Representations of E_8 and other algebras

Richard H. Capps
Department of Physics, Purdue University, West Lafayette, Indiana 47907

(Received 14 August 1986; accepted for publication 18 February 1987)

A simple procedure is given for determining whether or not an arbitrary weight of a simple Lie algebra is contained in an arbitrary irreducible representation. The procedure involves two steps, determining the Weyl class of the weight and determining the classes contained in the representation. The second step applies without alteration to all simple algebras, while the first step is given here only for E_8. The weights of the shortest 31 Weyl classes of E_8 are listed in a convenient, orthogonal basis.

I. INTRODUCTION

In the application of finite, simple Lie algebras a common problem is to find whether or not a weight M is contained in an irrep (irreducible representation) Λ. Dynkin developed a constructive method that may be used to solve this problem for any M and Λ (Ref. 1). Descriptions of the procedure are given in various reviews and texts.[2,3] Unfortunately, the method is cumbersome for all but the smallest representations, and the number of steps involved increases without limit as one considers very large irreps. One purpose of this paper is to present an alternate method involving a number of steps that is small for all irreps.

The new method is applied here to the algebra E_8. Recent developments in string theory suggest that this algebra is important in particle physics.[4] The properties of the smallest nontrivial irrep (the adjoint) have been discussed in detail in the literature. However, the other irreps are still unfamiliar. The main reason for this is that the irreps are very large; for example, the sixth smallest irrep has dimension 147 250. A second purpose of this paper is to provide a tractable treatment of E_8 irreps. The weights of the shortest 31 irreps are given in a convenient basis. (The length of an irrep is defined to be the length of the longest weights in the irrep.)

The method involves the Weyl reflection group. The weights that may be obtained from a given weight by a series of Weyl reflections are said to be in the same Weyl class (or Weyl orbit). The Weyl class is a convenient concept, for three well-known reasons. First, each weight is in one and only one Weyl class. Second, the multiplicity of a weight in an irrep is the same for all weights in a class. Third, the sizes of all Weyl classes for an algebra are bounded by the finite size of the Weyl group.

The method involves two separate algorithms. The first, the "weight algorithm," is for finding the Weyl class of an arbitrary weight. The second, the "representation algorithm" is for determining whether or not a given class is contained in a given irrep. The representation algorithm is discussed first, because the method used is general and may be applied to any finite simple Lie algebra. The weight algorithm involves the choice of a convenient basis. Although the method for selecting a basis is general, the details of application depend on the basis and thus are different for different algebras. Identifying the class of a weight is nontrivial only for the algebras G_2, F_4, E_6, E_7, and E_8. The algebra E_8 is treated here.

If a Weyl class is in an irrep, the multiplicity may be obtained from the Freudenthal recursion formula,[2,3,5] or from published tables.[6] Multiplicities are not given here.

The basic group-theoretical concepts and formulas that are used in the paper are listed in Sec. II. The representation algorithm is derived and applied in Sec. III. Section IV contains the weight algorithm for E_8, an iterative procedure involving either zero, one, or two iterations. The basic reasons, and a proof, that two iterations are sufficient are given in Sec. V.

II. BASIC CONCEPTS AND FORMULAS

The standard Cartan–Weyl construction is used. If n is the rank of a simple algebra, n commuting generators, denoted by H_1 to H_n, are diagonalized in each irrep. The weight vector M of a state in an irrep is a vector in an n-dimensional Euclidean space, with real components f_i given by $H_i M = f_i M$.

The roots are the weights of the adjoint representation. Following the Dynkin method, I label the orthogonal axes 1 through n, and define a weight as positive (or negative) if its first nonzero component is positive (or negative). A simple root is a positive root that cannot be written as a sum of two positive roots; there are n simple roots. If M is a weight in an irrep, and α is a nonzero root, a fundamental equation is,[7]

$$\langle M,\alpha\rangle(2/\alpha^2) = p_{\alpha-} - p_{\alpha+}, \tag{2.1}$$

where the non-negative integers $p_{\alpha-}$ and $p_{\alpha+}$ are the maximum numbers of times the root α may be subtracted from M, and may be added to M, to obtain other weights in the irrep. If α is a simple root R_j, the integer $p_{\alpha-} - p_{\alpha+}$ is the Dynkin component m_j, i.e.,

$$m_j = \langle M,R_j\rangle(2/R_j^2). \tag{2.2}$$

A dominant weight, denoted by M^+, is one for which all the Dynkin components are non-negative. Most irreps contain more than one dominant weight. However, the most positive weight in the irrep must be dominant, and is used to characterize the irrep. Each dominant weight is the most positive weight of a unique irrep.

The root-basis components and Dynkin components of a weight are denoted by capital and small letters, respectively. The root-basis components are the coefficients in an expansion in the simple roots, i.e., $M = \Sigma_i M_i R_i$. They are related to the Dynkin component m_j by[8]

 0022-2488/87/061215-08$02.50

$$m_j = \sum_i M_i A_{ij}, \tag{2.3a}$$

$$M_i = \sum_j m_j (A^{-1})_{ji}, \tag{2.3b}$$

where the elements of the Cartan matrix A are given by

$$A_{ij} = \langle R_i, R_j \rangle (2/R_j^2). \tag{2.4}$$

A scalar product of two vectors is expressed simply if one vector is in the Dynkin basis and the other is in the root basis. The expression is

$$\langle M, W \rangle = \sum_i m_i W_i (R_i^2/2). \tag{2.5}$$

The length squared of a vector M may be determined from the formula,

$$M^2 = \sum_{ij} m_i m_j G_{ij}, \tag{2.6}$$

where G is the symmetric metric tensor, related to A^{-1} by

$$G_{ij} = (A^{-1})_{ij} (R_j^2/2). \tag{2.7}$$

For every nonzero root α there is a Weyl reflection operator S_α, which permutes the weights in an irrep. The action of S_α on a weight M is given by[3]

$$S_\alpha(M) = M - (2/\alpha^2)\langle M, \alpha \rangle \alpha. \tag{2.8}$$

All the weights in a Weyl class (related by one or more Weyl reflections) are of the same length, called here the length of the class. The most positive weight in a class is dominant; all other weights are not dominant.

III. THE REPRESENTATION ALGORITHM

The set of weights for an algebra is the set of vectors with integral Dynkin components. The representation algorithm makes use of the following "containment criterion": the weight m is in the irrep with most positive weight Λ if and only if the root-basis components of $\Lambda - M^+$ are all non-negative integers, where M^+ is the dominant weight of the Weyl class of M. The validity of the criterion is proved at the end of this section.[9]

Equation (2.3b) may be used to base a simple algorithm on this criterion. I denote $\Lambda - M^+$ by Δ. The rule is that the root-basis components Δ_i must be non-negative integers, where

$$\Delta_i = \sum_j (\lambda_j - m_j^+)(A^{-1})_{ji}. \tag{3.1}$$

To my knowledge, this simple rule is not contained in the literature.

I illustrate the algorithm with an example from the algebra F_4. The A^{-1} matrix for F_4 is given by

$$A^{-1} = \begin{bmatrix} 2 & 3 & 4 & 2 \\ 3 & 6 & 8 & 4 \\ 2 & 4 & 6 & 3 \\ 1 & 2 & 3 & 2 \end{bmatrix}, \tag{3.2}$$

where the root numbering convention is that of Fig. 1. We consider the two dominant weights B and C, with Dynkin components $B = (2000)$ and $C = (0011)$, and ask whether or not either of the corresponding Weyl classes is the irrep of the other weight. From Eqs. (3.1) and (3.2), the root basis components of B and C are

$$B = [4684], \quad C = [3695].$$

Since the difference $B - C$ contains both positive and negative root-basis components, neither is in the irrep of the other. [The lengths of these Weyl classes, determined from Eqs. (2.6), (2.7), and (3.2), and the convention that the short and long root lengths are 1 and $\sqrt{2}$, are $B^2 = 8$ and $C^2 = 7$.]

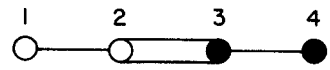

FIG. 1. Dynkin diagram for F_4, with the root-numbering convention. The shaded circles denote short roots.

For some algebras the weights are in two or more congruence classes. All weights in any Weyl class, or in any irrep, are in the same congruence class. If M^+ is in a different congruence class from Λ, the Δ_i of Eq. (3.1) are not all integers. This follows because every nontrivial congruence relation may be associated with a column of the A^{-1} matrix in which at least one of the elements is nonintegral.[10] Let such a column be the k th column. Then, if C is the smallest positive integer such that the product $C\,(A^{-1})_{ik}$ is an integer for each i, the congruence class of a weight with Dynkin indices a_i is

$$\sum_i a_i C (A^{-1})_{ik} \pmod{C}.$$

Clearly, $\Sigma_i a_i (A^{-1})_{ik}$ is an integer only for class C (class 0). The containment criterion is valid when M and Λ are in different congruence classes, but is not useful in such a case, since one recognizes the incompatibility of M and Λ from the congruence relation.

Another useful tool is the well-known rule: if M^+ and Λ are different dominant weights, and if the Weyl class of M^+ is in the irrep Λ, then,

$$\Lambda^2 > (M^+)^2. \tag{3.3}$$

This rule may be proved by using Eq. (2.5), i.e.,

$$\Lambda^2 - (M^+)^2 = 2\langle M^+, \Delta \rangle + \Delta^2 = \sum_i m_i^+ \Delta_i R_i^2 + \Delta^2. \tag{3.4}$$

Since the m_i^+ and Δ_i are all non-negative, the right-hand term of Eq. (3.4) is positive, proving the rule.

I will use the algorithm to list the Weyl classes in all irreps of E_8 that are not longer than $(32)^{1/2}$. The normalization convention is that the nonzero roots are of length $\sqrt{2}$. As seen from Eq. (2.7), the A^{-1} and G matrices are identical. The matrix is[11]

$$A^{-1} = \begin{bmatrix} 4 & 7 & 10 & 8 & 6 & 4 & 2 & 5 \\ 7 & 14 & 20 & 16 & 12 & 8 & 4 & 10 \\ 10 & 20 & 30 & 24 & 18 & 12 & 6 & 15 \\ 8 & 16 & 24 & 20 & 15 & 10 & 5 & 12 \\ 6 & 12 & 18 & 15 & 12 & 8 & 4 & 9 \\ 4 & 8 & 12 & 10 & 8 & 6 & 3 & 6 \\ 2 & 4 & 6 & 5 & 4 & 3 & 2 & 3 \\ 5 & 10 & 15 & 12 & 9 & 6 & 3 & 8 \end{bmatrix}. \tag{3.5}$$

The root-numbering convention may be obtained by delet-

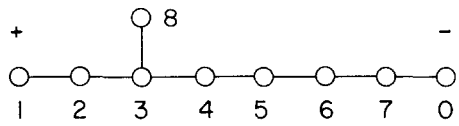

FIG. 2. Extended Dynkin diagram illustrating the $D_8 \rightarrow E_8$ basis. The E_8 simple roots are numbered 1 to 8.

ing the root 0 and its connecting line from Fig. 2. The elements of A^{-1} are all integral, so all irreps are in the same congruence class.

It is easy to use Eqs. (2.6) and (3.5) and the requirement that the Dynkin components of dominant weights are non-negative integers to determine all the Weyl classes of E_8 such that $(M^+)^2 \leqslant 32$. These classes are listed in order of increasing length as the boldface entries in Table I. The subscripts a, b, c, and d are used to label different classes of the same length, i.e., **14a** and **14b** are the two classes of length $(14)^{1/2}$. The symbols in curly brackets are an abbreviation for the Dynkin components; e.g., $\{1^2 7\}$ denotes the Dynkin components (2000 0010).

Because of the rule of Eq. (3.3) each irrep contains its own Weyl class (class of the most positive weight), and some other shorter classes. If B and C are dominant weights in the same congruence class, and $B^2 > C^2$, usually $B \supset C$ (irrep B contains class C). Thus I define an anomalous pair of Weyl classes as two classes of the same congruence of unequal length, such that the shorter is not contained in the irrep of the longer. In Table I the classes in parentheses are the shorter anomalous partners of the preceding boldface entries. An x indicates that there are no shorter anomalous partners.

I discuss briefly the calculation of the table entries. The easiest procedure is to consider the irreps in order of increasing length, checking each for shorter anomalous partners. One can use the fact, obvious from the contents criterion, that if $B \supset C$ and $C \supset D$, then $B \supset D$. I illustrate by considering the irrep **16a**, with Dynkin components $\{1^2\}$. I assume that the shorter irreps have all been examined and found to contain no shorter anomalous partners. The root-basis components of $\{1^2\}$ are twice the elements in the first row of A^{-1}, i.e.,

$$\textbf{16a}\{1^2\} = [8\ 14\ 20\ 16\ 12\ 8\ 4\ 10]. \tag{3.6}$$

One uses Eq. (2.3b) to find the corresponding root-basis components of the dominant weights immediately preceding **16a** in Table I. These are

$$\textbf{14a}\{2\} = [7\ 14\ 20\ 16\ 12\ 8\ 4\ 10], \tag{3.7a}$$

$$\textbf{14b}\{67\} = [6\ 12\ 18\ 15\ 12\ 9\ 5\ 9]. \tag{3.7b}$$

TABLE I. Anomalous pairs of E_8 Weyl classes with $M_2 \leqslant 32$.

0 $\{0\}x$; **2** $\{7\}x$; **4** $\{1\}x$; **6** $\{6\}x$; **8a** $\{8\}x$; **8b** $\{7^2\}x$;
10 $\{17\}x$; **12** $\{5\}x$; **14a** $\{2\}x$; **14b** $\{67\}x$; **16a**$\{1^2\}$(14*b*);
16b $\{78\}x$; **18a** $\{16\}x$; **18b** $\{7^3\}$(14a,16a,16b); **20a** $\{17^2\}x$;
20b $\{4\}$(18b); **22a** $\{18\}$(18b,20a); **22b** $\{57\}x$; **24a** $\{27\}x$;
24b $\{6^2\}$(22a); **26a** $\{1^2 7\}$(24b); **26b** $\{68\}x$; **26c** $\{67^2\}$(22a,24a);
28a $\{15\}$(26c); **28b** $\{7^2 8\}$(26a); **30a** $\{167\}x$; **30b** $\{3\}$(26c,28b);
32a $\{12\}$(26c,28b,30a); **32b** $\{47\}x$; **32c** $\{8^2\}$(26c,28b,30a);
32d $\{7^4\}$(22a,24a,26a,26b,28b,30a,30b).

Comparing Eq. (3.6) with Eqs. (3.7a) and (3.7b), one sees that the pair **16a** and **14b** is anomalous, but **14a** is contained in **16a**. Since all shorter classes are contained in **14a**, one concludes that **14b** is the only shorter anomalous partner of **16a**. Thus the irrep **16a** contains the class **16a** and all shorter classes except **14b**. We note that the root-basis difference between Eqs. (3.7a) and (3.7b) contains elements of both signs, but this follows automatically from the fact that the two classes are of the same length.

Finally, I will prove the validity of the containment criterion, stated at the beginning of the section. One-half of the proof is obvious. Since the standard procedure for finding weights in an irrep involves subtracting simple roots from Λ, it is clear that if the root-basis components of $\Delta = \Lambda - M^+$ are not all non-negative integers, M^+ (and hence M) is not in Λ. Hence we turn to the case where the Δ_i are all non-negative integers.

We neglect the trivial case $\Delta = 0$. Then Δ is a positive weight. It follows that at least one of its Dynkin components must be positive, since weights with all nonpositive Dynkin components are the most negative weights of irreps, and so are not positive. Let δ_k be a positive Dynkin component. Since the Δ_j are non-negative and the off-diagonal elements of the Cartan matrix are nonpositive, it follows from the transformation rule of Eq. (2.3a) that $\Delta_k > 0$. Since Δ_k is integral,

$$\Delta_k \geqslant 1. \tag{3.8}$$

The Dynkin component λ_k of Λ satisfies the equation, $\lambda_k = \delta_k + m_k^+$. Since m_k^+ is non-negative and δ_k is positive, $\lambda_k \geqslant 1$. It follows from Eq. (2.1) that one may subtract the simple root R_k from Λ, i.e., the weight $\Lambda' = \Lambda - R_k$ is in the irrep.

One next considers the pair of weights Λ' and M^+. The new difference weight $\Delta' = \Lambda' - M^+$ is related to Δ by the equation,

$$\Delta_i' = \Delta_i - \delta_{ik}.$$

It follows from Eq. (3.8) that the root-basis components Δ_i' are all non-negative integers, so that process may be repeated, until a path is traced from Λ to M^+, proving M^+ is in Λ. This completes the proof.

IV. THE WEIGHT ALGORITHM

A. The problem

In this section I discuss the problem of finding the E_8 Weyl class of an arbitrary weight, expressed in Dynkin components. There exists a standard, finite procedure for finding the answer. One makes a series of Weyl reflections S_R, always choosing R to be a simple root corresponding to a negative Dynkin component, so that the reflection leads to a more positive weight. Such a series is called a "Dynkin reflection series" in this paper. It is easy to make each reflection; if the jth Dynkin component is $-b$ (where b is positive) the S_j reflection adds b times the simple root R_j. Unfortunately, the number of reflections necessary to obtain a dominant weight may be as large as the number of positive roots, 120 in the case of E_8. Clearly, a shorter procedure is needed.

B. The D_8 basis

For each of the classical algebras a basis exists in which the Weyl class of a weight is obvious. Therefore it is convenient to express the E_8 weights in a classical-algebra basis. Such a basis has the additional advantage of giving a clearer picture of the E_8 weights. A convenient basis for E_8 is based on D_8. I will define this basis here, and apply it to E_8 in part C of this section.

In a standard, orthogonal D_8 basis each nonzero root has two components of magnitude 1, the other components being zero.[12] A shorthand notation is used for the roots in this basis: i.e., 3_+5_- denotes [0010 −1000]. The simple roots are 1_+2_-, 2_+3_-, 3_+4_-, 4_+5_-, 5_+6_-, 6_+7_-, 7_+8_-, and 7_+8_+. The Dynkin diagram is obtained by deleting the root labeled + and its connecting line from Fig. 2. The simple roots listed above correspond to the circles numbered 0, 7, 6, 5, 4, 3, 2, and 8, respectively.

The components of a weight in the orthogonal basis are denoted by f_i. It is seen from Eq. (2.2) that the set of weights for D_8 is the set of vectors that have integral scalar products with all roots. The weights are of two types. The vector weights are all vectors such that each component f_i is integral. The spinor weights are all vectors such that each component is half-odd-integral. Both the vector and spinor weights may be classified as even or odd, according to whether the component sum $\Sigma_{i=1}^{8} f_i$ is even or odd. These four classes are the four congruence classes of D_8. It is clear that adding a root to a weight leads to another weight of the same congruence class.

Next we examine the structure of the Weyl classes in the orthogonal basis. It follows from Eq. (2.8) that the Weyl reflection corresponding to the root j_+k_- (or j_-k_+) interchanges the j and k components of a weight. The reflection corresponding to the root j_+k_+ (or j_-k_-) interchanges the j and k components and changes both their signs. It is convenient to define the signature of a weight as zero, positive, or negative, corresponding to the component product $\Pi_i f_i$ being zero, positive, or negative. Weyl reflections preserve the signature.

