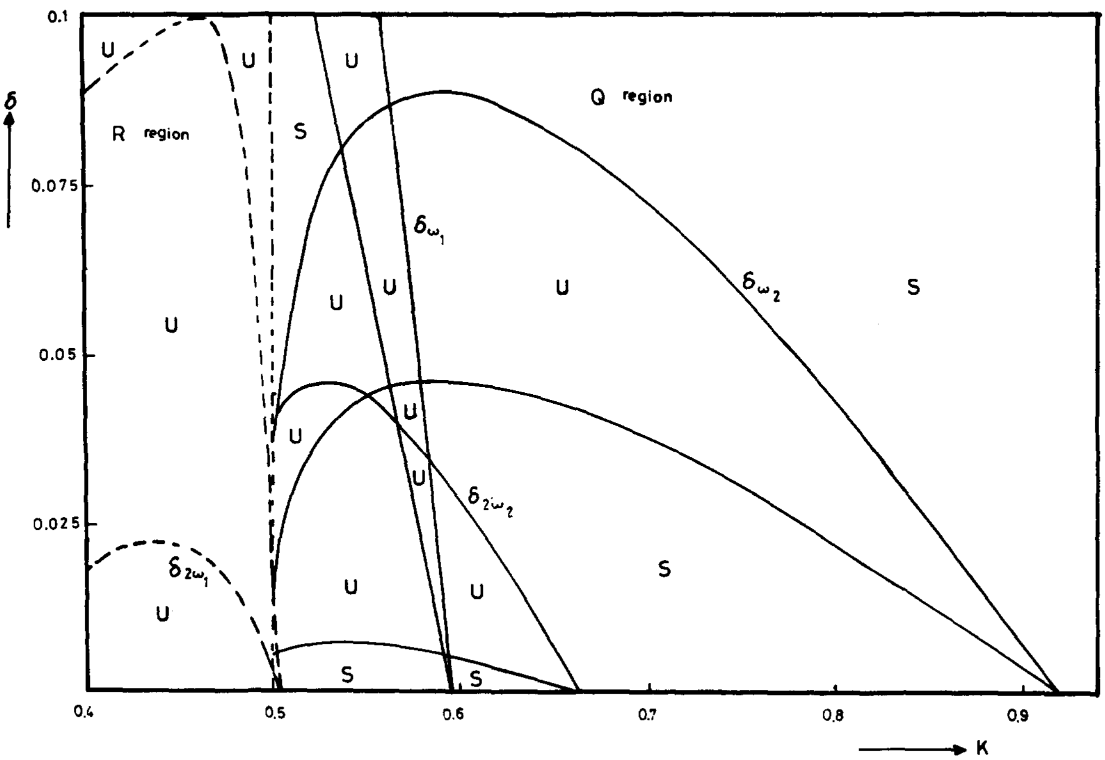


FIG. 5. For the system considered in Fig. 1, but $a = 2.0$ cm and $b = 3.0$ cm.

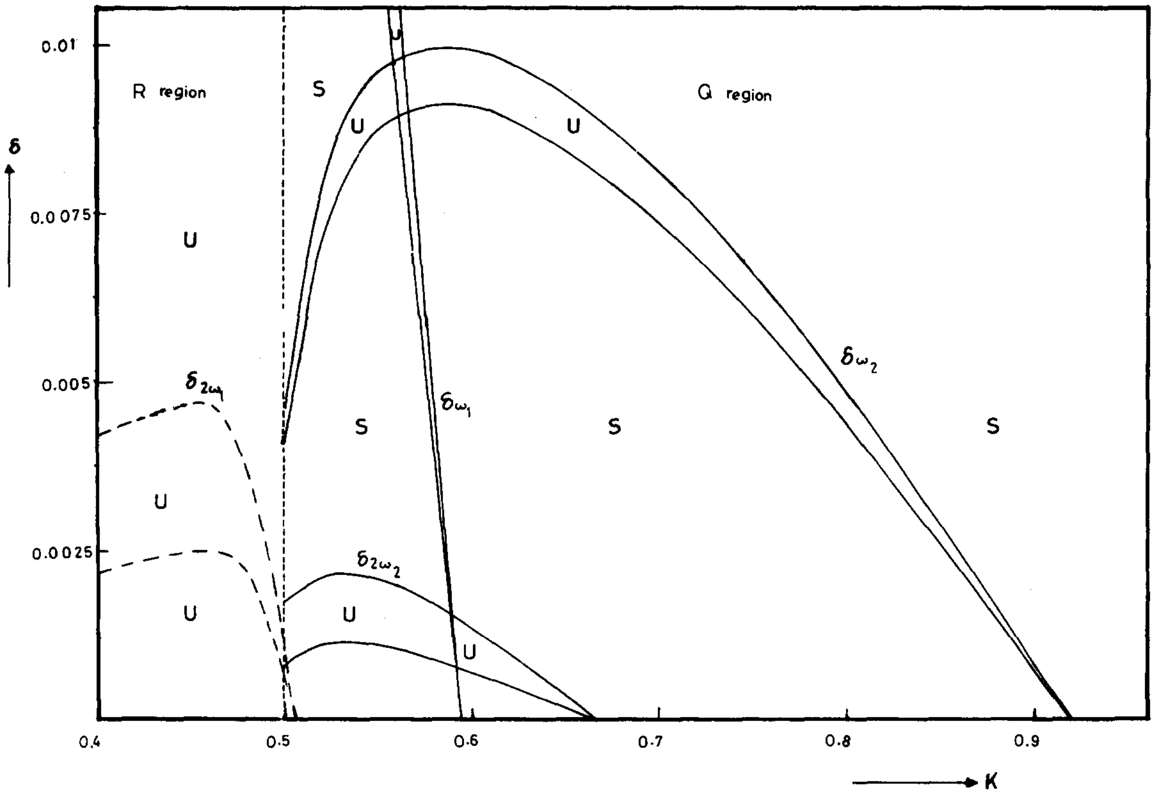


FIG. 6. For the system considered in Fig. 1, but $a = 2.0$ cm and $\hat{e} = 6.0$ V/cm.

$$\ddot{\gamma}_2 + [a_3 + 4\delta b_3(e \cos \omega t)^2]\gamma_2 + [a_4 + 4\delta b_4(e \cos \omega t)^2]\gamma_1 = 0. \tag{49}$$

This case can be treated in the same manner as those above, but the resonances (instability) are here ω near ω_1, ω near ω_2, and ω near $\frac{1}{2}(\omega_1 + \omega_2)$.

Figure 7 is for the same system considered in Fig. 1 but with $\hat{e} = 0.0$ V/cm, $e = 4.0$ V/cm, $a = 2.0$ cm, and $b = 3.0$ cm. The δ-k diagram is partitioned into two regions R, Q separated by the dotted line at $k = 0.5$, the R region is unstable and the Q region indicates a stable region except at $\omega \cong \omega_1$ and $\omega \cong \omega_2$ where the resonance mode exists. The resonance regions are plotted. The transition curves δ_{ω_1} and δ_{ω_2} separate the stable regions (S) from the unstable regions (U).

Figure 8 is for the same system considered in Fig. 7, but here $e = 2.0$ V/cm.

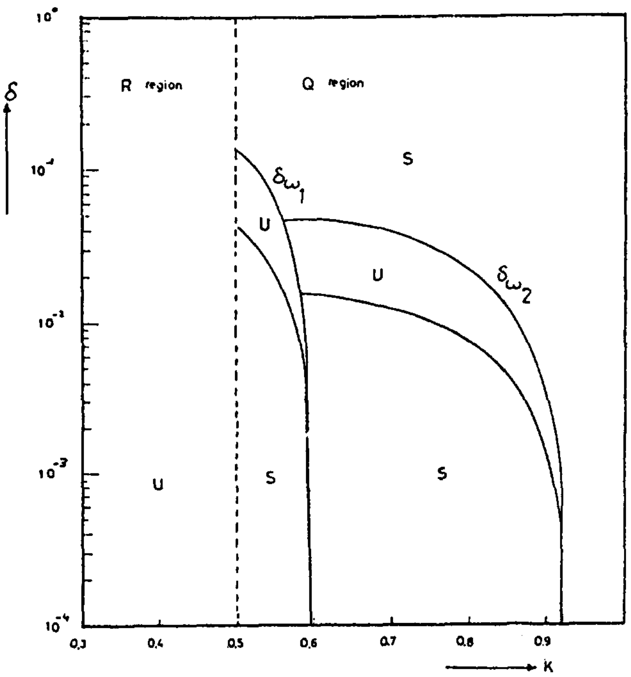


FIG. 7. For the same system in Fig. 1, but $a = 2.0$ cm, $b = 3.0$ cm, $\hat{e} = 0.0$ V/cm, and $e = 4.0$ V/cm.