It is clear from the above discussion that if one is given a D_8 weight in the orthogonal basis, he may determine the dominant weight of the Weyl class immediately. The components are arranged in order of decreasing magnitude, and the signs of the first seven components are chosen to be positive. If the eighth component is nonzero, its sign is chosen to be the signature of the weight. A weight is dominant if and only if it satisfies the conditions

$$|f_i| \geqslant |f_{i+1}|, \tag{4.1a}$$

$$f_i \geqslant 0 \quad \text{for } i < 8. \tag{4.1b}$$

C. Application to E_8

An E_8 basis based on D_8 may be found by the replacement prescription of Ref. 10. One writes the simple roots of D_8 and adds to this set the most negative weight of a D_8 irrep, such that the weight length is $\sqrt{2}$. If the irrep is the odd-signature fundamental spinor, the new root is [− − − − − − − +] (where the magnitudes are all $\frac{1}{2}$). One then discards the D_8 root 1_+2_- and reflects the first axis, so that the eight remaining roots are positive. The new root set is the simple root set of E_8. Ordered according to the numbers 1 to 8 of Fig. 2, the E_8 simple roots are

$$\begin{aligned} &[\,+\,-\,-\,-\;\;-\,-\,-\,+\,], \\ &7_+8_-,\;\; 6_+7_-,\;\; 5_+6_-, \\ &4_+5_-,\;\; 3_+4_-,\;\; 2_+3_-, \text{ and } 7_+8_+. \end{aligned} \tag{4.2}$$

The replacement procedure is useful because it allows one to consider all bases of a certain type. However, it has been known for years that a convenient E_8 basis may be obtained by taking as E_8 roots the D_8 roots plus the states of the even-signature fundamental spinor.[13] With the usual positivity definition (Sec. II) the simple roots of this set are those of Eq. (4.2).

Since only one of the E_8 simple roots (the spinor) is not a D_8 root, one may determine the E_8 weight set and Weyl classes by considering all the D_8 roots and one spinor. It is simplest to use the spinor

$$\chi^+ = [\,+\,+\,+\,+\;\;+\,+\,+\,+\,],$$

obtainable by combining [+ − − − − − − +] with D_8 roots. The weight set of E_8 includes all D_8 weights W such that $\langle W,\chi^+\rangle$ is an integer. This includes only two of the four D_8 congruence classes, the vectors of even $\Sigma_i f_i$ and the spinors of even $\Sigma_i f_i$. Since the spinor E_8 root connects these classes, there is only one E_8 congruence class.

It follows from Eq. (2.2) and the root set of Eq. (4.2) that the E_8 Dynkin components of a weight are given in terms of their components in the orthogonal basis by

$$a_1 = \tfrac{1}{2}(f_1 - f_2 - f_3 - f_4 - f_5 - f_6 - f_7 + f_8), \tag{4.3a}$$

$$a_2 = f_7 - f_8, \tag{4.3b}$$

$$a_3 = f_6 - f_7, \quad a_4 = f_5 - f_6, \tag{4.3c}$$

$$a_5 = f_4 - f_5, \quad a_6 = f_3 - f_4, \tag{4.3d}$$

$$a_7 = f_2 - f_3, \quad a_8 = f_7 + f_8. \tag{4.3e}$$

Frequently, one needs to determine the f's from the a's. The easiest procedure to follow is to use the inverse equations for f_1, f_7, and f_8,

$$f_1 = 2a_1 + \tfrac{7}{2}a_2 + 5a_3 + 4a_4 + 3a_5 + 2a_6 + a_7 + \tfrac{5}{2}a_8, \tag{4.4a}$$

$$f_8 = \tfrac{1}{2}(a_8 - a_2), \quad f_7 = \tfrac{1}{2}(a_8 + a_2), \tag{4.4b}$$

and then use Eqs. (4.3c)–(4.3e) to determine in order f_6 through f_2 (i.e., $f_6 = f_7 + a_3$, etc.) In order to simplify the determination of dominant weights in the orthogonal basis, I list below the eight fundamental E_8 weights in this basis; $\{k\}$ denotes the fundamental weight with Dynkin components $a_i = \delta_{ik}$

$$\begin{aligned} &\{1\}\ 2, && \{5\}\ 3111, \\ &\{2\}\ \tfrac{1}{2}(7111\ 111 - 1), && \{6\}\ 211, \\ &\{3\}\ 5111\ 11, && \{7\}\ 11, \\ &\{4\}\ 41111, && \{8\}\ \tfrac{1}{2}(5111\ 1111), \end{aligned} \tag{4.5}$$

where two or more zeros at the end of a weight are omitted, and the constant $\frac{1}{2}$ is factored out of the spinor weights.

The E_8 root set includes all the D_8 roots, so each E_8 Weyl class is a sum of complete D_8 Weyl classes. The D_8 classes in an E_8 class are those connected by the reflection S_{χ^+} associated with the spinor χ^+. It is seen from Eq. (2.8) that

$S_{\chi^+}(f) = f'$, where

$$f_i' = f_i - \frac{1}{4}\sum_{j=1}^{8} f_j. \tag{4.6}$$

When one uses Eq. (4.6) to determine the D_8 classes associated with a dominant class (class of the dominant E_8 weight) he does not have to worry about the order of the components, but only has to consider different combinations of the signs of the f_i.

I will illustrate the technique by determining the D_8 classes in the E_8 Weyl class {2}, with $M^2 = 14$. From Eq. (4.5) the dominant weight is $[\frac{1}{2}(7111\ 111 - 1)]$. From Eq. (4.6), the result of applying the reflection S_{χ^+} to this weight is $[2 - 1 - 1 - 1\ \ -1 - 1 - 1 - 2]$. The dominant weight of this D_8 Weyl class is $[2211\ 111 - 1]$. One next changes two signs in the original dominant weight, obtaining $[\frac{1}{2}(7111\ \ 1 - 1 - 1 - 1)]$. Reflecting by S_{χ^+}, one obtains a weight in the D_8 class with dominant weight $[\frac{1}{2}(5333\ 111 - 1)]$. Applying Eq. (4.6) to the weight $[\frac{1}{2}(711 - 1\ \ -1 - 1 - 1 - 1)]$ leads to the D_8 Weyl class $[3111\ 11]$.

No other D_8 classes may be obtained from one S_{χ^+} re-

TABLE II. The D_8 Weyl classes in the shorter E_8 classes.

0 {0} 1
0d

2 {7} $2^4 3 \cdot 5$
11d
$\frac{1}{2}(1111\ 1111)p$

4 {1} $2^4 3^3 5$
2d
1111s
$\frac{1}{2}(3111\ 111 - 1)p$

6 {6} $2^6 3 \cdot 5 \cdot 7$
211d
1111 11p
$\frac{1}{2}(3311\ 1111)p$

8a {8} $2^7 3^3 5$
21111p
1111 111 − 1p
$\frac{1}{2}(5111\ 1111)d$
$\frac{1}{2}(3331\ 111 - 1)p$

8b {7^2} $2^4 3 \cdot 5$
22d
1111 1111p

10 {17} $2^5 3^3 5 \cdot 7$
31d
2211s
2111 1110p
$\frac{1}{2}(5311\ 111 - 1)p$
$\frac{1}{2}(3333\ 1111)s$

12 {5} $2^6 3^3 5 \cdot 7$
3111d
222s
2211 11p
$\frac{1}{2}(5331\ 1111)p$
$\frac{1}{2}(3333\ 311 - 1)p$

14a {2} $2^9 3^3 5$
3111 11p
22211s
2211 111 − 1p
$\frac{1}{2}(7111\ 111 - 1)d$
$\frac{1}{2}(5333\ 111 - 1)p$

14b {67} $2^7 3 \cdot 5 \cdot 7$
321d
2211 1111p
$\frac{1}{2}(5511\ 1111)p$
$\frac{1}{2}(3333\ 3311)p$

16a {1^2}$2^4 3^3 5$
4d
3111 111 − 1p
2222s

16b {78} $2^{10} 3^3 5$
32111p
3111 1111p
2221 1110p
$\frac{1}{2}(7311\ 1111)d$
$\frac{1}{2}(5531\ 111 - 1)p$
$\frac{1}{2}(5333\ 3111)p$
$\frac{1}{2}(3333\ 333 - 1)p$

18a {16} $2^6 3^4 5 \cdot 7$
411d
3221s
3211 1110p
2222 11s
$\frac{1}{2}(7331\ 111 - 1)p$
$\frac{1}{2}(5533\ 1111)s$
$\frac{1}{2}(5333\ 333 - 1)p$

18b {7^3} $2^4 3 \cdot 5$
33d
$\frac{1}{2}(3333\ 3333)p$

20a {17^2} $2^5 3^3 5 \cdot 7$
42d
3311s
2222 1111s
$\frac{1}{2}(7511\ 111 - 1)p$
$\frac{1}{2}(5333\ 3331)p$

20b {4} $2^8 3^3 5 \cdot 7$
41111d
3221 11p
22222s
2222 111 − 1p
$\frac{1}{2}(7333\ 1111)p$
$\frac{1}{2}(5551\ 1111)s$
$\frac{1}{2}(5533\ 311 - 1)p$

22a {18} $2^{10} 3^3 5$
4111 1110p
32221s
3221 111 − 1p
$\frac{1}{2}(9111\ 1111)d$
$\frac{1}{2}(7333\ 311 - 1)p$
$\frac{1}{2}(5553\ 111 - 1)s$
$\frac{1}{2}(5333\ 333 - 3)p$

22b {57} $2^6 3^4 5 \cdot 7$
4211d
332s
3311 11p
3221 1111p
2222 2110p
$\frac{1}{2}(7531\ 1111)p$
$\frac{1}{2}(5533\ 3311)p$

24a {27} $2^9 3^3 5 \cdot 7$
4211 11p
33211s
3311 111 − 1p
3222 1110p
$\frac{1}{2}(9311\ 111 - 1)d$
$\frac{1}{2}(7533\ 111 - 1)p$
$\frac{1}{2}(7333\ 3311)p$
$\frac{1}{2}(5553\ 3111)s$
$\frac{1}{2}(5533\ 333 - 1)p$

24b {6^2} $2^6 3 \cdot 5 \cdot 7$
422d
3311 1111p
2222 22p

26a {$1^2 7$}$2^5 3^3 5 \cdot 7$
51d
4211 111 − 1p
3322s
$\frac{1}{2}(7333\ 333 - 1)p$
$\frac{1}{2}(5555\ 1111)s$

26b {68} $2^9 3^3 5 \cdot 7$
42211p
4211 1111p
3321 1110p
3222 21p
2222 221 − 1p
$\frac{1}{2}(9331\ 1111)d$
$\frac{1}{2}(7551\ 111 - 1)p$
$\frac{1}{2}(7533\ 3111)p$
$\frac{1}{2}(5553\ 331 - 1)p$

26c {67^2} $2^7 3 \cdot 5 \cdot 7$
431d
2222 2211p
$\frac{1}{2}(7711\ 1111)p$
$\frac{1}{2}(5533\ 3333)p$

28a {15} $2^7 3^3 5^2 7$
5111d
4222s
4221 1110p
3331s
3322 11s
3222 211 − 1p
$\frac{1}{2}(9333\ 111 - 1)p$
$\frac{1}{2}(7553\ 1111)s$
$\frac{1}{2}(7533\ 331 - 1)p$
$\frac{1}{2}(5555\ 311 - 1)s$

28b {$7^2 8$} $2^{10} 3^3 5$
43111p
3222 2111p
2222 2220p
$\frac{1}{2}(9511\ 1111)d$
$\frac{1}{2}(7731\ 111 - 1)p$
$\frac{1}{2}(7333\ 3333)p$
$\frac{1}{2}(5553\ 3331)p$

30a {167} $2^7 3^4 5 \cdot 7$
521d
4321s
4311 1110p
3322 1111s
3222 2210p
$\frac{1}{2}(9531\ 111 - 1)p$
$\frac{1}{2}(7733\ 1111)s$
$\frac{1}{2}(7533\ 3331)p$
$\frac{1}{2}(5555\ 3311)s$

30b {3} $2^9 3^3 5 \cdot 7$
5111 11d
4222 11p
3331 11s
33222s
3322 111 − 1p
$\frac{1}{2}(9333\ 3111)p$
$\frac{1}{2}(7553\ 311 - 1)p$
$\frac{1}{2}(5553\ 333 - 3)p$

32a {12} $2^{10} 3^3 5$
5111 111 − 1p
4222 111 − 1p
33321s
$\frac{1}{2}([11]\ 111\ 111 - 1)d$
$\frac{1}{2}(9333\ 331 - 1)p$
$\frac{1}{2}(7555\ 111 - 1)s$
$\frac{1}{2}(7533\ 333 - 3)p$

32b {47} $2^{10} 3^3 5 \cdot 7$
52111d
4321 11p
4222 1111p
3331 1111s
3322 2110p
$\frac{1}{2}(9533\ 1111)p$
$\frac{1}{2}(7751\ 1111)s$
$\frac{1}{2}(7733\ 311 - 1)p$
$\frac{1}{2}(7553\ 3311)p$
$\frac{1}{2}(5555\ 5111)s$
$\frac{1}{2}(5555\ 333 - 1)p$

32c {8^2} $2^7 3^3 5$
5111 1111d
42222p
3331 111 − 1p
2222 222 − 2p

32d {7^4} $2^4 3 \cdot 5$
44d
2222 2222p

flection of a member of the $\frac{1}{2}(7111\ 111-1)$ class. However, if S_{χ^+} is applied to all members of the new D_8 classes, one other D_8 class is found, with dominant weight [22211]. (For example, the weight [1100 0 −2 −2 −2] results from applying S_{χ^+} to [2211 1 −1 −1 −1].) There are five D_8 classes in the E_8 class {2}.

All the D_8 classes in the E_8 classes with $M^2 \leqslant 32$ are listed in Table II. The boldface symbols and curly bracket Dynkin labels (e.g., **14a** {2}) denote the E_8 classes with the same notation used in Table I. The figures after the curly brackets (i.e., $2^9 3^3 5$) are the numbers of the weights in the E_8 classes, expressed in prime factors. The contained D_8 classes are listed below each E_8 class, with the vector classes preceding the spinor classes. The letter d denotes the D_8 class containing the dominant E_8 weight, while p and s (primary and secondary) label classes that require one and two S_{χ^+} reflections, respectively, from a member of the dominant class. The fact that two S_{χ^+} reflections are sufficient to obtain all contained classes is proved in Sec. V.

We now return to the basic problem of finding the Weyl class of an arbitrary E_8 weight. We consider first a weight of length no greater than $(32)^{1/2}$, the weight $\mathscr{W}$ with Dynkin components,

$$\mathscr{W} = (1\ 0\ -1\ 2\ -3\ 3\ -4\ 1). \tag{4.7}$$

One finds from Eqs. (4.4a) and (4.4b) and (4.3c)–(4.3e) that the components in the orthogonal basis are $[\frac{1}{2}(1\ -5\ 3\ -3\ 3\ -1\ 1\ 1)]$. The length squared is 14 and the dominant weight of the D_8 class is $[\frac{1}{2}(5333\ 111-1)]$. From Table II this is the E_8 class **14a** {2}. [In this case if one used a Dynkin reflection series (Sec. IV A), 47 reflections would be required.]

D. The long-weight procedure

If the weight length exceeds $(32)^{1/2}$, one cannot use Table II, so an extension of the method is needed in order to find the E_8 Weyl class of the weight. One begins as before, expressing the weight in the orthogonal basis and considering the corresponding D-dominant weight (dominant weight of the D_8 Weyl class). Since all E_8 simple roots except R_1 are D_8 roots, the E_8 Dynkin components a_i of the D-dominant weight are non-negative for $i \geqslant 2$. One computes a_1 from Eq. (4.3a). If a_1 is non-negative the weight is E dominant (a dominant E_8 weight). If a_1 is negative one performs the Weyl reflection S_1 associated with R_1. The easiest way to make this reflection is to change the signs of the interior components (components 2 through 7), apply Eq. (4.6), and change the signs of the interior components again. Since one is interested only in the D_8 Weyl class of the reflected weight, the final sign-change operation may be omitted. One considers the dominant weight of the new D_8 Weyl class. If the new a_1 is negative one makes a second S_1 reflection. It is shown in Sec. V that two is the maximum number of S_1 reflections needed to produce an E-dominant weight. One may use Eqs. (4.3a)–(4.3e) to transform the e-dominant weight to the Dynkin basis.

We consider the example of a weight in the orthogonal basis that is a member of the D_8 Weyl class [7654 3210]. The length squared is 140. From Eq. (4.3a), the value of a_1 is -7, so an S_1 reflection is needed. One changes the signs of the interior components, yielding [7 −6 −5 −4 −3 −2 −1 0]. Application of Eq. (4.6) yields the weight $[\frac{1}{2}(21\ -5\ -3\ -1\ 1\ 3\ 5\ 7)]$. The dominant weight of this D_8 class is $[\frac{1}{2}(21\ 7\ 5\ 5\ 3\ 3\ 1\ -1)]$. For this weight $a_1 = -1$, so a second S_1 reflection is needed. The second reflection leads to the D_8 class with dominant weight [11 3 2 2 1 1 0 0], which must be E dominant. From Eqs. (4.3a)–(4.3e), the Dynkin components are (1010 1010).

We consider one more example, a weight M of the D_8 class [12 11 10 9 8 7 6 −5]. The value of M^2 is 620. The value of a_1 is −22, so an S_1 reflection is required. Changing the signs of the interior components and applying Eq. (4.6) yields a weight of the D_8 class,

$$[23\ 6\ 5\ 4\ 3\ 2\ 1\ 0]. \tag{4.8}$$

For this weight, $a_1 = 1$, so the weight is E dominant. The Dynkin indices are (1111 1111). This is the shortest class with a dimension equal to that of the full Weyl group (696, 729, 600) (Ref. 14). The dimensions of some classes are discussed in Sec. V.

V. CONVERGENCE RATES AND ASYMPTOTIC LIMITS

The long-weight procedure of Sec. IV D is related to a Dynkin reflection series (Sec. IV A) with all reflections except S_1 made automatically. However, in such a series the total number of reflections is as large as 120 in some cases. Therefore one might expect that the maximum number of S_1 reflections in the long-weight procedure would be on the order of 15 (120/8), rather than two. There are three reasons that the procedure converges so fast. These are discussed below.

(1) *R_1 is less active than most roots*: Since R_1 is the end root of the second shortest branch of the E_8 diagram of Fig. 2, this root is the second least active simple root of E_8. This is evidenced by the fact that the first root-basis components of most dominant weights [such as those of Eqs. (3.7a) and (3.7b)] are the second smallest components.

(2) *The S_1 reflections in the long-weight procedure are made with "bottom priority"*: In order to understand this point let us construct a Dynkin reflection series for an SU(3) weight with Dynkin components $(-a, -b)$, where a and b are positive. The two simple roots R_1 and R_2 have Dynkin components (2, −1) and (−1,2), as seen from Eq. (2.3a) and the Cartan matrix for SU(3). One can use either S_1 or S_2 for the first reflection. If S_1 is chosen, the Dynkin components of the reflected weight are $(a, -b-a)$. The next two reflections must then be S_2 and S_1, respectively, yielding successively, $(-b, b+a)$ and (b,a), which is dominant. Two S_1 and one S_2 reflections were used. Clearly, if one had applied S_2 first, he would have arrived at the same place in the same number of steps, but with only one S_1 reflection. In order to generalize this phenomenon I define a bottom priority, Dynkin reflection series as one in which a particular reflection S_k is made only if all Dynkin indices other than a_k are non-negative. One may minimize the number of S_k reflections by using a bottom-priority series. The long-weight procedure is related to a bottom-priority series, since the S_1 reflection is made only when all Dynkin components other than a_1 are non-negative.

(3) *The eighth D_8 simple root is used implicity:* In the bottom-priority series discussed above, one optimizes (reflects until the Dynkin indices are non-negative) with respect to the seven D_8 roots of the E_8 simple root set before and between S_1 reflections. This optimization does not affect the first component in the orthogonal basis, but leads to the optimal signs and ordering of the last seven components. However, in the procedure of Sec. IV D, one orders all eight orthogonal components. This is equivalent to optimizing with respect to all eight D_8 simple roots, the seven of the E_8 set and also 1_+2_-. This forces one closer to the top in positivity. The long-weight procedure is equivalent to starting fairly close to the top, and then using a bottom-priority Dynkin reflection series, with all reflections except S_1 made automatically.

I illustrate the above points by considering the weight $\mathscr{W}$ of Eq. (4.7). If one constructs a Dynkin reflection series choosing always for S_i the smallest i such that a_i is negative (top priority for S_1) 5 of the 47 reflections needed to obtain the dominant E_8 weight are S_1 reflections. If one uses S_1 with bottom priority, 47 reflections are still required, but only two are S_1 reflections. The weight preceding the first S_1 reflection is $[\frac{1}{2}(1533\ 311\ -1)]$. If one uses the procedure of Sec. IV D, the weight preceding the first S_1 reflection is $[\frac{1}{2}(5333\ 111\ -1)]$, and only one S_1 reflection is needed.

Next, I give the proof that the long-root procedure never requires more than two S_1 reflections. It is convenient to define indices g_i by the equations,

$$g_1 = f_1, \quad g_8 = f_8, \quad g_i = -f_i \quad (2 \leqslant i \leqslant 7). \tag{5.1}$$

As seen from Eq. (4.3a) the value of a_i is given simply as

$$a_1 = \frac{1}{2}\sum_{i=1}^{8} g_i. \tag{5.2}$$

If $S_1(g) = g'$, the transformation equations are

$$g_i' = g_i - \tfrac{1}{2}a_1. \tag{5.3}$$

If a weight is D dominant, the conditions of Eqs. (4.1a) and (4.1b) may be written

$$|g_i| \geqslant |g_{i+1}|, \tag{5.4a}$$

$$g_1 \geqslant 0, \quad g_i \leqslant 0, \quad \text{for } 2 \leqslant i \leqslant 7. \tag{5.4b}$$

If $g_8 \neq 0$, the sign of g_8 is the signature of the class.

We may start the iteration procedure with a D-dominant weight. In the following argument, F_i and G_i denote the f_i and g_i values of the original D-dominant weight, while G_i' and G_i'' denote the g_i values immediately after the first and second S_1 reflections, respectively.

If the original a_1, determined from Eq. (5.2) with $g_i = G_i$, is non-negative, no S_1 reflections are required, so we assume $a_1 < 0$. One performs an S_1 reflection. If the new weight (G') is not D dominant, one may make it D dominant by rearranging the coefficients 2 through 8 and changing pairs of signs. (At all stages of the procedure $g_1 \geqslant |g_i|$, for $i > 1$.) These changes correspond to the Weyl reflections S_2 through S_8, associated with the D_8 roots. Since the value of the new a_1 does not depend on the order of the g_i, I will not reorder the g_i's, but consider only the sign changes. A weight g (with $g_1 \geqslant |g_i|$) may be considered D dominant if either all g_i are nonpositive (for $i > 1$), or exactly one of these g_i is positive, and this g_i has a magnitude no greater than that of any other g_i.

We return to the S_1-reflected weight, G'. Because of Eqs. (5.4a), (5.4b), and (5.3),

$$G_i' \leqslant G_{i+1} \quad \text{for } (i > 1). \tag{5.5}$$

One may make G' D dominant by one of the four following sign-change procedures: (a) *no sign changes*, G' is already D dominant; (b) *two sign changes*, one changes the signs of G_7' and G_8'; (c) *four sign changes*, one changes the signs of G_5', G_6', G_7', and G_8'; or (d) *six sign changes*, one changes the signs of all G_i' except G_1' and G_2'. For example, suppose that G_2', G_3', G_4', and G_5' are negative, while G_6', G_7', and G_8' are positive. One then changes the signs of G_7' and G_8' and, if $G_6' > |G_5'|$, one also changes the signs of G_5' and G_6'.

We consider the four cases separately, and denote by a_1' the value after the appropriate sign changes are made. In case (a) it is easy to see that $a_1' = -a_1$. This is positive, so G' is E dominant. The one S_i reflection is sufficient. In case (b) a short calculation yields

$$a_1' = F_7 - F_8.$$

This is non-negative, so the D-dominant weight is also E dominant. The one S_1 reflection is sufficient. In case (d), after the six sign changes are made, a calculation shows that

$$a_1' = F_1 - F_2.$$

This is non-negative, so the one S_1 reflection is sufficient.

In case (c), after the four sign changes are made, a_1' may be negative. If this occurs one performs a second S_1 reflection, applying the rule of Eq. (5.3) after the sign changes are made. Because the last four G_i' have been reflected, the new g_2–g_7 indices form an asymmetric pyramid pattern, i.e.,

$$G_2'' \leqslant G_3'' \leqslant G_4'', \quad G_8'' \leqslant G_7'' \leqslant G_6'' \leqslant G_5''. \tag{5.6}$$

Again I do not reorder the columns. A calculation shows that the new g_3 and g_6 indices are

$$G_3'' = \tfrac{1}{2}[(F_2 - F_1) + (F_4 - F_3)],$$

$$G_6'' = \tfrac{1}{2}[(F_6 - F_5) + (F_8 - F_7)].$$

These are both nonpositive. Because of this and Eq. (5.6), the only components G_i'' (besides G_1'') that might be positive are G_4'' and G_5''. It follows that if G'' is not D dominant, it may be made D dominant with only one pair of sign changes. However, it was shown in case (b) above that when one pair of sign changes leads to D dominance, the resulting value of a_1 is non-negative. Therefore in case (c) the second S_1 reflection is sufficient. This concludes the proof.

The long-weight procedure is related closely to the procedure used in Sec. IV C to compile Table II. It is easy to show that weights requiring 0,1, and 2 S_1 reflections in the long-weight procedure are those labeled dominant, primary, and secondary, respectively, in Table II.

In order to illuminate some of the procedures discussed here, I will give a short discussion of the rates at which different quantities approach their asymptotic limits as one considers Weyl classes with longer and longer weights. A few definitions are useful. If $\mathscr{C}$ is a Weyl class of an algebra $\mathscr{A}$, $\mathscr{A}_0$ is defined as the algebra obtained by writing the Dynkin diagram for $\mathscr{A}$ and deleting each circle (with its connecting

lines) that corresponds to a positive Dynkin component of $\mathscr{C}$. Thus for the class {38} of E_8, $\mathscr{A}_0$ is $A_2 \times A_4$. It is well known that the number of weights in the Weyl class $\mathscr{C}$ is given by[14]

$$N(\mathscr{C}) = N(\mathscr{A})/N(\mathscr{A}_0), \tag{5.7}$$

where $N(\mathscr{A})$ is the number of reflections in the full Weyl group for the algebra $\mathscr{A}$. [If the diagram for $\mathscr{A}_0$ contains no circles, $N(\mathscr{A}_0) = 1$.] For convenience I list below the sizes of the Weyl groups for the A, D, and E algebras:

$$N(A_n) = (n+1)!,$$
$$N(D_n) = n!2^{n-1},$$
$$N(E_6) = 2^7 \cdot 3^4 \cdot 5, \quad N(E_7) = 2^{10} \cdot 3^4 \cdot 5 \cdot 7,$$
$$N(E_8) = 2^{14} \cdot 3^5 \cdot 5^2 \cdot 7.$$

The size of an E_8 Weyl class approaches its limit slowly. The sizes of all the classes in Table II are many times smaller than $N(E_8)$. The shortest class such that $N(\mathscr{C}) = N(E_8)$ is the $M^2 = 620$ class with dominant weight shown in Eq. (4.8).

Similarly, the number of D_8 classes in an E_8 class approaches its limit slowly. When all the E_8 Dynkin components are positive, this number is $N(E_8)/N(D_8) = 135$. It is seen that all the classes of Table II are far from this limit.

We next consider the Dynkin reflection series of Sec. IV A. Let $N_R(\mathscr{C})$ be the number of reflections necessary to proceed from the most negative weight of class $\mathscr{C}$ to the dominant weight. This number is given by

$$N_R(\mathscr{C}) = P(\mathscr{A}) - P(\mathscr{A}_0), \tag{5.8}$$

where $P(\mathscr{A})$ is the number of positive roots in the algebra $\mathscr{A}$, and $\mathscr{A}_0$ is defined as before. It is seen from Eqs. (5.7) and (5.8) that $N_R(\mathscr{C})$ is equal to its maximum value (120) only when $N(\mathscr{C})$ is equal to its maximum value. However, the nature of Eq. (5.8) is such that even for short classes N_R is on the order of 120. Consider, for example, the $M^2 = 6$ class {6}; $N_R\{6\} = 120 - P(E_6 \times A_1) = 83$.

Finally, we consider the asymptotic behavior of the classification procedure of Table II and related long-root procedure. I denote by F_d, F_p, and F_s the fractions of the weights in a class that are dominant, primary, and secondary, in the sense of Table II. An analysis shows that if $a_1 > 0$, $F_d = \frac{1}{135}$, $F_p = \frac{64}{135}$, $F_s = \frac{70}{135}$; if $a_1 = 0$, $F_d > \frac{2}{135}$, $F_s < \frac{70}{135}$. Thus the average number of S_1 reflections necessary in the long-root procedure $(F_p + 2F_s)$ is equal to its asymptotic value of 1.51 whenever $a_1 > 0$. This occurs for many short classes, such as the class with $M^2 = 4$. This argument shows why the long-root procedure does not get more difficult or time consuming as the weight lengths become large.

ACKNOWLEDGMENT

This paper was supported in part by the U.S. Department of Energy.

[1]E. B. Dynkin, Am. Math. Soc. Transl., Ser. 1, **9**, 328 (1962); Ser. 2, **6**, 111 (1957).
[2]R. Slansky, Phys. Rep. **79**, 1 (1981), Sec. 5.
[3]R. N. Cahn, *Semi-Simple Lie Algebras and Their Representations* (Benjamin/Cummings, Reading, MA, 1984), Chaps. X–XII.
[4]See, for example, D. Gross, J. Harvey, E. Martinec, and R. Rohm, Nucl. Phys. B **256**, 253 (1985).
[5]James E. Humphreys, *Introduction to Lie Algebras and Representation Theory* (Springer, New York, 1972).
[6]M. R. Bremner, R. V. Moody, and J. Patera, *Tables of Dominant Weight Multiplicities for Representations of Simple Lie Algebras* (Dekker, New York, 1985).
[7]Howard Georgi, *Lie Algebras in Particle Physics* (Benjamin/Cummings, Reading, MA, 1982), Eq. (VI. 19).
[8]The formulas of Eqs. (2.3a)–(2.7) may be obtained from Ref. 2, pp. 27 and 28.
[9]This theorem may be obtained by combining Secs. 13.4 and 21.3 of Ref. 5 with the well-known fact that all weights in a Weyl class may be obtained by subtracting simple roots from the dominant weight.
[10]R. H. Capps, J. Math. Phys. **27**, 914 (1986). The results concerning congruence are given on p. 920.
[11]See Ref. 2, Table 7.
[12]R. Gilmore, *Lie Groups, Lie Algebras, and Some of their Applications* (Wiley, New York, 1974), see Chap. 8.
[13]B. L. van der Waerden, Math. Z. **37**, 446 (1933). See also Ref. 12.
[14]The formula for the size of Weyl class is given in Ref. 6, p. 5.

Generalized Euler angles as intrinsic coordinates for nonlinear spinors

P. Furlan
International School for Advanced Studies (ISAS), Dipartimento di Fisica Teorica, Università di Trieste, Trieste, Italy and Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Trieste, Trieste, Italy

(Received 24 June 1986; accepted for publication 31 December 1986)

Among the nonlinear spinor representations of the pseudo-orthogonal rotation groups presented in a preceding work, those ones are considered, whose carrier spaces are isomorphic with group spaces. The general element of such groups is written as a product of one-parameter subgroups, allowing, for the $SO(\nu,\nu)$ and $SO(\nu+1,\nu)$ cases, to write explicitly the nonlinear spinor components in terms of independent intrinsic coordinates satisfying the Cartan's quadratic constraints.

I. INTRODUCTION

In a recent work,[1] done in collaboration with Rączka, we have introduced a whole class of nonlinear spinor representations of complex and pseudo-orthogonal groups, showing also that they give rise to a natural generalization of the concept of pure spinors introduced by Cartan.[2] The distinguishing characteristic of pure spinors is that their components are not all independent but satisfy some quadratic constraints. Exploiting this fact we have been able, in a subsequent work,[3] to write a set of nonlinear spinor wave equations for the $SO(\nu,\nu)$ and $SO(\nu+1,\nu)$ groups, which simplify in a drastic way when expressed in terms of the intrinsic spinor coordinates (which resemble a kind of "generalized Euler angles" for the considered groups). The key point in getting this result has been the possibility of *explicitly* writing the generic element of the complementary set C, by which the pure spinor space can be parametrized (see Refs. 1 and 3 for more details), as a product of one-parameter subgroups.

In this paper we want to show how this result has been achieved, allowing us to write explicitly all the pure spinor components in terms of the independent intrinsic ones in such a way that the quadratic Cartan relations are recovered.[2] We have limited our analysis to the pseudo-orthogonal groups $SO(\nu,\nu)$, $SO(\nu+1,\nu)$, $SO(\nu+1,\nu-1)$, and $SO(\nu+2,\nu-1)$. In fact, from Ref. 1 it follows that these are the only (real) cases in which the carrier space of the nonlinear spinor representations may be represented—up to a set of Haar measure zero—not simply as a homogeneous space but as a group space. This carrier space is obtained by the action of the considered pseudo-orthogonal group G on the standard spinor ψ_m (chosen to be the highest-weight eigenspinor). If we denote by g the Lie algebra of the group G, by h the Lie algebra of the stability subgroup H of ψ_m, and by c the set of generators complementary to h in g, we see that, when G coincides with one of the above-mentioned groups, c is also a Lie algebra and the complementary set C is a group, obtained—up to a set of Haar measure zero—by exponentiating c (see Ref. 1 for a more detailed analysis). Therefore the basic technical problem we are confronted with here is to write explicitly $\exp(c)$ as a product of the one-parameter subgroups obtained by exponentiating the operators of c.

In Sec. II we give a detailed analysis of the $SO(\nu+1,\nu)$ case, taken as a prototype, while we simply sketch the remaining cases in Sec. III. In Sec. IV for the $SO(\nu,\nu)$ and $SO(\nu+1,\nu)$ cases we write the pure spinor components explicitly in terms of the intrinsic coordinates, leaving apart the remaining two cases which present major technical complications but no new interesting features. Finally in Appendix A we give the main features of the Zassenhaus formula,[4] on which all our calculations are based, and in Appendix B we present some cumbersome relations giving the "generalized Euler angles" of the $SO(\nu+2,\nu-1)$ case. Notations and conventions are the ones introduced in Refs. 1 and 3.

II. THE SO(ν+1,ν) CASE

The complementary subalgebra c, which we are going to exponentiate in the form of a product of one-parameter subgroups, is a $[1+\binom{\nu+1}{2}]$-dimensional solvable Lie algebra with

$$\binom{\nu}{2} \text{ generators } \tilde{Q}^{kl}=-\tilde{Q}^{lk}, \quad k\neq l=1,\ldots,\nu, \tag{2.1a}$$

$$\nu \text{ generators } \tilde{F}^{k}, \quad k=1,\ldots,\nu, \tag{2.1b}$$

$$1 \text{ generator } \tilde{D}, \tag{2.1c}$$

satisfying the following commutation relations[1]:

$$[\tilde{Q}^{kl},\tilde{Q}^{mn}]=0, \quad k,l,m,n=1,\ldots,\nu, \tag{2.2a}$$

$$[\tilde{D},\tilde{Q}^{rs}]=0, \quad r\neq s=1,\ldots,\nu-1, \tag{2.2b}$$

$$[\tilde{D},\tilde{F}^{r}]=0, \quad r=1,\ldots,\nu-1, \tag{2.2c}$$

$$[\tilde{F}^{k},\tilde{Q}^{lm}]=0, \quad k,l,m=1,\ldots,\nu, \tag{2.2d}$$

$$[\tilde{D},\tilde{Q}^{r\nu}]=-\tilde{Q}^{r\nu}, \quad r=1,\ldots,\nu-1, \tag{2.2e}$$

$$[\tilde{D},\tilde{F}^{\nu}]=-\tilde{F}^{\nu}, \tag{2.2f}$$

$$[\tilde{F}^{k},\tilde{F}^{l}]=2\tilde{Q}^{kl}, \quad k\neq l=1,\ldots,\nu. \tag{2.2g}$$

In order to get an explicit realization of this algebra we can express the generators (2.1a)–(2.1c) in terms of the basis elements $\{H_0,H_j,H_{j'}\}$, $j=1,\ldots,\nu$, of the Clifford algebra $\mathbb{R}_{\nu+1,\nu}$, satisfying the anticommutation relations

$$\{H_r,H_s\}=2g_{rs}1, \quad r,s,=0,1,\ldots,\nu,1',\ldots,\nu', \tag{2.3a}$$

where

 0022-2488/87/061223-08$02.50

$$g=\left\|\begin{array}{ccc} 1 & & 0 \\ & 0 & \tfrac{1}{2}\mathbb{1}_\nu \\ 0 & & \\ & \tfrac{1}{2}\mathbb{1}_\nu & 0 \end{array}\right\|, \tag{2.3b}$$

i.e., we let

$$\tilde{Q}^{kl}=-\tfrac{1}{2}[H_{k'},H_{l'}], \quad k\neq l=1,...,\nu, \tag{2.4a}$$

$$\tilde{F}^{k}=-\tfrac{1}{2}[H_{k'},H_0], \quad k=1,...,\nu, \tag{2.4b}$$

$$\tilde{D}=-\tfrac{1}{2}[H_{\nu'},H_\nu]. \tag{2.4c}$$

[This basis, connected with the nondiagonal metric tensor (2.3b), is the so-called Witt or isotropic basis used by Cartan.[2]]

Following Cartan,[2] we provide an explicit realization of the so($\nu+1,\nu$) algebra, and then of its subalgebra c, by representing the Clifford algebra's generators H's by $2^\nu\times 2^\nu$ matrices as follows:

$$(H_j)_{i_1i_2\cdots i_p}{}^{i_1i_2\cdots i_pj}=\begin{cases}0, & \text{for } j\in i_1,i_2,...,i_p,\\ 1, & \text{for } j\notin i_1,i_2,...,i_p,\end{cases} \tag{2.5a}$$

$$(H_{j'})_{i_1i_2\cdots i_pj}{}^{i_1i_2\cdots i_p}=\begin{cases}0, & \text{for } j\in i_1,i_2,...,i_p,\\ 1, & \text{for } j\notin i_1,i_2,...,i_p,\end{cases} \tag{2.5b}$$

$$(H_0)_{i_1i_2\cdots i_p}{}^{i_1i_2\cdots i_p}=(-1)^p, \tag{2.5c}$$

all other matrix elements being zero. Here we have labeled the 2^ν rows and columns *à la* Cartan,[2] i.e., using the completely skew-symmetric set of indices

$$i_1i_2\cdots i_p, \quad p=0,1,...,\nu, \quad i_1,i_2,...,i_\nu=1,2,...,\nu \tag{2.6}$$

(where the index "0" means that no index appears).

Let us define the general element of the complementary subalgebra c as

$$\tilde{X}:=\alpha\tilde{D}+\sum_{k<l=1}^{\nu}c_{kl}\tilde{Q}^{kl}+\sum_{k=1}^{\nu}f_k\tilde{F}^k, \tag{2.7}$$

and let us separate the generators commuting with $\tilde{D}$ from the other ones, i.e., let us write

$$\tilde{X}=U+V, \tag{2.8}$$

with

$$U:=\alpha\tilde{D}+\sum_{r<s=1}^{\nu-1}c_{rs}\tilde{Q}^{rs}+\sum_{r=1}^{\nu-1}f_r\tilde{F}^r, \tag{2.9a}$$

$$V:=\sum_{r=1}^{\nu-1}c_{r\nu}\tilde{Q}^{r\nu}+f_\nu\tilde{F}^\nu. \tag{2.9b}$$

If we now apply the Zassenhaus formula (A1) to $\exp\tilde{X}$, we get

$$\exp\tilde{X}=\exp(U+V)=e^Ue^Ve^{C_2}e^{C_3}\cdots, \tag{2.10}$$

where $C_2,C_3,...$ are defined in terms of U and V in Appendix A. Then the following lemmas hold.