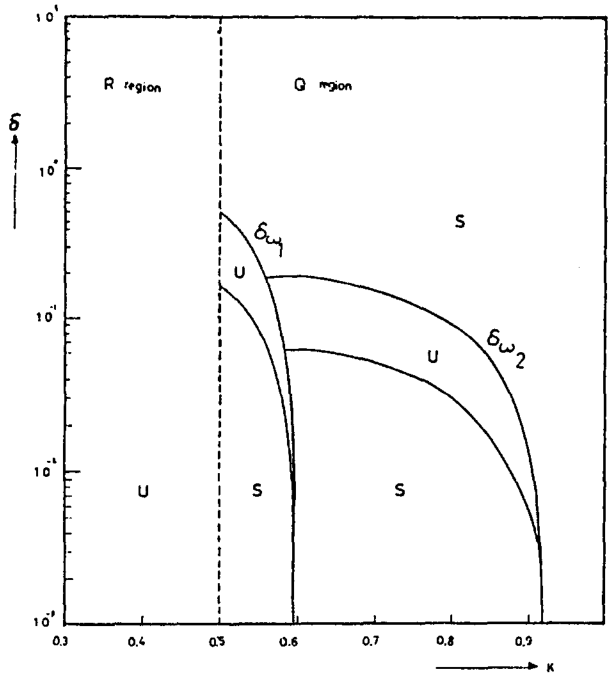


FIG. 8. For the same system as in Fig. 1, but $a = 2.0$ cm, $b = 3.0$ cm, $\hat{e} = 0.0$ V/cm, and $e = 2.0$ V/cm.

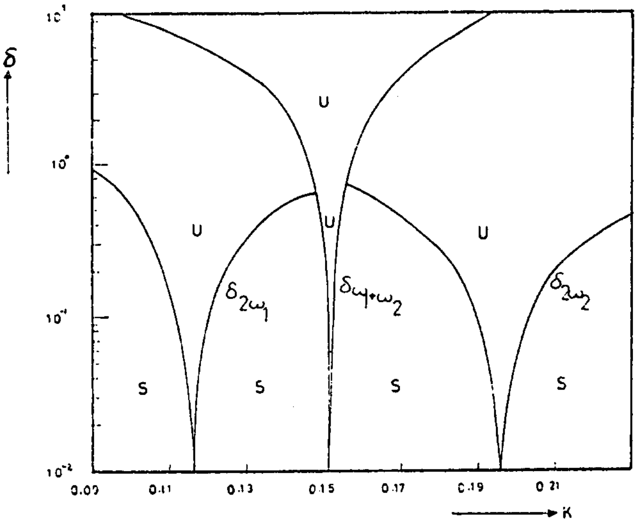


FIG. 9. A representation of the same system as in Fig. 1, but $\hat{e} = 0.0$ V/cm, and $\omega = 8.0$ Hz. Here $E_0 = 5.0$ V/cm.

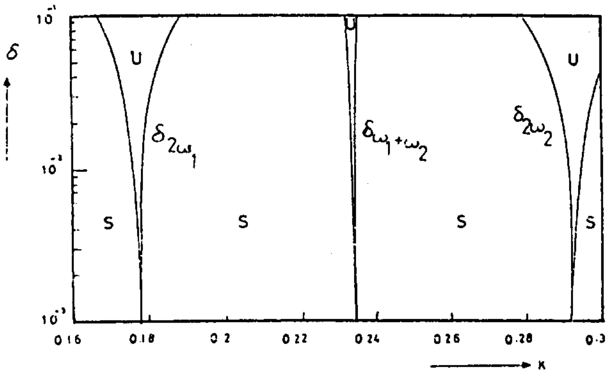


FIG. 10. For the system considered in Fig. 9, but $E_0 = 3.0$ V/cm.