Lemma 2.1: All multiple commutators with any number of U's and more than one V are identically zero.

Proof: From Eqs. (2.2a)–(2.2g) we have

$$\underbrace{[[[\cdots[V,U],U],...],U]}_{n\text{ times}}=\alpha^nV-2n\alpha^{n-1}\sum_{r=1}^{\nu-1}f_rf_\nu\tilde{Q}^{r\nu}, \tag{2.11}$$

but from Eqs. (2.2a) and (2.2d) we have that

$$[\tilde{Q}^{r\nu},V]=0, \quad r=1,...,\nu-1, \tag{2.12}$$

and thus the lemma is proved. ▼

Lemma 2.2: The C_n's appearing in Eq. (2.10) are given explicitly by

$$C_{n+1}=\frac{1}{(n+1)!}\left[\alpha^nf_\nu\tilde{F}^\nu+\sum_{r=1}^{\nu-1}\alpha^{n-1}(\alpha c_{r\nu}-2nf_rf_\nu)\tilde{Q}^{r\nu}\right], \quad n=1,2,.... \tag{2.13}$$

Proof: From Eq. (A16), with x and y replaced by U and V, and from Eq. (2.11), Eq. (2.13) follows. ▼

Furthermore, from Refs. 1 and 2 we know that

$$(H_{\nu'})^2=0; \tag{2.14}$$

therefore the following lemma holds.

Lemma 2.3: We have that

$$V^2=0=C_rC_s=0=VC_r, \quad \forall r,s=2,3,.... \tag{2.15}$$

Proof: From the definitions (2.1a) and (2.1b) and from Eqs. (2.9b), (2.13), and (2.14), Eq. (2.15) follows. ▼

Now we are ready to write explicitly the infinite product part of Eq. (2.10), by means of the following proposition.

Proposition 2.1: The infinite product part of Eq. (2.10) is given by

$$e^Ve^{C_2}e^{C_3}\cdots=\mathbb{1}+\frac{e^\alpha-1}{\alpha}f_\nu\tilde{F}^\nu+\sum_{r=1}^{\nu-1}\left[\frac{e^\alpha-1}{\alpha}c_{r\nu}-2\frac{(\alpha-1)e^\alpha+1}{\alpha^2}f_rf_\nu\right]\tilde{Q}^{r\nu}. \tag{2.16}$$

Proof: From Eqs. (2.15) and (2.13) we have

$$e^{C_2}e^{C_3}\cdots=\mathbb{1}+\sum_{n=1}^{\infty}C_{n+1}=\mathbb{1}+\frac{e^\alpha-1-\alpha}{\alpha}f_\nu\tilde{F}^\nu+\sum_{r=1}^{\nu-1}\left[\frac{e^\alpha-1-\alpha}{\alpha}c_{r\nu}-2\frac{(\alpha-1)e^\alpha+1}{\alpha^2}f_rf_\nu\right]\tilde{Q}^{r\nu}; \tag{2.17}$$

then, taking into account Eq. (2.15) and (2.9b), Eq. (2.16) follows. ▼

For the remaining $\exp U$ factor we have the following proposition.

Proposition 2.2: The $\exp U$ factor in Eq. (2.10) is given by

$$\exp U=\left(\prod_{r=1}^{\nu-1}e^{f_r\tilde{F}_r}\right)\left(\prod_{s<t=1}^{\nu-1}e^{(c_{st}-f_sf_t)\tilde{Q}^{st}}\right)e^{\alpha\tilde{D}}. \tag{2.18}$$

Proof: Let us write

$$U=U_1+U_2 \tag{2.19}$$

with

$$U_1:=\alpha\tilde{D}+\sum_{r<s=1}^{\nu-1}c_{rs}\tilde{Q}^{rs} \tag{2.20a}$$

and

$$U_2 := \sum_{r=1}^{\nu-1} f_r \tilde{F}^r; \tag{2.20b}$$

then, since from Eqs. (2.2c) and (2.2d)

$$[U_1, U_2] = 0, \tag{2.21}$$

we have

$$e^U = e^{U_1} e^{U_2} = e^{U_2} e^{U_1}. \tag{2.22}$$

Taking into account Eqs. (2.2a) and (2.2b) we get

$$e^{U_1} = \left(\prod_{r<s=1}^{\nu-1} e^{c_{rs}\tilde{Q}^{rs}} \right) e^{\alpha \tilde{D}}, \tag{2.23a}$$

while, using Eq. (2.2g), we have

$$e^{U_2} = \left(\prod_{r=1}^{\nu-1} e^{f_r \tilde{F}^r} \right) \left(\prod_{s<t=1}^{\nu-1} e^{-f_s f_t \tilde{Q}^{st}} \right). \tag{2.23b}$$

Inserting Eqs. (2.23a) and (2.23b) into Eq. (2.22), and taking into account Eqs. (2.2b) and (2.2c), the proposition is proved. ▼

Now we are ready to give the final form of $\exp \tilde{X}$ in the following theorem.

Theorem 2.1: The general group element obtained by exponentiating the complementary subalgebra c of the Lie algebra $so(\nu+1,\nu)$ is expressed in terms of one-parameter subgroups as follows:

$$e^{\tilde{X}} = \left[\prod_{k=1}^{\nu} \exp(g_k \tilde{F}^k) \right] \left[\prod_{l<m=1}^{\nu} \exp(\tilde{d}_{lm} \tilde{Q}^{lm}) \right] e^{\alpha \tilde{D}}, \tag{2.24}$$

where

$$g_r := f_r, \quad r = 1,\ldots,\nu-1, \tag{2.25a}$$

$$g_\nu := [(1-e^{-\alpha})/\alpha] f_\nu, \tag{2.25b}$$

$$\tilde{d}_{rs} = -\tilde{d}_{sr} := c^{rs} - \epsilon(s-r) f^r f^s, \quad r \neq s = 1,\ldots,\nu-1, \tag{2.25c}$$

$$\tilde{d}_{r\nu} = -\tilde{d}_{\nu r} := [(1-e^{-\alpha})/\alpha] c_{r\nu} + 2[(1-\alpha-e^{-\alpha})/\alpha^2] f_r f_\nu, \quad r = 1,\ldots,\nu-1. \tag{2.25d}$$

Proof: Inserting Eqs. (2.16) and (2.18a) into Eq. (2.10) and using the relation

$$e^{\alpha \tilde{D}} \begin{pmatrix} \tilde{Q}^{r\nu} \\ \tilde{F}^\nu \end{pmatrix} e^{-\alpha \tilde{D}} = e^{-\alpha} \begin{pmatrix} \tilde{Q}^{r\nu} \\ \tilde{F}^\nu \end{pmatrix}, \quad r = 1,\ldots,\nu-1, \tag{2.26}$$

obtained from Eqs. (2.2e) and (2.2f), the theorem is proved. ▼

Remark: The relations (2.25a)–(2.25d) connecting the parameters $(g_k, \tilde{d}_{lm})$ to the parameters (f_k, c_{lm}) are analogs to the ones expressing the Euler angles in terms of the standard parameters of the SO(3) group elements taken as exponentials of the so(3) Lie algebra (keeping in mind the qualitative differences between the two cases).

III. THE SO(ν,ν), SO(ν+1,ν−1), AND SO(ν+2,ν−1) CASES

In this section we want just to sketch the main features of the remaining pseudo-orthogonal cases (in which the complementary set C has a group structure), leaving apart as much as possible all those technical details which can be recovered by analogy in the preceding section.

A. The SO(ν,ν) case

This case can be considered as a subcase of the $SO(\nu+1,\nu)$ case in which all $\tilde{F}^k$ generators are deleted. Let us denote the general element of the complementary subalgebra c as

$$\tilde{Y} := \alpha \tilde{D} + \sum_{k<l=1}^{\nu} c_{kl} \tilde{Q}^{kl}. \tag{3.1}$$

Then we have the following theorem.

Theorem 3.1: The general group element obtained by exponentiating the complementary subalgebra c of the Lie algebra $so(\nu,\nu)$ is expressed in terms of one-parameter subgroups as follows:

$$e^{\tilde{Y}} = \left[\prod_{k<l=1}^{\nu} \exp(\tilde{c}_{kl} \tilde{Q}^{kl}) \right] e^{\alpha \tilde{D}}, \tag{3.2}$$

where

$$\tilde{c}_{rs} = -\tilde{c}_{sr} := c_{rs}, \quad r \neq s = 1,\ldots,\nu-1, \tag{3.3a}$$

$$\tilde{c}_{r\nu} = -\tilde{c}_{\nu r} := [(1-e^{-\alpha})/\alpha] c_{r\nu}, \quad r = 1,\ldots,\nu-1. \tag{3.3b}$$

Proof: Put all $f_r \equiv 0$, $r = 1,\ldots,\nu$, in Eqs. (2.24), (2.25c), and (2.25d) of Theorem 2.1. ▼

B. The SO(ν+1,ν−1) case

In this case the complementary subalgebra c is a $[1 + \binom{\nu+1}{2}]$-dimensional solvable Lie algebra with

$\binom{\nu-1}{2}$ generators $\tilde{Q}^{kl} = -\tilde{Q}^{lk} = -H_{k'} H_{l'}$, $k \neq l = 2,\ldots,\nu$, (3.4a)

$\nu-1$ generators $\tilde{C}^{1l} = -\frac{1}{2}(H_1 + H_{1'}) H_{l'}$,
$\nu-1$ generators $\tilde{D}^{1l} = \frac{i}{2}(H_1 - H_{1'}) H_{l'}$, $\Big\}$ $l = 2,\ldots,\nu$, (3.4b) (3.4c)

1 generator $\tilde{B} = -(i/2)[H_1, H_{1'}]$, (3.4d)

1 generator $\tilde{D} = \frac{1}{2}[H_\nu, H_{\nu'}]$, (3.4e)

satisfying the following commutation relations:

$$[\tilde{D}, \tilde{Q}^{r\nu}] = -\tilde{Q}^{r\nu}, \quad r = 2,\ldots,\nu-1, \tag{3.5a}$$

$$[\tilde{D}, \tilde{C}^{1\nu}] = -\tilde{C}^{1\nu}, \tag{3.5b}$$

$$[\tilde{D}, \tilde{D}^{1\nu}] = -\tilde{D}^{1\nu}, \tag{3.5c}$$

$$[\tilde{B}, \tilde{C}^{1l}] = \tilde{D}^{1l}, \tag{3.5d}$$

$$[\tilde{B}, \tilde{D}^{1l}] = -\tilde{C}^{1l}, \tag{3.5e}$$

$$[\tilde{C}^{1k}, \tilde{C}^{1l}] = \tfrac{1}{2} \tilde{Q}^{kl}, \quad k,l = 2,\ldots,\nu, \tag{3.5f}$$

$$[\tilde{D}^{1k}, \tilde{D}^{1l}] = \tfrac{1}{2} \tilde{Q}^{kl}, \tag{3.5g}$$

all other commutators being zero.

Let us define the general element of the complementary subalgebra c as

$$\tilde{W} := \alpha \tilde{D} + \beta \tilde{B} + \sum_{k<l=2}^{\nu} c_{kl} \tilde{Q}^{kl} + \sum_{m=2}^{\nu} c_m \tilde{C}^{1m} + \sum_{n=2}^{\nu} d_n \tilde{D}^{1n}. \tag{3.6}$$

Then we have the following theorem.

Theorem 3.2: The general group element, obtained by exponentiating the complementary subalgebra c of the Lie algebra $so(\nu+1,\nu-1)$, is expressed in terms of one-parameter subgroups as follows:

$$e^{\tilde{W}} = e^{\beta\tilde{B}}\left[\prod_{k=2}^{\nu}\exp(\hat{c}_k\tilde{C}^{1k})\right]\left[\prod_{l=2}^{\nu}\exp(\hat{d}_l\tilde{D}^{1l})\right] \times\left[\prod_{m<n=2}^{\nu}\exp(\hat{c}_{mn}\tilde{Q}^{mn})\right]e^{\alpha\tilde{D}}, \tag{3.7}$$

where

$$\hat{c}_r := \frac{1}{2}\left(\frac{\sin\beta}{\beta}c_r + \frac{1-\cos\beta}{\beta}d_r\right), \quad r=2,\ldots,\nu-1, \tag{3.8a}$$

$$\hat{c}_\nu := \frac{1}{\alpha^2+\beta^2}\{[\alpha\cos\beta+\beta\sin\beta-\alpha e^{-\alpha}]c_\nu + [\alpha\sin\beta-\beta\cos\beta+\beta e^{-\alpha}]d_\nu\}, \tag{3.8b}$$

$$\hat{d}_s := \frac{1}{2}\left(\frac{\sin\beta}{\beta}d_s - \frac{1-\cos\beta}{\beta}c_s\right), \quad s=2,\ldots,\nu-1, \tag{3.8c}$$

$$\hat{d}_\nu := \frac{1}{\alpha^2+\beta^2}\{[\alpha\cos\beta+\beta\sin\beta-\alpha e^{-\alpha}]d_\nu - [\alpha\sin\beta-\beta\cos\beta+\beta e^{-\alpha}]c_\nu\}, \tag{3.8d}$$

$$\hat{c}_{rs} = -\hat{c}_{sr} := c_{rs} + \frac{\beta-\sin\beta}{2\beta^2}(d_rc_s - d_sc_r) - \frac{1-\cos\beta}{2\beta^2}\epsilon(s-r)(c_rc_s + d_rd_s), \quad r\neq s=2,\ldots,\nu-1, \tag{3.8e}$$

$$\hat{c}_{r\nu} = -\hat{c}_{\nu r} := \frac{1-e^{-\alpha}}{\alpha}\left[c_{r\nu} - \frac{1}{2\beta}(c_rd_\nu - c_\nu d_r)\right] + \frac{1}{2\beta(\alpha^2+\beta^2)}\{-[\alpha\sin\beta-\beta\cos\beta+\beta e^{-\alpha}] \times(c_rc_\nu + d_rd_\nu) + [\alpha\cos\beta+\beta\sin\beta-\alpha e^{-\alpha}] \times(c_rd_\nu - c_\nu d_r)\}, \quad r=2,\ldots,\nu-1, \tag{3.8f}$$

with $\epsilon(x) = \pm 1$ for $x \gtrless 0$.

Proof: Following the same procedure used in Sec. II and by repeated use of the Zassenhaus formula (A1) the theorem is proved. ▼

C. The SO(ν+2,ν−1) case

This case differs from the preceding ones since the complementary subalgebra c is no more a solvable Lie algebra but a semidirect sum of the so(3) simple Lie algebra with a solvable Lie algebra. Its $[1+\binom{\nu+2}{2}]$ generators are those given by Eqs. (3.4a)–(3.4e) plus

$(\nu-1)$ generators $\tilde{F}^k = -H_{k'}H_0, \quad k=2,\ldots,\nu,$ (3.9a)

1 generator $\tilde{H}^1 = -\frac{1}{2}(H_1+H_{1'})H_0,$ (3.9b)

1 generator $\tilde{K}^1 = i/2(H_1-H_{1'})H_0.$ (3.9c)

They satisfy the commutation relations (3.5a)–(3.5g) plus

$$[\tilde{D},\tilde{F}^\nu] = -\tilde{F}^\nu, \tag{3.10a}$$

$$[\tilde{B},\tilde{H}^1] = \tilde{K}^1, \tag{3.10b}$$

$$[\tilde{H}^1,\tilde{K}^1] = \tilde{B}, \tag{3.10c}$$

$$[\tilde{K}^1,B] = \tilde{H}^1, \tag{3.10d}$$

$$\left.\begin{aligned}[\tilde{H}^1,\tilde{F}^k] &= 2\tilde{C}^{1k}, &(3.10e)\\ [\tilde{K}^1,\tilde{F}^k] &= 2\tilde{D}^{1k}, &(3.10f)\\ [\tilde{F}^k,\tilde{F}^l] &= 2\tilde{Q}^{kl}, &(3.10g)\\ [\tilde{H}^1,\tilde{C}^{1l}] &= -\tfrac{1}{2}\tilde{F}^l, &(3.10h)\\ [\tilde{K}^1,\tilde{D}^{1l}] &= -\tfrac{1}{2}\tilde{F}^l, &(3.10i)\end{aligned}\right\} \quad k,l=2,\ldots,\nu,$$

all other commutators being zero.

Let us define the general element of the complementary subalgebra c as

$$\tilde{Z} := \alpha\tilde{D}+\beta\tilde{B}+\sum_{k<l=2}^{\nu}c_{kl}\tilde{Q}^{kl}+\sum_{k=2}^{\nu}c_k\tilde{C}^{1k} + \sum_{l=2}^{\nu}d_l\tilde{D}^{1l}+\sum_{m=2}^{\nu}f_m\tilde{F}^m+hH^1+kK^1. \tag{3.11}$$

Then we have the following theorem.

Theorem 3.3: The general group element, obtained by exponentiating the complementary subalgebra c of the Lie algebra $so(\nu+2,\nu-1)$, is expressed in terms of one-parameter subgroups as follows:

$$e^{\tilde{Z}} = e^{J}\left[\prod_{k=2}^{\nu}\exp(\mathring{c}_k\tilde{C}^{1k})\right]\left[\prod_{l=2}^{\nu}\exp(\mathring{d}_l\tilde{D}^{1l})\right] \times\left[\prod_{m=2}^{\nu}\exp(\mathring{f}_m\tilde{F}^m)\right]\left[\prod_{k<l=2}^{\nu}\exp(\mathring{c}_{kl}\tilde{Q}^{kl})\right]e^{\alpha\tilde{D}}, \tag{3.12a}$$

where J denotes the general element of the so(3) simple Lie subalgebra

$$J := \beta\tilde{B}+h\tilde{H}^1+k\tilde{K}^1, \tag{3.12b}$$

and the coefficients $\mathring{c}_k$, $\mathring{d}_l$, $\mathring{f}_m$, $\mathring{c}_{kl}$ are given explicitly in terms of the coefficients of Eq. (3.11) in Appendix B.

Proof: Following the same procedure used in Sec. II and by repeated use of the Zassenhaus formula (A1) the theorem is proved. ▼

Remark: The group element e^J is easily factorized into one-parameter subgroups by means of the usual expression involving the standard Euler angles.

IV. INTRINSIC SPINOR COMPONENTS FOR PURE SPINORS

In this section we give the pure spinor components for the $SO(\nu,\nu)$ and $SO(\nu+1,\nu)$ cases in terms of the "generalized Euler angles" appearing in the complementary set C, once it is expressed as a product of one-parameter subgroups as in Eqs. (2.24) and (3.2). We prefer to start with the $SO(\nu+1,\nu)$ case, since the $SO(\nu,\nu)$ case can be easily deduced from it, taking into account Sec. III A.

A. The SO(ν+1,ν) case

Choosing the same spinor representation appearing in Ref. 1, we see that the generic $SO(\nu+1,\nu)$ pure spinor is given by

$$\psi := e^{\tilde{X}}\psi_m, \tag{4.1}$$

where $e^{\tilde{X}}$ is given by Eq. (2.24) and ψ_m is a 2^ν-component spinor with all components zero but the first one equal to unity. Following Refs. 1 and 2, we label the 2^ν components of ψ with a set of completely antisymmetric indices, i.e., $\psi_{i_1\cdots i_r}$ with $0\leqslant r\leqslant\nu$ and $1\leqslant i_1,...,i_r\leqslant\nu$. Then we have the following lemmas.