VIII. STABILITY OF THE SMALL MODULATION

This special case means that $E_0 \gg E_1^*$. We set $E^* = \delta e$ and hence

$$(E_0 + E^* \cos \omega t)^2 = E_0^2 + 2\delta e \cos \omega t.$$

So, the corresponding coupled equations are reduced to

$$\ddot{\gamma}_1 + [a_1' + 2\delta b_1'(e \cos \omega t)]\gamma_1 + [a_2' + 2\delta b_2'(e \cos \omega t)]\gamma_2 = 0, \tag{50}$$

$$\ddot{\gamma}_2 + [a_3' + 2\delta b_3'(e \cos \omega t)]\gamma_2 + [a_4' + 2\delta b_4'(e \cos \omega t)]\gamma_1 = 0, \tag{51}$$

where

$$a_j' = a_j + 4b_j E_0^2, \quad b_j' = 4E_0 b_j, \quad j = 1,2,3,4, \tag{52}$$

and a_j, b_j are as defined in Appendix A. Noting that, the resonance here corresponds to ω near $2\omega_1'$, ω near $2\omega_2'$ and ω near $(\omega_1' \pm \omega_2')$, where the $\omega_{1,2}'^2$ are the roots of the equation

$$\omega_0^4 - (a_1' + a_3')\omega_0^2 + (a_1' a_3' - a_2' a_4') = 0. \tag{53}$$

If the special case of small modulation is the characteristic of the system, then the resonance curves for the present system are $\delta_{2\omega_1'}$, $\delta_{2\omega_2'}$, and $\delta_{\omega_1' + \omega_2'}$ only.

Figure 9 represents the δ-k plane for a system having $\rho_i = 0.001\,293$ g/cm^3, $\rho_{be} = 0.998\,23$ g/cm^3, $\rho_o = 0.879$ g/cm^3, $\epsilon^i = 1.000\,59$, $\epsilon^{be} = 80.37$, $\epsilon^o = 2.284$, $T_1 = 72.75$ dyn/cm, $T_2 = 35.0$ dyn/cm, $a = 2.0$ cm, $b = 3.0$ cm, $E_0 = 5.0$ V/cm, and $\omega = 8.0$ Hz. The system here is different from the other cases, where initially we have a constant tangential field and its influence is observed here when the critical value for $(x = ka)$ is less than unity. The δ-k plane indicates stable regions except at ω near $2\omega_1'$, ω near $2\omega_2'$, and ω near $\omega_1' + \omega_2'$. All the regions are very thin. This means that the unstable effect of the periodicity is greatly reduced.

Figure 10 is similar to Fig. 9, but with $E_0 = 3.0$ V/cm. The resonance regions for the case of large modulation are relatively wider than those of small modulation. This illustrates the destabilizing role of periodicity. The role is enhanced by the domination of the modulation term.

APPENDIX A: THE VALUES OF THE COEFFICIENTS a_j, b_j, f_j

The values of the a_j's appearing in Eqs. (19) and (20) are

$$a_1 = \frac{T_1}{a^2}(1 - m^2 - x^2)\frac{\rho_{be} L_x^m K_m'(y) - \rho_o K_m(y) L_{xy}^m}{\Delta . K_m'(y)}, \quad a_2 = \frac{T_2}{b^2}(y^2 + m^2 - 1)\frac{\rho_{be}}{\Delta . x},$$

$$a_3 = \frac{T_2}{b^2}(1 - m^2 - y^2)\frac{\rho_i I_m(x) L_{xy}^m - \rho_{be} L_y^m I_m'(y)}{\Delta . I_m'(x)}, \quad a_4 = \frac{T_1}{a^2}(x^2 + m^2 - 1)\frac{\rho_{be}}{\Delta . y},$$

where

$$\Delta = (1/k L_{xy}^m)\,[\rho_{be}^2/xy\,(\rho_{be} L_x^m K_m'(y) - \rho_o K_m(y) L_{xy}^m)\,(\rho_i I_m(x) L_{xy}^m - \rho_{be} L_x^m I_m'(x)/(K_m'(y) I_m'(x)))].$$