Lemma 4.1: The group element $e^{\tilde{X}}$ of Eq. (2.10) can be rewritten as

$$e^{\tilde{X}}=\left(1+\sum_{k=1}^{\nu}g_k\tilde{F}^k+\sum_{r=1}^{\nu-1}g_rg_\nu\tilde{Q}^{r\nu}\right)\times\left[\prod_{k,l=1}^{\nu}(1+\omega_{kl}\tilde{Q}^{kl})\right]e^{\alpha\tilde{D}},\tag{4.2a}$$

where

$$\omega_{rs}=-\omega_{sr}:=\tfrac{1}{2}(\tilde{d}_{rs}+g_rg_s),\quad r\neq s=1,...,\nu-1,\tag{4.2b}$$

$$\omega_{r\nu}=-\omega_{\nu r}:=\tfrac{1}{2}\tilde{d}_{r\nu},\quad r=1,...,\nu-1.\tag{4.2c}$$

Proof: From Eqs. (2.2g) and (2.20b) we have that

$$e^{U_2}=\mathbb{1}+\sum_{r=1}^{\nu-1}f_r\tilde{F}^r,\tag{4.3a}$$

and from Eqs. (2.2a), (2.2b), and (2.20a)

$$e^{U_1}=\left[\prod_{r,s=1}^{\nu-1}\left(1+\frac{1}{2}c_{rs}\tilde{Q}^{rs}\right)\right]e^{\alpha\tilde{D}}.\tag{4.3b}$$

Furthermore, from Eqs. (2.16) and (2.26)

$$e^{\alpha\tilde{D}}(e^Ve^{C_2}e^{C_3}\cdots)e^{-\alpha\tilde{D}}=(1+g_\nu\tilde{F}^\nu)\left[\prod_{r=1}^{\nu-1}\exp(\omega_{r\nu}\tilde{Q}^{r\nu})\right]\left[\prod_{r=1}^{\nu-1}\exp(\omega_{\nu r}\tilde{Q}^{\nu r})\right].\tag{4.4}$$

Inserting Eqs. (4.3a) and (4.3b) into Eq. (2.22) and then, together with Eq. (4.4), into Eq. (2.10), with the help of the relation

$$\tilde{F}^k\tilde{F}^l=\tilde{Q}^{kl},\quad k\neq l=1,...,\nu,\tag{4.5}$$

and of Eqs. (2.25a)–(2.25d), the lemma is proved. ▼

Lemma 4.2: The *even* components of ψ can be written as

$$\psi_{i_1\cdots i_{2p}}=e^{\alpha/2}\left\{\left[\prod_{k,l=1}^{\nu}(1+\eta_{kl}\tilde{Q}^{kl})\right]\psi_m\right\}_{i_1\cdots i_{2p}},\tag{4.6a}$$

where

$$\eta_{rs}=-\eta_{sr}:=\omega_{rs},\quad r\neq s=1,...,\nu-1,\tag{4.6b}$$

$$\eta_{r\nu}=-\eta_{\nu r}:=\omega_{r\nu}+\tfrac{1}{2}g_rg_\nu,\quad r=1,...,\nu-1,\tag{4.6c}$$

while the *odd* components of ψ can be written as

$$\psi_{i_1\cdots i_{2q+1}}=e^{\alpha/2}\left\{\left(\sum_{k=1}^{\nu}g_k\tilde{F}^k\right)\left[\prod_{k,l=1}^{\nu}(1+\omega_{kl}\tilde{Q}^{kl})\right]\psi_m\right\}_{i_1\cdots i_{2q+1}}.\tag{4.7}$$

Proof: From Refs. 1 and 2 we have that

$$\tilde{D}\psi_m=\tfrac{1}{2}\psi_m.\tag{4.8}$$

Furthermore, from Refs. 1 and 2 we have that

$$(H_{l'})^2=0,\quad\forall l=1,...,\nu,\tag{4.9a}$$

and then, from Eqs. (2.1a) and (2.1b),

$$\tilde{Q}^{kl}\tilde{Q}^{ml}=0=\tilde{Q}^{kl}\tilde{F}^l=0=(\tilde{F}^l)^2,\quad\forall k,l,m=1,...,\nu.\tag{4.9b}$$

Since we know from Refs. 1 and 2 that the only nonzero matrix elements of H_j have the form

$$(H_{j'})_{k_1\cdots k_r j}{}^{k_1\cdots k_r}=1,\quad j\notin(k_1,...,k_r);\quad j,k_1,...,k_r=1,...,\nu,\quad r=0,...,\nu-1,\tag{4.10}$$

we see from Eqs. (2.1a) and (2.1b) that each $\tilde{Q}^{kl}$ in Eq. (4.2a) "creates" a pair of distinct indices for ψ, while each $\tilde{F}^k$ (since H_0 is a diagonal matrix) "creates" only one index for ψ. After separation of the even and odd pieces in Eq. (4.2a), the lemma is proved. ▼

Let us introduce the notation for totally antisymmetric tensors

$$T_{[i_1\cdots i_r]}:=\frac{1}{r!}\sum_{\pi(i_1\cdots i_r)}(-1)^\pi T_{i_{\pi_1}\cdots i_{\pi_r}},\tag{4.11}$$

where the summation is extended over all permutations of the indices $i_1,...,i_r$ and $(-1)^\pi$ denotes the signature of the given permutation. Then we are ready to present the main proposition of this subsection.

Proposition 4.1: The generic non-null component of the spinor ψ of Eq. (4.1) with an even number of indices is given by

$$\psi_{i_1\cdots i_{2p}}=(2p-1)!!e^{\alpha/2}d_{[i_1i_2}d_{i_3i_4}\cdots d_{i_{2p-1}i_{2p}]},\tag{4.12a}$$

while that one with an odd number of indices is given by

$$\psi_{i_1\cdots i_{2q+1}}=-(2q+1)!!e^{\alpha/2}g_{[i_1}d_{i_2i_3}\cdots d_{i_{2q}i_{2q+1}]},\tag{4.12b}$$

where

$$d_{kl}=-d_{lk}:=\tilde{d}_{kl}+g_kg_l\epsilon(l-k),\quad k\neq l=1,...,\nu,\tag{4.12c}$$

with $\tilde{d}_{kl}$ and g_k defined by Eqs. (2.25a)–(2.5d) [and $(-1)!!:=1$].

Proof: Let us begin by proving Eq. (4.12a). Taking into account Eq. (4.9b) we can write

$$\left[\prod_{k,l=1}^{\nu}(1+\eta_{kl}\tilde{Q}^{kl})\right]\psi_m=\left\{1+2\sum_{\substack{k,l=1\\k<l}}^{\nu}\eta_{kl}\tilde{Q}^{kl}+2^2\times\sum_{\substack{k_1\neq l_1\neq k_2\neq l_2=1\\k_1<l_1,k_2<l_2\\l_1<l_2}}^{\nu}\eta_{k_1l_1}\eta_{k_2l_2}\tilde{Q}^{k_1l_1}\tilde{Q}^{k_2l_2}+\cdots+2^{[\nu/2]}\sum_{\substack{k_1\neq l_1\neq\cdots\neq k_{[\nu/2]}\neq l_{[\nu/2]}=1\\k_1<l_1,...,k_{[\nu/2]}<l_{[\nu/2]}\\l_1<l_2<\cdots<l_{[\nu/2]}}}^{\nu}\eta_{k_1l_1}\cdots\eta_{k_{[\nu/2]}l_{[\nu/2]}}\times\tilde{Q}^{k_1l_1}\cdots\tilde{Q}^{k_{[\nu/2]}l_{[\nu/2]}}\right\}\psi_m.\tag{4.13}$$

Inserting Eq. (4.13) into Eq. (4.6a) and taking into account Eqs. (2.1a) and (4.10) we see that $\psi_{i_1\cdots i_{2p}}$ gets a contribution only from those terms of Eq. (4.13) having p $\tilde{Q}$'s with the indices being a permutation of $(i_1\cdots i_{2p})$, i.e., only from the terms of the kind

$$\tilde{Q}^{k_1 l_1}\cdots\tilde{Q}^{k_p l_p} = (-1)^p \pi\begin{pmatrix} i_1 i_2\cdots i_{2p} \\ k_1 l_1\cdots k_p l_p \end{pmatrix} H_{i_1' i_1} H_{i_2' i_2}\cdots H_{i_{2p}' i_{2p}}, \tag{4.14a}$$

where $\pi\left(\begin{smallmatrix} i_1\cdots i_{2p} \\ k_1 l_1\cdots k_p l_p \end{smallmatrix}\right)$ denotes the parity of the permutation transforming the sequence $(i_1\cdots i_{2p})$ into the sequence $(k_1 l_1\cdots k_p l_p)$. Taking into account Eq. (4.10), we have then

$$(\tilde{Q}^{k_1 l_1}\cdots\tilde{Q}^{k_p l_p}\psi_m)_{i_1\cdots i_{2p}} = \pi\begin{pmatrix} i_1 i_2\cdots i_{2p} \\ k_1 l_1\cdots k_p l_p \end{pmatrix}, \tag{4.14b}$$

and, using Eqs. (4.6a), (4.13), (4.14b), and the definition (4.11), the first part of the proposition is proved, with the help of the relation

$$\sum_{\substack{(k_1,l_1,\ldots,k_p,l_p)\in(i_1,\ldots,i_{2p}) \\ k_1\neq l_1\neq\cdots\neq k_p\neq l_p \\ k_1<l_1,\ldots,k_p<l_p \\ l_1<l_2<\cdots<l_p}} \pi\begin{pmatrix} i_1 i_2\cdots i_{2p} \\ k_1 l_1\cdots k_p l_p \end{pmatrix}\eta_{k_1 l_1}\cdots\eta_{k_p l_p} = (2p-1)!!\eta_{[i_1 i_2}\cdots\eta_{i_{2p-1} i_{2p}]}, \tag{4.14c}$$

and of the definitions (4.6b), (4.6c), (4.2b), (4.2c), and (4.12c).

Let us now rewrite Eq. (4.7) as follows:

$$\psi_{i_1\cdots i_{2q+1}} = \sum_{r=0}^{\nu}\sum_{j_1,\ldots,j_r=1}^{\nu}\left(\sum_{k=1}^{\nu} g_k\tilde{F}^k\right)_{i_1\cdots i_{2q+1}}^{\quad j_1\cdots j_r}\tilde{\psi}_{j_1\cdots j_r}, \tag{4.15a}$$

where

$$\tilde{\psi}_{j_1\cdots j_r} := e^{\alpha/2}\left\{\left[\prod_{k,l=1}^{\nu}(1+\omega_{kl}\tilde{Q}^{kl})\right]\psi_m\right\}_{j_1\cdots j_r}. \tag{4.15b}$$

We see that $\tilde{\psi}_{j_1\cdots j_r}$ is identical to $\psi_{i_1\cdots i_{2p}}$ of Eq. (4.6a) (non-null contributions come for r even only), once we substitute η_{kl} with ω_{kl}. Its explicit form is given by

$$\tilde{\psi}_{j_1\cdots j_r} = 2^{r/2}(r-1)!!e^{\alpha/2}\omega_{[j_1 j_2}\cdots\omega_{j_{r-1} j_r]}. \tag{4.16}$$

Since H_0 is a diagonal matrix with

$$(H_0)_{i_1\cdots i_s}^{\quad i_1\cdots i_s} = (-1)^s, \quad i_1,\ldots,i_\nu = 1,\ldots,\nu, \quad s = 0,\ldots,\nu, \tag{4.17}$$

taking into account the definition (2.1b) and Eqs. (4.10), (4.16), and (4.17), we can write

$$\psi_{i_1\cdots i_{2q+1}} = \sum_{k=1}^{2q+1}(-1)^k g_{i_k}\tilde{\psi}_{i_1\cdots\hat{i}_k\cdots i_{2q+1}}, \tag{4.18}$$

where a $\hat{}$ sign over an index means that that particular index is missing from the sequence in which it appears. If we now express the $\tilde{\psi}$ components by means of Eq. (4.16) and we take into account definitions (4.11), (4.2b), (4.2c), and (4.12c), we see that Eq. (4.12b) is also proved. ▼

Remark: From the explicit expressions (4.12a) and (4.12b) it is easy to check that the pure spinor's components satisfy the quadratic relations presented in Ref. 2.

B. The SO(ν,ν) case

In this case we define the generic pure spinor as

$$\psi := e^{\bar{Y}}\psi_{m^+}, \tag{4.19}$$

where $e^{\bar{Y}}$ is given by Eq. (3.2) and ψ_{m^+} is identical to ψ_m of the preceding subsection. From Refs. 1 and 2 we see that ψ is effectively a semispinor of the first kind, having all components with an odd number of indices equal to zero. [Any second kind semispinor can be obtained by reflections of first kind semispinors (see Ref. 2).]

As already mentioned in Sec. III A, this case can be derived from the SO($\nu+1,\nu$) case, deleting all $\tilde{F}^k$ generators. We have then the following proposition.

Proposition 4.2: The generic non-null component of the semispinor ψ of Eq. (4.19) is given by

$$\psi_{i_1\cdots i_{2p}} = (2p-1)!!e^{\alpha/2}\tilde{c}_{[i_1 i_2}\cdots\tilde{c}_{i_{2p-1} i_{2p}]}, \tag{4.20}$$

with $\tilde{c}_{kl}$ given by Eqs. (3.3a) and (3.3b).

Proof: From Eqs. (2.3) and (3.1) we see that, if we let $f_k \equiv 0$, $k = 1,\ldots,\nu$, we have

$$e^{\bar{X}}|_{\{f_k\equiv 0\}} = e^{\bar{Y}}; \tag{4.21a}$$

then, since $\psi_{m^+} \equiv \psi_m$, from Eqs. (4.1), (4.19), and (4.21a) we have that

$$\psi_{\mathrm{SO}(\nu,\nu)} = \psi_{\mathrm{SO}(\nu+1,\nu)}|_{\{f_k\equiv 0\}}. \tag{4.21b}$$

Then the only nonzero components of $\psi_{\mathrm{SO}(\nu,\nu)}$ are given by Eq. (4.12a), with the parameters d_{kl} evaluated with $f_k \equiv 0$. Taking into account Eqs. (4.12c), (2.25a)–(2.25d) and (3.3a), (3.3b), we see that the proposition is proved. ▼

Remark: We want to stress that in Eqs. (4.12a), (4.12b), and (4.20) we have been able to write the nonlinear (pure) spinor components in terms of the independent intrinsic coordinates, just because we succeeded in writing both $e^{\bar{X}}$ and $e^{\bar{Y}}$ as a product of one-parameter subgroups. It can be easily seen also that such a "factorization" of the group elements of the complementary set C is absolutely necessary for the realization of minimal covariant nonlinear spinor wave equations associated with our nonlinear spinor representations (see Ref. 3).

APPENDIX A: THE ZASSENHAUS FORMULA

In order to make the paper as self-contained as possible, we give here a brief account on what is presented in Ref. 4 on the Zassenhaus formula, and derive some straightforward consequences to be used in our analysis in the main text.

Let R be a free ring with two generators x, y and with rational coefficients. Then we have the following theorem.

Theorem A1 (Zassenhaus): There exist uniquely determined Lie elements C_n ($n = 2,3,4,\ldots$) in R, which are exactly of degree n in x, y, such that

$$e^{x+y} = e^x e^y e^{C_2} e^{C_3}\cdots e^{C_n}\cdots. \tag{A1}$$

▼

Here a rough definition of C_n is that it is any linear combination of multiple commutators of x and y, in which the total power of x and y is n (see Ref. 4 for a more precise definition).

In order to get the C_n's explicitly we have to introduce the "curly bracket operator" $\{\ \}$ having the following properties.

(i) It is a linear operator, i.e., for any two elements $F_1, F_2 \in R$ and for any two elements c_1, c_2 in the field of coefficients

$$\{c_1 F_1 + c_2 F_2\} = c_1\{F_1\} + c_2\{F_2\}. \tag{A2}$$

(ii) Let x_ν for $\nu = 1,2,...,n$ be any one of the generators. Then for any monomial $x_1x_2\cdots x_n$ we define

$$\{x_1x_2\cdots x_n\} := [[\cdots[[x_1,x_2],x_3],...],x_n], \tag{A3}$$

$$\{x_\nu\} := x_\nu, \tag{A4}$$

$$\{1\} := 0. \tag{A5}$$

Furthermore, the two following propositions are needed.

Proposition A1: Let G be a homogeneous Lie element in R which is of degree n; then

$$\{G\} = nG. \tag{A6}$$

▼

Proposition A2: If G is a homogeneous Lie element and F is any element of R then

$$\{G^2F\} = 0. \tag{A7}$$

▼

Then, applying the operator $\{\ \}$ to both sides of Eq. (A1) we get

$$\{e^{x+y}\} = x + y \tag{A8}$$

and

$$\{e^xe^ye^{C_2}e^{C_3}...\} = x + y + \{xy\} + \{C_2\} + \{xy^2\}/2! + \{xC_2\} + \{yC_2\} + \{C_3\} + \cdots. \tag{A9}$$

By comparing terms of the same degree in Eqs. (A8) and (A9), we get

$$\{C_2\} + \{xy\} = 0, \tag{A10}$$

$$\{C_3\} + \{xC_2\} + \{yC_2\} + \tfrac{1}{2}\{xy^2\} = 0, \tag{A11}$$

and so on, giving

$$C_2 = -\tfrac{1}{2}[x,y], \tag{A12}$$

$$C_3 = -\tfrac{1}{3}[[x,y],y] - \tfrac{1}{6}[[x,y],x], \tag{A13}$$

and so on.