The values of the b_j's involved in Eqs. (19) and (20) are

$$b_1 = \frac{k}{\psi(x,y)}\Bigg\{\frac{\rho_{be}}{x}\Bigg[\epsilon^o \epsilon^{be}(\epsilon^i - \epsilon^{be}) K_m(y)\frac{I_m(x)}{y} - \epsilon^{be}(\epsilon^o - \epsilon^{be})(\epsilon^i - \epsilon^{be}) I_m(y) K_m(y) I_m(x) K_m'(y)$$
$$+ \epsilon^{be}(\epsilon^i - \epsilon^{be})(\epsilon^o K_m'(y) I_m(y) - \epsilon_{be} K_m(y) I_m'(y)) K_m(y) I_m(x)\Bigg] - \frac{(\rho_{be} L_x^m K_m'(y) - \rho_o K_m(y) L_{xy}^m)}{K_m'(y)}$$
$$\times [\epsilon^{be}(\epsilon^o - \epsilon^{be})(\epsilon^i - \epsilon^{be}) I_m^2(x) K_m(y) K_m'(y) - \epsilon^{be}(\epsilon^i - \epsilon^{be})(\epsilon^o K_m'(y) I_m(y) - \epsilon^{be} K_m(y) I_m'(y)) K_m(x) I_m(x)$$
$$- \epsilon^i(\epsilon^i - \epsilon^{be})(\epsilon^{be} L_y^m K_m(y) - \epsilon^o L^m K_m'(y)) I_m(x)]\Bigg\}\Delta^{-1},$$

$$b_2=\frac{-k}{\psi(x,y)}\left\{\frac{(\rho_{be}L_x^m K'_m(y)-\rho_o K_m(y)L_{xy}^m)}{K'_m(y)}\left[\epsilon^{be}(\epsilon^{be}-\epsilon^o)(\epsilon^i K_m(x)I'_m(x)-\epsilon^{be}K'_m(x)I_m(x))I_m(x)K_m(y)\right.\right.$$

$$\left.+\epsilon^{be}(\epsilon^i-\epsilon^{be})(\epsilon^o-\epsilon^{be})K_m(x)I_m(x)K_m(y)I'_m(x)+\epsilon^i\epsilon^{be}(\epsilon^o-\epsilon^{be})\frac{I_m(x)K_m(y)}{x}\right]-\frac{\rho_{be}}{x}\left[-\epsilon^o(\epsilon^o-\epsilon^{be})(\epsilon^i L^m I'_m(x)\right.$$

$$-\epsilon^{be}L_x^m I_m(x))K_m(y)+\epsilon^{be}(\epsilon^o-\epsilon^{be})(\epsilon^i K_m(x)I'_m(x)-\epsilon^{be}I_m(x)K'_m(x))I_m(y)K_m(y)$$

$$\left.\left.-\epsilon^{be}(\epsilon^i-\epsilon^{be})(\epsilon^o-\epsilon^{be})K_m^2(y)I_m(x)I'_m(x)\right]\right\}\Delta^{-1},$$

$$b_3=\frac{k}{\psi(x,y)}\left\{\frac{\rho_{be}}{y}\left[-\epsilon^{be}(\epsilon^o-\epsilon^{be})(\epsilon^i K_m(x)I'_m(x)-\epsilon^{be}K'_m(x)I_m(x))I_m(x)K_m(y)\right.\right.$$

$$+\left[\epsilon^{be}(\epsilon^i-\epsilon^{be})(\epsilon^o-\epsilon^{be})K_m(x)I_m(x)K_m(y)I'_m(x)+\epsilon^i\epsilon^{be}(\epsilon^o-\epsilon^{be})\frac{I_m(x)K_m(y)}{x}\right]-\left(\frac{\rho_i I_m(x)L_{xy}^m-\rho_{be}L_y^m I'_m(x)}{I'_m(x)}\right)$$

$$\times\left[\epsilon^o(\epsilon^{be}-\epsilon^o)(\epsilon^i L^m I'_m(x)-\epsilon^{be}L_x^m I_m(x))K_m(y)+\epsilon^{be}(\epsilon^o-\epsilon^{be})(\epsilon^i K_m(x)I'_m(x)\right.$$

$$\left.\left.\left.-\epsilon^{be}I_m(x)K'_m(x))I_m(y)K_m(y)-\epsilon^{be}(\epsilon^i-\epsilon^{be})(\epsilon^o-\epsilon^{be})K_m^2(y)I_m(x)I'_m(x)\right]\right]\right\}\Delta^{-1},$$