If we now look at the relations (A10), (A11), and to the analogous ones of higher degree, we see that

$$\{C_n\} = -\{xC_{n-1}\} + (\text{terms at least of degree 2 in } y), \tag{A14}$$

i.e., from Eq. (A3) and (A6),

$$nC_n = [C_{n-1},x] + (\text{terms at least of degree 2 in } y). \tag{A15}$$

Iterating Eq. (A15) and using Eq. (A12), we finally get

$$C_n = (1/n!)\underbrace{[[[\cdots[y,x],x],...],x]}_{n-1} + (\text{multiple commutators with } y \text{ appearing at least twice}). \tag{A16}$$

APPENDIX B: EXPLICIT FORM OF THE COEFFICIENTS IN THE SO(ν+2,ν−1) CASE

The relations connecting the coefficients $\mathring{c}_k, \mathring{d}_l, \mathring{f}_m, \mathring{c}_{kl}$ appearing in Eq. (3.12a) to the coefficients $\alpha, \beta, c_{kl}, c_k, d_l$, of Eq. (3.11) are rather cumbersome; therefore we prefer to write them explicitly in this appendix as follows:

$$\mathring{c}_r := \left(2h\frac{\cos\rho-1}{\rho^2} - \beta k\frac{\sin\rho-\rho}{\rho^3}\right)f_r + \left[1 + \frac{\sin\rho-\rho}{\rho^3}(h^2+\beta^2)\right]c_r + \left(hk\frac{\sin\rho-\rho}{\rho^3} - \beta\frac{\cos\rho-1-\frac{1}{2}\rho^2}{\rho^2}\right)d_r, \quad r = 2,...,\nu-1, \tag{B1a}$$

$$\mathring{c}_\nu := \frac{1-e^{-\alpha}}{\alpha}c_\nu + \frac{1}{\alpha^2+\rho^2}\left(\alpha\frac{\sin\rho}{\rho} - \cos\rho + e^{-\alpha}\right)(\beta d_\nu - 2hf_\nu) - \left[\frac{1}{\alpha\rho^2} - \frac{1}{\alpha^2+\rho^2}\left(\alpha\frac{\cos\rho}{\rho^2} + \frac{\sin\rho}{\rho} + \frac{e^{-\alpha}}{\alpha}\right)\right][(\beta^2+h^2)c_\nu + hkd_\nu + 2k\beta f_\nu], \tag{B1b}$$

$$\mathring{d}_r := \left(2k\frac{\cos\rho-1}{\rho^2} + \beta h\frac{\sin\rho-\rho}{\rho^3}\right)f_r + \left(hk\frac{\sin\rho-\rho}{\rho^3} + \beta\frac{\cos\rho-1-\frac{1}{2}\rho^2}{\rho^2}\right)c_r + \left[1 + (\beta^2+k^2)\frac{\sin\rho-\rho}{\rho^3}\right]d_r, \quad r = 2,...,\nu-1, \tag{B1c}$$

$$\mathring{d}_\nu := \frac{1-e^{-\alpha}}{\alpha}d_\nu - \frac{1}{\alpha^2+\rho^2}\left(\alpha\frac{\sin\rho}{\rho} - \cos\rho + e^{-\alpha}\right)(\beta c_\nu + 2kf_\nu) - \left[\frac{1}{\alpha\rho^2} - \frac{1}{\alpha^2+\rho^2}\left(\alpha\frac{\cos\rho}{\rho^2} + \frac{\sin\rho}{\rho} + \frac{e^{-\alpha}}{\alpha}\right)\right][(\beta^2+k^2)d_\nu + hkc_\nu - 2h\beta f_\nu], \tag{B1d}$$

$$\mathring{f}_r := \left[1 - \frac{1}{2}(h^2+k^2)\frac{\sin\rho-\rho}{\rho^3}\right]f_r + \tfrac{1}{2}\left(\beta k\frac{\sin\rho-\rho}{\rho^3} - h\frac{\cos\rho-1-\frac{1}{2}\rho^2}{\rho^2}\right)c_r - \frac{1}{2}\left(\beta h\frac{\sin\rho-\rho}{\rho^3} + k\frac{\cos\rho-1-\frac{1}{2}\rho^2}{\rho^3}\right)d_r, \quad r = 2,...,\nu-1, \tag{B1e}$$

$$\mathring{f}_\nu := \frac{1-e^{-\alpha}}{\alpha}f_\nu + \frac{1}{2}\frac{1}{\alpha^2+\rho^2}\left(\alpha\frac{\sin\rho}{\rho} - \cos\rho + e^{-\alpha}\right)(hc_\nu + kd_\nu)$$

$$+\left[\frac{1}{\alpha\rho^2}-\frac{1}{\alpha^2+\rho^2}\left(\alpha\frac{\cos\rho}{\rho^2}+\frac{\sin\rho}{\rho}+\frac{e^{-\alpha}}{\alpha}\right)\right]\left[\frac{1}{2}\beta(hd_\nu-kc_\nu)-(h^2+k^2)f_\nu\right], \tag{B1f}$$

$$\mathring{c}_{rs}=-\mathring{c}_{sr}:=c_{rs}-2\left(h\frac{\sin\rho-\rho}{\rho^3}+k\beta\frac{\cos\rho-1+\frac{1}{2}\rho^2}{\rho^4}\right)c_{[r}f_{s]}+2d_{[r}\left[h\frac{\cos\rho-1+\frac{1}{2}\rho^2}{\rho^4}\left(\beta f_{s]}-\frac{1}{2}kc_{s]}\right)\right.$$

$$\left.-\frac{\sin\rho-\rho}{\rho^3}\left(\frac{1}{2}\beta c_{s]}+kf_{s]}\right)\right]-\left[\frac{1}{4}(\mathring{c}_r\mathring{c}_s+\mathring{d}_r\mathring{d}_s)+\mathring{f}_r\mathring{f}_s\right]\epsilon(s-r),\quad r\neq s=2,\ldots,\nu-1, \tag{B1g}$$

$$\mathring{c}_{r\nu}=-\mathring{c}_{\nu r}:=\frac{1-e^{-\alpha}}{\alpha}c_{r\nu}-\frac{1}{2}\frac{1}{\alpha^2+\rho^2}\left(\alpha\frac{\sin\rho}{\rho}-\cos\rho+e^{-\alpha}\right)(c_rc_\nu+d_rd_\nu+4f_rf_\nu)$$

$$+\left[\frac{1}{\alpha\rho^2}-\frac{1}{\alpha^2+\rho^2}\left(\alpha\frac{\cos\rho}{\rho^2}+\frac{\sin\rho}{\rho}+\frac{e^{-\alpha}}{\alpha}\right)\right](2hc_{[r}f_{\nu]}+2kd_{[r}f_{\nu]}-\beta c_{[r}d_{\nu]})$$

$$-\frac{1}{2\rho^2}\left[\frac{1}{\alpha^2+\rho^2}\left(\alpha\frac{\sin\rho}{\rho}-\cos\rho+e^{-\alpha}\right)+\frac{e^{-\alpha}-1+\alpha}{\alpha^2}\right][hk(c_rd_\nu+c_\nu d_r)$$

$$+2\beta k(c_rf_\nu+c_\nu f_r)-2\beta h(d_rf_\nu+d_\nu f_r)-k^2(c_rc_\nu+d_rd_\nu)-4\beta^2f_rf_\nu],$$

$$r=2,..,\nu-1, \tag{B1h}$$

where

$$\rho^2:=\beta^2+h^2+k^2 \tag{B2a}$$

and

$$c_{[r}f_{s]}:=\tfrac{1}{2}(c_rf_s-c_sf_r). \tag{B2b}$$

[1]P. Furlan and R. Rączka, "Nonlinear spinor representations," J. Math. Phys. **26**, 3021 (1985).

[2]E. Cartan, *The Theory of Spinors* (Hermann, Paris, 1966).

[3]P. Furlan and R. Rączka, "Intrinsic nonlinear spinor wave equations associated with nonlinear spinor representations," J. Math. Phys. **27**, 1883 (1986).

[4]W. Magnus, "On the exponential solution of differential equations for a linear operator," Commun. Pure Appl. Math. **7**, 649 (1954).

Decomposition of the SO*(8) enveloping algebra under U(4)⊃U(3)

R. Le Blanc and D. J. Rowe
Department of Physics, University of Toronto, Toronto, Ontario, Canada M5S 1A7

(Received 29 September 1986; accepted for publication 7 January 1987)

The existence of a complete set of SU(3) tensor operators in the enveloping algebra of SO*(8) is demonstrated. The analysis recasts a parallel analysis by Biedenharn and Flath [Commun. Math. Phys. **93**, 143 (1984)] concerning an SO(6,2) model for SU(3) in the isomorphic but simpler framework of an SO*(8) model.

I. INTRODUCTION

It has been shown independently by Biedenharn and Flath[1] and Bracken and MacGibbon[2] that the fundamental unirrep of SO(6,2) defines a model for SU(3) in the sense of Bernšteĭn *et al.*[3] According to their definition, a representation space is called a model for a group G if it contains precisely one irrep from every equivalence class of irreps of G.

Exploiting the local isomorphism SO(6,2)~SO*(8), Le Blanc and Rowe[4] gave a Bargmann representation of the SO(6,2) model for SU(3) which is simple, naturally unitary with respect to the Bargmann measure, and readily expressed in a Cartan basis. Furthermore, taking advantage of the natural embedding of the su(3) algebra in the so*(8) Lie algebra through the subgroup chain

$$SO^*(8)\supset U(4)\supset U(3)\supset SU(3),$$

where U(4)~SO(6)×SO(2) is the maximal compact subgroup of SO*(8), it was realized therein that the Cartan generators of the so*(8) Lie algebra are the fundamental Wigner operators as defined by Biedenharn and Louck (see, e.g., Louck[5]) in contrast with the SO(6,2) model which call for more complicated linear combinations of the generators of its Lie algebra for their realization.

Since the fundamental Wigner operators of SU(3) are the elementary building blocks for a complete set of basic SU(3) tensor operators, it is appropriate to try to identify the elements of this set in the enveloping algebra of SO*(8). Such an analysis has been carried out by Biedenharn and Flath[1] for the SO(6,2) model. The purpose of this paper is to recast their analysis in the much simpler SO*(8) model framework. Using only the elegant properties of the Gel'fand patterns required for the classification of the U(4)⊃U(3) tensors arising in the enveloping algebra of SO*(8), we rederive all their results in a simple deductive and therefore more transparent way.

II. REVIEW OF THE SO*(8) MODEL FOR SU(3)

A basis for the complexification of the so*(8) Lie algebra is given (Le Blanc and Rowe[4]) in terms of two four-dimensional Bargmann vectors ($g_{\alpha\mu}$; $\alpha=1,2$; $\mu=1,\ldots,4$) and with summation over repeated indices by

$$C_{\mu\nu}=\frac{1}{2}\left(g_{\alpha\mu}\frac{\partial}{\partial g_{\alpha\nu}}+\frac{\partial}{\partial g_{\alpha\nu}}g_{\alpha\mu}\right),\quad \mu,\nu=1,4,$$

$$=g_{\alpha\mu}\frac{\partial}{\partial g_{\alpha\nu}}+\delta_{\mu\nu}, \tag{2.1a}$$

$$A_{\mu\nu}=-A_{\nu\mu}=\begin{vmatrix} g_{1\mu} & g_{1\nu}\\ g_{2\mu} & g_{2\nu}\end{vmatrix}, \tag{2.1b}$$

$$B_{\mu\nu}=-B_{\nu\mu}=A^{\dagger}_{\mu\nu}=\begin{vmatrix} \dfrac{\partial}{\partial g_{1\mu}} & \dfrac{\partial}{\partial g_{1\nu}}\\ \dfrac{\partial}{\partial g_{2\mu}} & \dfrac{\partial}{\partial g_{2\nu}}\end{vmatrix}, \tag{2.1c}$$

where $(C_{\mu\nu})$ span the maximal compact subalgebra u(4) of so*(8) and $(A_{\mu\nu})$ and $(B_{\mu\nu})$ are, respectively, Cartan raising and lowering operators. The u(3) subalgebra (C_{ij}) is given by the restriction of the indices μ,ν in (2.1a) to $i,j=1,2,3$.

We have the following commutation relations:

$$\begin{aligned}[C_{\mu\nu},A_{\gamma\delta}]&=\delta_{\nu\gamma}A_{\mu\delta}+\delta_{\nu\delta}A_{\gamma\mu},\\ [C_{\mu\nu},B_{\gamma\delta}]&=-\delta_{\mu\gamma}B_{\nu\delta}-\delta_{\mu\delta}B_{\gamma\nu},\\ [B_{\mu\nu},A_{\gamma\delta}]&=\delta_{\delta\nu}C_{\gamma\mu}+\delta_{\gamma\mu}C_{\delta\nu}-\delta_{\nu\gamma}C_{\delta\mu}-\delta_{\mu\delta}C_{\gamma\nu}.\end{aligned} \tag{2.2}$$

Under u(4), A is a $\{1100\}$ tensor, B is a $\{00-1-1\}$ tensor while the elements of su(4) are the components of a $\{100-1\}$ self-conjugate tensor.

The lowest weight state for this fundamental representation of so*(8) is clearly given by the Bargmann vacuum

$$\langle g|0\rangle=1, \tag{2.3}$$

which carries an unirrep $\{1111\}$ of u(4).

Since the raising operators $(A_{\mu\nu})$ transform under the adjoint action of u(4) as the components of a $\{1100\}$ tensor, the polynomials of degree h_1 in $(A_{\mu\nu})$ transform as the components of a $\{h_1h_100\}$ tensor. These tensors then reduce under u(3) according to the branching rule

$$u(4)\downarrow u(3):\ \{h_1h_100\}\downarrow\sum_{h_2=0}^{h_1}\{h_1h_20\} \tag{2.4}$$

given by the usual betweenness conditions of the corresponding Gel'fand patterns. Since there is no multiplicity involved in the SO*(8)↓U(4) reduction, it follows that a U(3) invariant subspace of this so*(8) representation space can be uniquely labeled by its u(4)⊃u(3) quantum numbers

$$\left|\begin{matrix} h_1 & & h_1 & & 0 & & 0\\ & h_1 & & h_2 & & 0 & \\ & & & \eta & & & \end{matrix}\right\rangle=P^{\{h_1h_2\}}_{\eta}(A)|0\rangle, \tag{2.5a}$$

where $P^{\{h_1h_2\}}_{\eta}(A)$ is a polynomial of degree h_1 in the raising operators $(A_{\mu\nu})$ and η stands for any appropriate scheme to

 0022-2488/87/061231-06$02.50

label components of the corresponding su(3) unirrep $\{h_1h_2\}$. For example, η can stand for either the usual lower Gel'fand pattern or for the basis labels $(\delta)LM$ (Le Blanc and Rowe[6]) corresponding to a canonical SU(3)↓SO(3) reduction. When η stands for an SU(3) lowest weight, we have, from Eqs. (4.7) and Eq. (6.13) of Ref. 4,

$$\left|\begin{matrix} h_1 & & h_1 & & 0 & & 0 \\ & h_1 & & h_2 & & 0 & \\ & & & lw & & & \end{matrix}\right\rangle$$

$$= \left[\frac{h_2!}{h_1!(h_1-h_2)!}\right]^{1/2} C^{(h_1-h_2)}_{42} \left\{\frac{1}{h_1!\sqrt{h_1+1}} A^{h_1}_{12}\right\}|0\rangle$$

$$= \frac{1}{[(h_1+1)!h_2!(h_1-h_2)!]^{1/2}} A^{h_2}_{12} A^{h_1-h_2}_{14}|0\rangle. \quad (2.5b)$$

It does follow that the so*(8) unirrep $\{1111\}$ decomposes under the successive restrictions

$$\text{so*}(8)\downarrow\text{u}(4)\downarrow\text{u}(3):\ \{1111\}\downarrow\sum_{h_1}\{h_1h_100\}\downarrow\sum_{h_1h_2}\{h_1h_20\}. \quad (2.6)$$

Thus the $\{1111\}$ representation space for so*(8) contains precisely one representative of every equivalence of SU(3) irreducible representations. It is therefore, by definition, an SU(3) model space (Bernšteĭn *et al.*[3]).

In the following sections, we identify in the enveloping algebra of so*(8) a complete set of SU(3) tensor operators which act on this model space and which have well-defined shift properties. In classifying these tensors, we shall make use of the convenient and insightful labeling scheme by upper Gel'fand (operator) patterns introduced by Biedenharn and Louck (see, e.g., Louck[5]) who showed that a complete set of basic SU(3) tensors can be classified by means of operator patterns of the type

$$\begin{bmatrix} & & \delta_1 & & \\ & \gamma_1 & & \gamma_2 & \\ h_{13} & & h_{23} & & h_{33} \end{bmatrix} \quad (2.7)$$

with the usual betweenness conditions for γ_1, γ_2, and δ_1. Such a pattern indicates that the corresponding tensor maps the states of a U(3) representation $\{\lambda_1,\lambda_2,\lambda_3\}$ into a representation $\{\lambda_1+\Delta_1,\lambda_2+\Delta_2,\lambda_3+\Delta_3\}$, where

$$\Delta_1=\delta_1,\quad \Delta_2=\gamma_1+\gamma_2-\delta_1, \qquad \Delta_3=h_{13}+h_{23}+h_{33}-\gamma_1-\gamma_2. \quad (2.8)$$

Thus it maps an SU(3) representation $\{\Lambda_1\Lambda_2\}$ into a representation $\{\Lambda_1+\Delta_1-\Delta_3,\Lambda_2+\Delta_2-\Delta_3\}$.

To label a set of unit tensors for SU(3), Biedenharn and Flath restricted the U(3) patterns (2.7) to the subset of the type

$$\begin{bmatrix} & & \Gamma_{11} & & \\ & \Gamma_{12} & & \Gamma_{22} & \\ h_1 & & h_2 & & 0 \end{bmatrix}, \quad (2.9)$$

and declared the equivalence of the U(3) unirreps

$$\{\lambda_1\lambda_2\lambda_3\}\equiv\{\lambda_1+\Delta_3,\lambda_2+\Delta_3,\lambda_3+\Delta_3\}$$

in their model space. Thus they restricted to the subset of patterns (2.7) with $h_{33}=0$. For our purposes, however, it is more appropriate to restrict to the subset with $\Delta_3=0$. This is because the SU(3) states of our model space carry U(3) representations strictly of the type $\{\lambda_1,\lambda_2,0\}$; cf. Eq. (2.6). It follows that any tensor operator acting on the model space must have shift $\Delta_3=0$. We therefore classify tensors acting within the so*(8) representation space by patterns of the type (2.7) with the constraint

$$\sum_i h_{i3}=\gamma_1+\gamma_2. \quad (2.10)$$

The corresponding shifts (2.8) are then given by

$$\Delta_1=\delta_1,\quad \Delta_2=\gamma_1+\gamma_2-\delta_1\equiv\delta_2, \qquad \Delta_3=0. \quad (2.11)$$

Our labeling is clearly related to that of Biedenharn and Flath by

$$\begin{bmatrix} & & \delta_1 & & \\ & \gamma_1 & & \gamma_2 & \\ h_{13} & & h_{23} & & h_{33} \end{bmatrix} = \begin{bmatrix} & & \Gamma_{11}+h_{33} & & \\ & \Gamma_{12}+h_{33} & & \Gamma_{22}+h_{33} & \\ h_1+h_{33} & & h_2+h_{33} & & h_{33} \end{bmatrix}, \quad (2.12)$$

with

$$h_{33}=\Gamma_{12}+\Gamma_{22}-h_1-h_2\leqslant 0. \quad (2.13)$$

Note that, from (2.10) and (2.11),

$$\gamma_1+\gamma_2=\delta_1+\delta_2=\sum_i h_{i3}. \quad (2.14)$$

Note also that for given values of $h_1=h_{13}-h_{33}$, $h_2=h_{23}-h_{33}$, δ_1, and δ_2, there generally corresponds a multiplicity set of tensors

$$\begin{bmatrix} & & \delta_1 & & \\ & \gamma_1(\rho) & & \gamma_2(\rho) & \\ h_{13} & & h_{23} & & h_{33} \end{bmatrix}, \quad (2.15)$$

which can be indexed by an integer $0\leqslant\rho\leqslant\rho_{\max}$ with

$$\gamma_1(\rho)=\gamma_1^s-\rho,\quad \gamma_2(\rho)=\gamma_2^s+\rho, \quad (2.16)$$

$$\rho_{\max}=\min(\gamma_1^s-\delta_1,\gamma_1^s-h_{23},h_{23}-\gamma_2^s,\delta_1-\gamma_2^s), \quad (2.17)$$

where γ_1^s and γ_2^s, respectively, denote the maximum and minimum (stretched) values of γ_1 and γ_2 allowed by the Gel'fand betweenness conditions. Thus a basic SU(3) tensor can be identified by either of the two sets of labels

$$(h_{13},h_{23},h_{33},\delta_1,\rho), \quad (2.18a)$$

$$(h_1,h_2,\delta_1,\delta_2,\rho), \quad (2.18b)$$

from which one readily reconstructs the corresponding operator pattern.

It can be easily verified (cf. Proposition 4.1) that the generators of the so*(8) Lie algebra have the following U(3) tensorial and shift properties:

$$A^{\{110\}}_{jk}:\ \delta=(11),\quad B^{\{0-1-1\}}_{jk}:\ \delta=(-1-1),$$
$$A^{\{100\}}_{i4}:\ \delta=(10),\quad B^{\{00-1\}}_{i4}:\ \delta=(-10),\qquad (2.19)$$
$$C^{\{100\}}_{i4}:\ \delta=(01),\quad C^{\{00-1\}}_{4i}:\ \delta=(0-1).$$

Therefore the expressions for the fundamental Wigner operators are seen to be extremely simple in this model. Their Hermiticity relations are also clearly apparent.

III. DECOMPOSITION OF A UNDER SO*(8)⊃U(4)⊃U(3)

We now proceed to the decomposition of the SO*(8) enveloping algebra A under SU(3).

Proposition 3.1: Under SO*(8), A decomposes as

$$\{A\}\downarrow \sum_{p=0}^{\infty}\{00-p-p\}, \qquad (3.1)$$

where $\{00-p-p\}$ is a nonunitary finite-dimensional lowest weight irrep of SO*(8).