$$b_4=\frac{-k}{\psi(x,y)}\left\{\frac{(\rho_i I_m(x)L_{xy}^m-\rho_{be}L_y^m I'_m(x))}{I_m(x)}\left[\epsilon^o\epsilon^{be}(\epsilon^i-\epsilon^{be})\frac{K_m(y)I_m(x)}{y}-\epsilon^{be}(\epsilon^{be}-\epsilon^o)(\epsilon^i-\epsilon^{be})I_m(y)K_m(y)I'_m(x)K'_m(y)\right.\right.$$

$$\left.+\epsilon^{be}(\epsilon^i-\epsilon^{be})(\epsilon^o K'_m(y)I_m(y)-\epsilon^{be}K_m(y)I'_m(y))K_m(y)I_m(x)\right]-\frac{\rho_{be}}{y}\left[\epsilon^{be}(\epsilon^o-\epsilon^{be})(\epsilon^i-\epsilon^{be})I_m^2(x)K_m(y)K'_m(y)\right.$$

$$+\epsilon^{be}(\epsilon^{be}-\epsilon^i)(\epsilon^o K'_m(y)I_m(y)-\epsilon^{be}K_m(y)I'_m(y))K_m(x)I_m(x)$$

$$\left.+\{[\epsilon^i(\epsilon^{be}-\epsilon^i)(\epsilon^{be}L_y^m K_m(y)-\epsilon^o L'_y(y)K'_m(y))I_m(x)]\right\}\Delta^{-1},$$

where

$$L^m=I_m(y)K_m(x)-I_m(x)K_m(y),\quad L_{xy}^m=I'_m(y)K'_m(x)-I'_m(x)K'_m(y),$$

$$L_y^m=I'_m(y)K_m(x)-I_m(x)K'_m(y),\quad L_x^m=I_m(y)K'_m(x)-I'_m(x)K_m(y),$$

and

$$\psi(x,y)=\epsilon^i\epsilon^o I'_m(x)K'_m(y)L^m-[\epsilon^o I_m(x)K'_m(y)L_x^m+\epsilon^i K_m(y)I'_m(x)L_y^m]\epsilon^{be}+I_m(x)K_m(y)L_{xy}^m(\epsilon^{be})^2.$$

The values of the f_j's appearing in Eq. (45) are

$$f_1=\{[\sigma^*-\nu_2(e^2+2\hat{e}^2)]^2+[\sigma^*-\nu_1(e^2+2\hat{e}^2)]^2-2\mu_1\mu_2e^4\},$$

and

$$f_2=\{[\sigma^*-\nu_1(e^2+2\hat{e}^2)]\,[\sigma^*-\nu_2(e^2+2\hat{e}^2)]-\mu_1\mu_2e^4\},$$

where

$$\mu_j=[a_2(b_4a_2-b_3a_1)+a_3(b_2a_1-b_1a_2)+b_2(a_1a_3-a_2a_4)+(a_2b_3-a_3b_2)\omega_j^2+(b_1a_3-b_3a_1)\omega_{j+1}^2]$$
$$\times\{2a_2\omega_{j+1}\,[2\omega_{j+1}^2-a_1a_3]\}^{-1}.$$

APPENDIX B: PROOF OF $\omega_{1,2}^2$ REAL

We show that the $\omega_{1,2}^2$, in Eq. (31), are real [we consider the case of the axisymmetric mode ($m=0$) which is the most unstable one]. The radical in Eq. (31) is

$$(a_1-a_3)^2+4a_2a_4=0,$$

by using the values of the a_j's that are given in Appendix A, the above equation can be rewritten as follows:

$$\left(\frac{\rho_{be}L_x^o K_1(x)+\rho_o K_o(y)L_{xy}^o}{K_1(y)}\right)\left[\frac{T_1(1-x^2)b^2}{T_2(1-y^2)a^2}\right]^2+2\left\{\frac{\rho_{be}^2}{xy}-\frac{[\rho_{be}L_x^o K_1(x)+\rho_o K_o(y)L_{xy}^o]}{K_1(y)}\right.$$

$$\left.\times\frac{[\rho_i I_o(x)L_{xy}^o+\rho_{be}L_y^o I_1(x)]}{I_1(x)}\right\}\left[\frac{T_1(1-x^2)b^2}{T_2(1-y^2)a^2}\right]+\frac{[\rho_i I_o(x)L_{xy}^o+\rho_{be}L_y^o I_1(x)]_2}{I_1(x)}=0,$$

which gives a quadratic in $[T_1(1-x^2)b^2/T_2(1-y)^2b^2]$, with the discriminant

$$D=-(16\rho_{be}^2/xy)\{\rho_{be}^2L^oL_{xy}^o+(\rho_{be}\,\rho_iL_x^o\,L_{xy}^oK_1(y)I_o(x)$$
$$+\rho_o\,\rho_iK_o(y)I_o(x)(L_{xy}^o)^2+\rho_o\,\rho_{be}\,K_o(y)I_1(x)L_{xy}^o\,L_y^o)/I_1(x)K_1(y)\}<0.$$

Since the coefficient of $[T_1(1-x^2)b^2/T_2(1-y^2)a^2]^2$ is positive definite, the result follows.

Note that

$$L_{xy}^{0} = -\lim_{m\to 0} L_{xy}^{m} = I_1(y)K_1(x) - I_1(x)K_1(y), \quad L_x^0 = -\lim_{m\to 0} L_x^m = I_0(y)K_1(x) + I_1(x)K_0(y),$$

$$L_y^0 = \lim_{m\to 0} L_y^m = I_1(y)K_0(x) + I_0(x)K_1(y), \quad L^0 = \lim_{m\to 0} L^m = I_0(y)K_0(x) - I_0(x)K_0(y).$$

It follows directly from the above relations that

$$L_x^0 L_y^0 = L^0 L_{xy}^0 + 1/xy.$$

[1]A. B. Bassat, Am. J. Math. **16**, 93 (1894).
[2]J. W. S. Rayleigh, *Theory of Sound* (Dover, New York, 1945), Vol. II.
[3]J. H. Dumbleton and J. J. Hermans, Phys. Fluids **13**, 12 (1970).
[4]N. K. Nayyar and G. S. Murty, Proc. Phys. Soc., London **75**, 369 (1960).
[5]A. A. Mohamed and N. K. Nayyar, J. Phys. A **4**, 296 (1970).
[6]A. A. Mohamed and N. T. El Dabe, Proc. Math. Phys. Soc., Egypt **45**, 53 (1978).
[7]H. H. Woodson and J. R. Melcher, *Electromechanical Dynamics* (Wiley, New York, 1968).
[8]J. R. Melcher, *Field Coupled Surface Waves* (MIT, Cambridge, MA, 1963).
[9]C. A. Erdely, *Higher Transcendental Functions* (McGraw–Hill, New York, 1953), Vol. II.
[10]N. W. McLachlan, *Theory and Application of Mathieu Functions* (Dover, New York, 1964).
[11]A. H. Nayfeh, J. Acoust. Soc. Am. **62**, 375 (1977).
[12]A. H. Nayfeh, *Nonlinear Oscillation* (Wiley, New York, 1979).
[13]A. A. Mohamed and N. K. Nayyar, Arab. J. Sci. Eng. **8**, 159 (1983).

ERRATA

Erratum: A novel class of Bessel function integrals [J. Math. Phys. 25, 2933 (1984)]

M. L. Glasser
Department of Mathematics and Computer Science, Clarkson College of Technology, Potsdam, New York 13676

(Received 29 October 1984; accepted for publication 16 November 1984)

(1) α and β are reversed in Eq. (12).

(2) The exponent -1 should be deleted from α in the second line of (17).

(3) The argument of si should be doubled in (18) and (19).

(4) A plus sign should be inserted between α and β on the right-hand side of (20).

(5) "be" should be replaced by "give" on the line above (24), which should read

$$B_0 = 2[\lambda^{-1}\cos(2\lambda) + \mathrm{si}(2\lambda)].$$

Erratum: Triality principle and G_2 group in spinor language [J. Math. Phys. 26, 6 (1985)]

Z. Hasiewicz and A. K. Kwasniewski
Institute of Theoretical Physics, University of Wrocław, ul. Cybulskiego 36, 50-205 Wrocław, Poland

(Received 22 January 1985; accepted for publication 22 January 1985)

The 12th line from the bottom, left column, p. 8 should read

$$= (h_{v,\eta} \circ i_v)(h_{v,\eta}(\psi) h_{v,\eta}^{-1} \circ h_{v,\eta}(\psi')) .$$

 0022-2488/85/082082-01$02.50