Proof: B^{p}_{34} is easily verified to be the unique lowest weight polynomial of degree p in the generators of the so*(8) Lie algebra. It satisfies the equations

$$\begin{aligned}&[B_{\mu\nu},B^{p}_{34}]=0,\\ &[C_{\mu\nu},B^{p}_{34}]=0,\quad \mu<\nu,\\ &[C_{11},B^{p}_{34}]=[C_{22},B^{p}_{34}]=0,\\ &[C_{33},B^{p}_{34}]=[C_{44},B^{p}_{34}]=-pB^{p}_{34}.\end{aligned}\qquad (3.2)$$

Q.E.D.

(See Biedenharn and Flath,[1] Proposition 7.3, for an equivalent statement.)

Proposition 3.3: The decomposition of $\{00-p-p\}$ under U(4) is given by the branching rule

SO*(8)↓U(4):

$$\{00-p-p\}\downarrow \sum_{p_1+p_2+p_3+p_4=p}\{p_1+p_3,p_1,-p_2,-p_2-p_3\}. \qquad (3.3)$$

Proof: The SO*(8) irreducible tensor $\{00-p-p\}$ decomposes as a sum of U(4) Racah tensors $T^{\{h_{\mu 4}\}}$ for which the U(4) lowest weight component is given by the product of commuting factors

$$T^{\{h_{\mu 4}\}}_{lw}=A^{p_1}_{12}B^{p_2}_{34}C^{p_3}_{14}C^{p_4}_{\mu\mu},\quad \sum_{\nu=1}^{4}p_\nu=p. \qquad (3.4)$$

Here A_{12}, B_{34}, C_{14}, and $C_{\mu\mu}$ are, respectively, lowest weight components of U(4) tensors of rank $\{1100\}$, $\{00-1-1\}$, $\{100-1\}$, and $\{0000\}$. It follows that the coupling is stretched and the resulting tensor $T^{\{h_{\mu 4}\}}$ is of rank $\{h_{\mu 4}\}$ with

$$\begin{aligned}&h_{14}=p_1+p_3,\quad h_{24}=p_1,\quad h_{34}=-p_2,\\ &h_{44}=-p_2-p_3.\end{aligned}\qquad (3.5)$$

Q.E.D.

The decomposition of a U(4) tensor into a sum of U(3) tensors is given by the betweenness conditions for the associated U(4)⊃U(3) Gel'fand patterns:

$$\mathrm{U}(4)\downarrow\mathrm{U}(3):\{h_{\mu 4}\}\downarrow\sum\{h_{i3}\},\quad h_{i-1,4}<h_{i3}<h_{i+1,4}. \qquad (3.6)$$

Observe, however, that from any given p level U(3) tensor in A, we can construct other ("old") tensors at a higher p level and of the same U(3) rank by multiplying it by arbitrary polynomials in (a) the U(1)⊂U(1)×SU(4)⊂U(4) scalar

$$C_{\mu\mu} \qquad (3.7a)$$

and (b) the U(1)⊂U(1)×SU(3)⊂SU(4) scalar

$$C^{\mathrm{tr}}_{ii}=C_{ii}-\tfrac{3}{4}C_{\mu\mu}. \qquad (3.7b)$$

It is therefore sufficient to restrict consideration to the subset of "new" p level U(3) tensors having no such factors. Let I denote the set of polynomials in the above two U(1) scalars. Evidently I is the U(1)×U(1) enveloping algebra which is an Abelian subalgebra of A. The set of new U(3) tensors in A is therefore isomorphic to the factor space A/I and we can identify A/I with the space of SU(3) basic tensors.

Factors of the type (3.7a) are easily suppressed by setting

$$p_4=0 \qquad (3.8)$$

in Eq. (3.4); thus we restrict the set of p $(=\Sigma_{\nu=1}^{4}p_\nu)$ level tensors to be subset with $p=\Sigma_{i=1}^{3}p_i$.

The suppression of factors of the type (3.7b) can be carried out as follows. Let

$$\begin{Bmatrix}h_{\nu 4}\\ h_{i3}\end{Bmatrix} \qquad (3.9)$$

denote a U(3)⊂U(4) tensor in A. Since C^{tr}_{ii} is the U(3) scalar component of the su(4) Lie algebra of (self-conjugate) rank $\{100-1\}$, the tensor (3.9) is an old U(3) tensor if it can be expressed as a product

$$\begin{Bmatrix}1 & 0 & 0 & -1\\ 0 & 0 & 0 & \end{Bmatrix}\times\begin{Bmatrix}h_{14}-1 & & h_{24} & & h_{34} & & h_{44}+1\\ & h_{13} & & h_{23} & & h_{33} & \end{Bmatrix}. \qquad (3.10)$$

Thus only the U(3) tensors $\{h_{i3}\}\subset\{h_{\nu 4}\}$ for which

$$h_{13}=h_{14}\quad\text{and/or}\quad h_{33}=h_{44} \qquad (3.11)$$

are new.

Finally, each U(3) tensor is also an SU(3) tensor of rank

$$\{h_1h_2\}\equiv\{h_{13}-h_{33},h_{23}-h_{33}\}. \qquad (3.12)$$

As an example, we trivially have for the $p=0$ level the identity operator

$$T^{\{0000\}}\sim 1. \qquad (3.13)$$

For the $p=1$ level, we have the following three cases:

$$(p_i)=(100),\quad T^{\{1100\}}_{lw}\sim A_{12},$$
$$\mathrm{U}(4)\downarrow\mathrm{U}(3):\{1100\}\downarrow\{110\}\oplus\{100\}; \qquad (3.14)$$
$$(p_i)=(010),\quad T^{\{00-1-1\}}_{lw}\sim B_{34},$$
$$\mathrm{U}(4)\downarrow\mathrm{U}(3):$$
$$\{00-1-1\}\downarrow\{0-1-1\}\oplus\{00-1\}; \qquad (3.15)$$
$$(p_i)=(001),\quad T^{\{100-1\}}_{lw}\sim C_{14},$$
$$\mathrm{U}(4)\downarrow\mathrm{U}(3):$$
$$\{100-1\}\downarrow\{100\}\oplus\{10-1\}\oplus\{000\}\oplus\{00-1\}. \qquad (3.16)$$

Therefore

$$\{1100\}_{\text{new}} \downarrow \{110\} \oplus \{100\},$$
$$\{00-1-1\}_{\text{new}} \downarrow \{0-1-1\} \oplus \{00-1\}, \tag{3.17}$$
$$\{100-1\}_{\text{new}} \downarrow \{100\} \oplus \{10-1\} \oplus \{00-1\}.$$

The identification of all new SU(3) basic tensors for $p=1$ is unambiguous and their lowest weight components are given by

$$\left\{\begin{matrix} 1 & 1 & 0 & 0 \\ & 1 & 1 & 0 \\ & & lw & \end{matrix}\right\} \sim A_{12}, \quad \left\{\begin{matrix} 0 & 0 & -1 & -1 \\ & 0 & -1 & -1 \\ & & lw & \end{matrix}\right\} \sim B_{23},$$
$$\left\{\begin{matrix} 1 & 1 & 0 & 0 \\ & 1 & 0 & 0 \\ & & lw & \end{matrix}\right\} \sim A_{14}, \quad \left\{\begin{matrix} 0 & 0 & -1 & -1 \\ & 0 & 0 & -1 \\ & & lw & \end{matrix}\right\} \sim B_{34},$$
$$\left\{\begin{matrix} 1 & 0 & 0 & -1 \\ & 1 & 0 & 0 \\ & & lw & \end{matrix}\right\} \sim C_{14}, \quad \left\{\begin{matrix} 1 & 0 & 0 & -1 \\ & 0 & 0 & -1 \\ & & lw & \end{matrix}\right\} \sim C_{43}, \tag{3.18}$$
$$\left\{\begin{matrix} 1 & 0 & 0 & -1 \\ & 1 & 0 & -1 \\ & & lw & \end{matrix}\right\} \sim C_{13},$$

which, incidentally, close upon an $hw(3)$ algebra.

Observe that altogether, there are nine SU(3) lowest weight operators, old and new, at level $p=1$;

$$C_{\mu\mu}, C^{\text{tr}}_{ii}, A_{12}, A_{14}, B_{34}, B_{23}, C_{14}, C_{43}, C_{13}. \tag{3.19}$$

In Biedenharn and Flath's[1] SO(6,2) ~ SO*(8) realization, the two old $p=1$ SU(3) [also SU(3) invariant] tensors are given, to within additive constants, by their "quark" and "antiquark" number operators (see also Bracken and McGibbon[2]). Ignoring $C_{\mu\mu}$ and C^{tr}_{ii}, there are seven new SU(3) operators at the $p=1$ level.

Consider now the $p=2$ level. Since there are a total of nine $p=1$ SU(3) tensors, we expect, at first sight, a total of $9\times 10/2 = 45$ (old and new) symmetrical $p=2$ SU(3) tensors. In fact, listing the complete set of $U(4) \supset U(3)$ $p=2$ tensors using Proposition 3.3 and the betweenness conditions of the corresponding Gel'fand patterns, we find only 43 tensors. This implies the existence of two linear relationships between the SU(3) quadratic tensors. Fortunately, there is no need to seek these relationships as the 43 linearly independent tensors are obtained unambiguously from their corresponding U(4) parents [see Eqs. (4.7) and (4.9) below]. This obviates the necessity of imposing constraints, to restrict a linearly independent set, as in Biedenharn and Flath's analysis. Since our analysis is based on the $U(4)\downarrow U(3)$ branching rules, we automatically obtain linearly independent tensors and thus confirm a previous conjecture (Le Blanc and Rowe[4]) that a primary classification of SO*(8) tensors under U(4) would eliminate any ambiguities in the identification of the SU(3) tensors in A.

Applying the restrictions (3.8) and (3.11), we get the following new SU(3) tensors for the $p=2$ level:

$(p_i) = (200)$: $\{2200\}_{\text{new}} \downarrow \{220\} \oplus \{210\} \oplus \{200\}$,

$(p_i) = (110)$: $\{11-1-1\}_{\text{new}} \downarrow \{11-1\} \oplus \{10-1\} \oplus \{1-1-1\}$,

$(p_i) = (101)$: $\{210-1\}_{\text{new}} \downarrow \{210\} \oplus \{21-1\} \oplus \{200\} \oplus \{20-1\} \oplus \{11-1\} \oplus \{10-1\}$,

$(p_i) = (020)$: $\{00-2-2\}_{\text{new}} \downarrow \{00-2\} \oplus \{0-1-2\} \oplus \{0-2-2\}$,

$(p_i) = (011)$: $\{10-1-2\}_{\text{new}} \downarrow \{10-1\} \oplus \{10-2\} \oplus \{1-1-1\} \oplus \{1-1-2\} \oplus \{00-2\} \oplus \{0-1-2\}$,

$(p_i) = (002)$: $\{200-2\}_{\text{new}} \downarrow \{200\} \oplus \{20-1\} \oplus \{20-2\} \oplus \{10-2\} \oplus \{00-2\}$.

There are thus 26 new $p=2$ SU(3) tensors (as in Biedenharn and Flath[1]).

The classification of SU(3) tensors at any p level is thus very straightforward.

IV. SU(3) SHIFT PROPERTIES OF THE SO*(8)⊃U(4)⊃U(3) TENSORS

Proposition 4.1: The $SO^*(8) \supset U(4) \supset U(3)$ tensors (3.9) of the factor space A/I have well-defined U(3) shift properties given by

$$\delta_1 = p_1 - p_2 = h_{14} + h_{44} = h_{24} + h_{34}, \tag{4.1a}$$

$$\delta_2 = \sum_i h_{i3} - \delta_1. \tag{4.1b}$$

Proof: First, note that when acting on a state (2.5) of the model space, the tensor (3.9)

$$\left\{\begin{matrix} h_{\mu 4} \\ h_{i3} \end{matrix}\right\} = \left\{\begin{matrix} p_1+p_3 & p_1 & -p_2 & -p_2-p_3 \\ & h_{13} & h_{23} & h_{33} \end{matrix}\right\}$$

has a δ_1 shift given by the difference between its number of Cartan raising operators [the number of A's in (3.4)] and its number of Cartan lowering operators [the number of B's in (3.4)],

$$\delta_1 = p_1 - p_2.$$

Then, note that the U(4) weight operator C_{44} has eigenvalue $h_1 - h_2$,

$$C_{44}\left|\begin{matrix} h_1 & & h_1 & & 0 & & 0 \\ & h_1 & & h_2 & & 0 & \\ & & & \eta & & & \end{matrix}\right\rangle$$
$$= (h_1 - h_2)\left|\begin{matrix} h_1 & & h_1 & & 0 & & 0 \\ & h_1 & & h_2 & & 0 & \\ & & & \eta & & & \end{matrix}\right\rangle,$$

on a model space state. There is therefore no C_{44} weight multiplicity in the model space as a U(3) invariant subspace is unambiguously identified by the U(4) label $\{h_1 h_1 00\}$ and the C_{44} weight $(h_1 - h_2)$.

Now the tensor (3.9) has a C_{44} weight given by

$$\sum_\mu h_{\mu 4} - \sum_i h_{i3} = 2(p_1 - p_2) - \sum_i h_{i3}.$$

The tensor will therefore take a U(3) irreducible subspace $\{h_1 h_1 00\} \supset \{h_1 h_2 0\}$ to a new U(3) irreducible subspace with U(4) $\supset$ U(3) labels given by

$$\{h_1 + \delta_1, h_1 + \delta_1, 0, 0\} \supset \{h_1 + \delta_1, h_2 + \delta_2, 0\},$$

where, from the additivity of the C_{44} weight,

$$h'_1 - h'_2 = h_1 - h_2 + \sum_\mu h_{\mu 4} - \sum_i h_{i3},$$

which leads to the result

$$\delta_2 = \sum_i h_{i3} - \delta_1$$

[cf. Eq. (2.14)]. We thus conclude that the tensor (3.9) has unambiguous U(3) shift properties.

Q.E.D.

We now proceed to prove that the tensors (3.9), classified in the last section according to their SO*(8) $\supset$ U(4) $\supset$ U(3) properties, can be put in one-to-one correspondence with the abstract set (2.15) of U(3) shift tensors.

We first consider, as in Biedenharn and Flath,[1] the subset of p level tensors of SU(3) rank $\{h_1 = p, h_2\}$ and shifts $(\delta_1 \delta_2)$. [Recall from (2.18) that an SU(3) tensor can be identified by either $(h_{13}, h_{23}, h_{33}, \delta_1, \rho)$ or $(h_1, h_2, \delta_1, \delta_2, \rho)$ with

$$h_1 = h_{13} - h_{33}, \quad h_2 = h_{23} - h_{33}, \quad \sum_i h_{i3} = \delta_1 + \delta_2.]$$

From (4.1a) and condition $h_1 = p$, we have two equations for (p_1, p_2, p_3), namely,

$$p_1 + p_2 + p_3 = h_{13} - h_{33}, \quad p_1 - p_2 = \delta_1. \tag{4.2}$$

The third equation is given by the constraint (3.11). We have either

$$h_{13} = h_{14} = p_1 + p_3 \Rightarrow h_{33} = h_{34} = -p_2 \tag{4.3a}$$

or

$$h_{33} = h_{44} = -p_2 - p_3 \Rightarrow h_{13} = h_{24} = p_1. \tag{4.3b}$$

Solving for the p_i's, we obtain a unique solution for a given $\{h_1, h_2, \delta_1, \delta_2, \rho = 0\}$ which we identify with the stretched ($\rho = 0$) tensors of the abstract set (2.15) of shift tensors by the following proposition.

Proposition 4.4: To every SU(3) tensor of rank $\{h_1 h_2\}$, of given shift (δ_1, δ_2) and with stretched operator pattern, we can associate a $p^s = h_1$ level tensor

$$\begin{Bmatrix} h_{\mu 4} \\ h_{i3} \end{Bmatrix} = \begin{Bmatrix} p_1 + p_2 & & p_1 & & -p_2 & & -p_2 - p_3 \\ & h_{13} & & h_{23} & & h_{33} & \end{Bmatrix},$$

with U(4) $\supset$ U(3) labels given by

$$\begin{Bmatrix} h_{13} & & \delta_1 - h_{33} & & h_{33} & & -(h_{13} - \delta_1) \\ & h_{13} & & h_{23} & & h_{33} & \end{Bmatrix}, \tag{4.4a}$$

if $\delta_1 - h_{33} \leqslant h_{13}$ and by

$$\begin{Bmatrix} \delta_1 - h_{33} & & h_{13} & & -(h_{13} - \delta_1) & & h_{33} \\ & h_{13} & & h_{23} & & h_{33} & \end{Bmatrix}, \tag{4.4b}$$

if $\delta_1 - h_{33} \geqslant h_{13}$.

Now, for a $(p = h_1 - \rho)$ level SO*(8) $\supset$ U(4) $\supset$ U(3) tensor (3.9) with fixed SU(3) quantum labels $\{h_1 h_2\}$ to also have fixed U(3) shift properties (δ_i), we must have, from (3.11) and (4.1),

$$p_1(\rho) = p_1^s - \rho, \quad p_2(\rho) = p_2^s - \rho, \quad p_3(\rho) = p_3^s + \rho. \tag{4.5}$$

We thus deduce the following proposition.

Proposition 4.6: A multiplicity set of SU(3) tensors (2.5) will have a unique representative on the p $(= p_1 + p_2 + p_3)$ levels starting with the first (stretched) tensor on the level $p = p^s = h_1$ and the following tensors, indexed by the integer ρ [Eq. (2.17)], on the consecutive $p = p^s - \rho$ levels. The corresponding U(4) $\supset$ U(3) ($\{h_{\mu 4}\} \supset \{h_{i3}\}$) labels for these representatives are given by

$$\begin{Bmatrix} h_{13} & & \delta_1 - h_{33} - \rho & & h_{33} + \rho & & -(h_{13} - \delta_1) \\ & h_{13} & & h_{23} & & h_{33} & \end{Bmatrix}, \tag{4.6a}$$

if $\delta_1 - h_{33} \leqslant h_{13}$ and by

$$\begin{Bmatrix} \delta_1 - h_{33} & & h_{13} - \rho & & -(h_{13} - \delta_1) + \rho & & h_{33} \\ & h_{13} & & h_{23} & & h_{33} & \end{Bmatrix}, \tag{4.6b}$$

if $\delta_1 - h_{33} \geqslant h_{13}$.

It can be verified that for $\rho \leqslant \rho_{\max}$, all labels are fully consistent with the betweenness conditions of all [upper and lower, U(4) and U(3)] Gel'fand patterns involved. Propositions 4.4 and 4.6 are equivalent to Theorem 8.12 of Biedenharn and Flath.[1]

Acting with the raising operator C_{42} on the expression (3.4) for the U(4) *lw* tensor component, we find that the lowest weight component of the shift tensor (4.4a) is given by

$$C_{42}^{(\delta_1 - h_{33} - h_{23})}\{A_{12}^{\delta_1 - h_{33}} B_{34}^{-h_{33}}\} C_{14}^{h_{13} - \delta_1 + h_{33}}, \tag{4.7}$$

where

$$\begin{aligned} C_{42}^{(0)}\{X\} &= X, \\ C_{42}^{(1)}\{X\} &= [C_{42}, X], \\ C_{42}^{(2)}\{X\} &= [C_{42}, [C_{42}, X]], \quad \text{etc.} \end{aligned} \tag{4.8}$$

Similarly, acting with C_{42} and C_{43}, we find that the lowest weight component of the shift tensor (4.6a) is given by

$$C_{42}^{(\delta_1 - h_{33} - h_{23} - \rho)}\{A_{12}^{\delta_1 - h_{33} - \rho} B_{34}^{-h_{33} - \rho}\} C_{14}^{h_{13} - \delta_1 + h_{33}} C_{13}^{\rho}. \tag{4.9}$$

Similar expressions can be obtained for the tensors (4.4b) and (4.6b). Equations (4.4), (4.6), (4.7), and (4.9) are also equivalent to Eqs. (8.21) and (8.22) of Biedenharn and Flath. The group theoretical structure of the tensors (4.9) guarantees their linear independence (Draayer and Akiyama,[7] Le Blanc[8]).

Finally, note that although extremely elegant and simple, the SO*(8) ~ SO(6,2) model for SU(3) does not offer (as yet) a group theoretical interpretation for the U(2) aspects of the upper pattern. Such an interpretation has been given elsewhere by the authors (Le Blanc and Rowe[9]) using an alternative model for SU(3) defined in a U(2) × SU(3) Bargmann space.

[1] L. C. Biedenharn and D. E. Flath, Commun. Math. Phys. **93**, 143 (1984).
[2] A. J. Bracken and J. H. MacGibbon, J. Phys. A: Math. Gen. **17**, 2581 (1984); A. J. Bracken, Commun. Math. Phys. **94**, 377 (1984).
[3] I. N. Bernšteĭn, I. M. Gel'fand, and S. I. Gel'fand, Funct. Anal. Appl. **9**, 332 (1975).
[4] R. Le Blanc and D. J. Rowe, J. Phys. A: Math. Gen. **19**, 1111 (1986).
[5] J. D. Louck, Am. J. Phys. **38**, 3 (1970).
[6] R. Le Blanc and D. J. Rowe, J. Phys. A: Math. Gen. **18**, 1891, 1905 (1985).
[7] J. P. Draayer and Y. Akiyama, J. Math. Phys. **14**, 1904 (1973).
[8] R. Le Blanc, Ph.D. thesis, University of Toronto, 1985.
[9] R. Le Blanc and D. J. Rowe, J. Phys. A: Math. Gen. **19**, 1093, 2913 (1986).

Indecomposable modules of the Poincaré algebra in an energy-cyclic angular momentum basis

Romuald Lenczewski
Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901

(Received 7 August 1986; accepted for publication 11 February 1987)

The universal enveloping algebra $\mathcal{U} = \mathcal{U}_+ \mathcal{U}_- \mathcal{H}$ of the Poincaré algebra iso(3,1) is considered. Infinite-dimensional induced modules $\mathcal{U}_+(\Gamma)$ are studied. Explicit formulas are obtained for $\mathcal{U}_+(\Gamma)$ in the Poincaré–Birkhoff–Witt basis. Then a change of basis is performed to an energy-cyclic angular momentum basis. For any $\Gamma \in \mathbb{C}^2$, $\mathcal{U}_+(\Gamma)$ turns out to be indecomposable. It can be represented as an infinite family of interacting Verma modules of the Lorentz algebra so(3,1). Finite-dimensional modules can be obtained as quotient modules of $\mathcal{U}_+(\Gamma)$ if $2\Gamma \in \mathbb{Z}^2$.

I. INTRODUCTION

Indecomposable modules of physically relevant Lie algebras have been suggested[1-3] for the description of unstable particles. In particular, indecomposable modules of the Poincaré algebra $\mathcal{P} = \text{iso}(3,1)$ may be useful in modeling unstable particles of arbitrary spin and their interactions. However, the Poincaré algebra, being a nonsemisimple Lie algebra, does not lend itself as easily to methods of the representation theory as semisimple Lie algebras (even in the case of finite-dimensional modules). On the other hand, its maximal solvable subalgebra is Abelian, hence it can be viewed as mathematically relatively easy to work with.

In fact, $\mathcal{P}$ is a semidirect product of the algebra of translations $\mathcal{K} = t(4)$ and the Lorentz algebra $\mathcal{L} = \text{so}(3,1)$. The Lie products satisfy $[\mathcal{L}, \mathcal{L}] = \mathcal{L}$, $[\mathcal{L}, \mathcal{K}] = \mathcal{K}$, $[\mathcal{K}, \mathcal{K}] = 0$. Therefore $\mathcal{K}$ is an Abelian ideal in $\mathcal{P}$ and this makes the analysis simpler than for other nonsemisimple Lie algebras (except, maybe, some low-dimensional ones).

Hence with both mathematical and physical motivations we resume[4] the study of the indecomposable modules of $\mathcal{P}$ (infinite dimensional as well as finite dimensional). This work is a natural extension of the study conducted by Gruber and Lenczewski.[4,5] Thus the results contained therein will be basic for the analysis that is carried out in this paper. Moreover, the modules considered previously[4,5] turn out to be incorporated into the present study as special cases. Therefore to a great extent we tried to use the same notation, although in several places it has been changed to avoid confusion.

In Sec. II we define the Poincaré algebra and introduce the raising and lowering operators.

In Sec. III we define the universal enveloping algebra of $\mathcal{P}$, $\mathcal{U}(\mathcal{P}) = \mathcal{U}_+ \mathcal{U}_- \mathcal{H}$, as a regular left $\mathcal{P}$ module. The induced modules $\mathcal{U}_+(\Gamma)$ (the analogs of Verma[6] modules of semisimple Lie algebras) are explicitly derived in the Poincaré–Birkhoff–Witt (PBW) basis. Although the PBW basis seems to be very natural, a basis more useful in physical applications, especially in the case of the chain $\text{so}(3) \subset \text{so}(3,1) \subset \text{iso}(3,1)$, is an angular momentum basis[5] (AMB). Therefore we need to introduce an AMB to the formalism.

In Sec. IV a certain AMB is defined. It turns out that it is not unique (in opposition to the Lorentz algebra case[5] or some less general modules of the Poincaré algebra[4]). We choose a certain energy-cyclic AMB which allows us to obtain the modules $\mathcal{U}_+(\Gamma)$ in a closed form.

In Sec. V we give explicit formulas for $\mathcal{U}_+(\Gamma)$ in the energy-cyclic AMB mentioned above. They are derived by using the induction method. Modules $\mathcal{U}_+(\Gamma)$ turn out to be indecomposable for all $\Gamma \in \mathbb{C}^2$. Each $\mathcal{U}_+(\Gamma)$ forms, in fact, a composition series. Moreover, when treated as a vector space, it can be written as

$$\mathcal{U}_+(\Gamma) = \sum_{\substack{M,\, m \in \mathbb{N} \\ m \le M}} \mathcal{U}_M^m$$

such that each $\mathcal{U}_M^m$ is a Verma $\mathcal{L}$ module. The module $\mathcal{U}_+(\Gamma)$ can be viewed as a sum of interacting Verma $\mathcal{L}$ modules.

In Sec. VI we give some examples of finite-dimensional indecomposable modules as quotient modules of $\mathcal{U}_+(\Gamma)$ [or its extension $\tilde{\mathcal{U}}_+(\Gamma)$] for certain $2\Gamma \in \mathbb{Z}^2$. Bases for the modules of dimensions 4, 5, 7, 8, 8′ (two different eight dimensional ones) are given as well as their $\text{so}(3) \subset \text{so}(3,1)$ decomposition. Thus we identify all but one of the finite-dimensional indecomposable modules of dimensions less than 9 according to their recent classification for low dimensions.[7]

The main result of this paper (contained in Sec. V) is of significance in its own right taking into account that the analogous result for the Lorentz algebra[5] reproduced the Gel'fand–Naimark[8,9] representations as a special case. At the same time it provides an important step forward in our quest for a unified approach to finite-dimensional indecomposable modules of $\mathcal{P}$.

In the sequel we use the abbreviations introduced above as well as the following: inf(fin)-in(ir)-mod [infinite(finite) dimensional indecomposable (irreducible) module]. All algebras are considered to be over $\mathbb{C}$. For general references one is referred to Humphreys[10] and Dixmier.[11]

II. PRELIMINARIES

The defining relations for the Poincaré algebra $\mathcal{P}$ are usually given as the following set of Lie products:

 0022-2488/87/061237-06$02.50

$$
\begin{aligned}
&[P_\mu, P_\nu] = 0\,,\\
&[M_{\mu\nu}, P_\alpha] = i(g_{\mu\alpha}P_\nu - g_{\nu\alpha}P_\mu)\,,\\
&[M_{\mu\nu}M_{\alpha\beta}]\\
&\quad = i(g_{\mu\alpha}M_{\nu\beta} - g_{\nu\alpha}M_{\mu\beta} + g_{\mu\beta}M_{\alpha\nu} - g_{\nu\beta}M_{\alpha\mu})\,,
\end{aligned}
\tag{1}
$$

where $M_{\mu\nu}, P_\alpha$ represent infinitesimal generators of rotations and translations, respectively, in the four-dimensional Minkowski space $[\mu,\nu,\alpha\in\{0,1,2,3\}$ and $g_{\mu\nu} = \mathrm{diag}(-1,1,1,1)]$. In our formalism it appears to be necessary to introduce another basis for $\mathscr{P}$, consisting of root vectors, vectors corresponding to positive roots (raising operators), and vectors corresponding to negative roots (lowering operators). The root vectors span the Cartan subalgebra of $\mathscr{P}$, i.e., the nilpotent subalgebra that equals its normalizer in (it is equal to the Cartan subalgebra of $\mathscr{L}$). They can be chosen as $h_3 = M_{12}$ and $f_3 = -iM_{03}$. Raising operators (with a subscript $+$) and lowering operators (with a subscript $-$) can be chosen as

$$
\begin{aligned}
&h_+ = M_{23} + iM_{13}, \quad h_- = M_{23} - iM_{13}\,,\\
&f_+ = -M_{02} - iM_{01}, \quad f_- = M_{02} - iM_{01}\,,\\
&k_+ = P_1 - iP_2, \quad k_- = P_1 + iP_2\,,\\
&r_+ = P_3 + P_0, \quad r_- = P_3 - P_0\,.
\end{aligned}
\tag{2}
$$

The advantage of using this basis stems from the fact that all structure constants lie in $\mathbb{Z}$ as can be seen below. In the modules investigated by Lenczewski and Gruber[4] the following basis was used: $p_+ = if_+, p_- = if_-, p_3 = if_3$ and $k_3 = P_3$, $k_0 = -iP_0$.

Introducing r_+ and r_- allows us to extend the modules obtained therein as will be seen throughout the paper. For notational convenience we are going to use $\mathscr{R}$ to represent so(3) and $\mathscr{F}$ to represent the $\mathbb{C}$-linear subspace spanned by f_+, f_-, f_3 (boosts). Clearly, $\mathscr{F}$ is not a subalgebra of $\mathscr{P}$ contrary to $\mathscr{K}$ and $\mathscr{L}$.

Using the above notation one can derive the following commutation relations in the new basis:

$$
\begin{aligned}
&[\mathscr{R},\mathscr{R}] \quad \begin{cases}[h_3, h_\pm] = \pm h_\pm\,,\\ [h_+, h_-]\end{cases} = 2h_3,\\
&[\mathscr{F},\mathscr{F}] \quad \begin{cases}[f_+, f_-] = 2h_3,\\ [f_3, f_\pm]\end{cases} = \pm h_\pm\,,\\
&[\mathscr{R},\mathscr{F}] \quad \begin{cases}[h_3, f_\pm] = \pm f_\pm\,,\\ [f_\pm, h_\mp] = \pm 2f_3,\\ [f_3, h_\pm] = \pm f_\pm\,,\end{cases}\\
&[\mathscr{F},\mathscr{K}] \quad \begin{cases}[f_-, r_\pm] = \pm k_-\,,\\ [f_+, r_\pm] = \pm k_+,\\ [f_\mp, k_\pm] = -r_- + r_+,\\ [f_3, r_\pm] = \pm r_\pm\,,\end{cases}\\
&[\mathscr{R},\mathscr{K}] \quad \begin{cases}[h_3, k_\pm] = \pm k_\pm\,,\\ [h_\pm, k_\mp] = \pm(r_- + r_+)\,,\\ [h_+, r_\pm] = -k_+\,,\\ [h_-, r_\pm] = k_-\,.\end{cases}
\end{aligned}
\tag{3}
$$

It is understood that either all upper or all lower signs (whether as subscripts or arithmetic symbols) are taken simultaneously.

III. INDUCED MODULES $\mathscr{U}_+(\Gamma)$ IN THE PBW BASIS

We come to the main object of interest: the universal enveloping algebra (UEA) of the algebra $\mathscr{P}$, $\mathscr{U}(\mathscr{P})$. It is defined as the quotient algebra $\mathscr{U}(\mathscr{P}) = \mathscr{T}(\mathscr{P})/\mathscr{I}$, where $\mathscr{T}(\mathscr{P})$ is the tensor algebra of $\mathscr{P}$ and $\mathscr{I}$ is the ideal generated by $x\otimes y - y\otimes x - [x,y]$, $x,y\in\mathscr{P}$. It is well known[10,11] that there exists a basis for $\mathscr{U}(\mathscr{P})$ that consists of standard monomials. This is the Poincaré–Birkhoff–Witt (PBW) basis.

The left tensor multiplication by $\mathscr{P}$ makes out of $\mathscr{U}(\mathscr{P})$ a regular left module. Various quotients of $\mathscr{U}(\mathscr{P})$ can be considered. For example, for given $\Gamma\in\mathbb{C}^2$ one can define a module $\mathscr{U}_+(\Gamma)$ as a quotient module $\mathscr{U}(\mathscr{P})/\mathscr{J}$, where the ideal $\mathscr{J}$ is generated by lowering operators and $h_3 - \Gamma_1, f_3 - \Gamma_2$. Similarly, $\mathscr{U}_-(\Gamma)$ is a quotient module $\mathscr{U}(\mathscr{P})/\mathscr{G}$ where the ideal $\mathscr{G}$ is generated by raising operators and $h_3 - \Gamma_1, f_3 - \Gamma_2$. We will restrict our attention to modules $\mathscr{U}_+(\Gamma)$ later on giving a connection between $\mathscr{U}_+(\Gamma)$ and $\mathscr{U}_-(\Gamma)$ through an automorphism of $\mathscr{P}$. The following relations are obtained on $\mathscr{U}_+(\Gamma)$:

$$
\begin{aligned}
&h_3X = (\Gamma_1 + n + s + p)X\,, \quad h_+X = X(n+1),\\
&f_+X = X(s+1), \quad k_+X = X(p+1)\,,\\
&h_-X = n(-2\Gamma_1 - 2p - 2s - n + 1)X(n-1)\\
&\qquad - 2s(v + \Gamma_2)X(s-1)\\
&\qquad - s(s-1)X(s+1, n-2) + pX(p-1, v+1)\,,\\
&f_-X = s(-2\Gamma_1 - 2p - s + 1 - 2n)X(s-1)\\
&\qquad + 2n(-v - \Gamma_2)X(n-1)\\
&\qquad - n(n-1)X(s+1, n-2) + pX(p-1, v+1),
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
&f_3X = (\Gamma_2 + v)X + nX(s+1, n-1) + sX(n+1, s-1)\,,\\
&k_-X = -sX(s-1, v+1) - nX(n-1, v+1)\\
&\qquad + s(s-1)X(s-2, p+1)\\
&\qquad - n(n-1)X(n-2, p+1)\,,\\
&r_-X = sX(s-1, p+1) + nX(n-1, p+1),\\
&r_+X = X(v+1) - sX(s-1, p+1) + nX(n-1, p+1)\,,
\end{aligned}
$$

where $X = X(n,s,p,v) = h_+^n f_+^s k_+^p r_+^v$, $n,s,p,v\in\mathbb{N}$, and only the parameters that are altered by the action of the operator on the left-hand side of each equation are indicated. The basis elements represent the natural basis induced by the PBW basis of the UEA, hence we refer to it also as a PBW basis of $\mathscr{U}_+(\Gamma)$.

The above equations show that $\mathscr{U}_+(\Gamma)$ is an inf-in-mod for all $\Gamma\in\mathbb{C}^2$. The value of v does not decrease. Hence, on quotients $\mathscr{U}_+(\Gamma)/\mathscr{O}_i$, where the ideal $\mathscr{O}_i$ is generated by r_+^i, one obtains induced relations similar to those exhibited by Eq. (4). In particular, if we consider the ideal $\mathscr{O}_1$ we obtain the modules $\Omega_+(\Gamma)$ (Ref. 4) recently examined. Presently, however, the structure under consideration is richer and will give new results. The next step consists in performing a change of basis from the PBW basis into an AMB.

IV. AN ENERGY-CYCLIC ANGULAR MOMENTUM BASIS

In this section we are going to concentrate on finding an AMB for $\mathscr{U}_+(\Gamma)$. We will use the fact that $[h_-, k_0] = 0$ to generate an AMB that is energy cyclic.

To find an AMB we define an $\mathscr{R}$-maximal ($\mathscr{R}$-minimal) vector to be an eigenvector Y of h_3 that satisfies $h_+Y = 0$ ($h_-Y = 0$). They generate cyclic $\mathscr{R}$ modules, whose bases will span $\mathscr{P}$ modules $\mathscr{U}_+(\Gamma)$ [$\mathscr{U}_-(\Gamma)$].

Let $\mathscr{Y} = \{Y_j,\ j\in J\}$ be the set of all $\mathscr{R}$-minimal vectors, where J is a certain index set (or a multi-index set). Then, $\mathrm{AMB}\subset\{h_+^n\mathscr{Y}\} = \{h_+^n Y_j,\ j\in J\}$. In the case of the Lorentz algebra[5] or the modules $\Omega_+(\Gamma)$ of the Poincaré algebra,[4] $\mathrm{AMB} = \{h_+^n\mathscr{Y}\}$. However, in $\mathscr{U}_+(\Gamma)$ we have $\mathrm{AMB}\neq\{h_+^n\mathscr{Y}\}$. In fact, $\mathrm{card}(\mathrm{AMB}) = \aleph_0$, whereas $\mathrm{card}\{h_+^n\mathscr{Y}\} = \mathfrak{c}$. This leads to the nonuniqueness of the AMB. In this paper we shall choose an AMB which is energy cyclic.

One arrives at the following form of minimal vectors:

$$\{Y_{NM}\} = \left\{ \sum_{\substack{s+p<N \\ p<M}} c_{sp}^{NM} X(N-s-p,s,p,M-p) \right\},$$

where $N\geqslant 0$, $M\geqslant 0$, and $h_3 Y_{NM} = (\Gamma_1 + N)Y_{NM}$. The constants c_{sp}^{NM} satisfy the following recurrence relation:

$$\begin{aligned}(N-s-p)(-2\Gamma_1 - p - s - N + 1)c_{sp}^{NM} \\ - 2(M+\Gamma_2 - p)(s+1)c_{s+1,p}^{NM} \\ - (s+2)(s+1)c_{s+2,p}^{NM} - (p+1)c_{s,p-1}^{NM} = 0 . \end{aligned} \tag{5}$$

Because of the abundance of solutions of the recurrence relation we need to choose certain representatives of the sets of $\mathscr{R}$-minimal vectors in the form given. We are going to do it in a fairly natural manner. We will try to reduce the problem to the Lorentz algebra case (where the solutions are unique and generate the bases of $\mathscr{L}$ modules) in a sufficient number of cases to generate the whole set of $\mathscr{R}$-minimal vectors that will be needed by applying the energy operator.

Let us set $M = 0$ first. This is the prototypical Lorentz algebra case. We obtain the following recurrence relation:

$$\begin{aligned}(N-s)(-2\Gamma_1 - s - N + 1)c_{s0}^{N0} - 2\Gamma_2(s+1)c_{s+1,0}^{N0} \\ - (s+2)(s+1)c_{s+2,0}^{N0} = 0 . \end{aligned} \tag{6}$$

Then Y_{N0} can be given as

$$Y_{N0} = \sum_{s<N} c_{s0}^{N0} X(N-s,s,0,0) \tag{7}$$

and the constants c_{s0}^{N0} can be uniquely determined. Their explicit form is fairly complicated and will not be needed in the sequel since our derivation of the modules $\mathscr{U}_+(\Gamma)$ in the new basis proceeds by induction.

Let us set $N = 0$ now. Then the only possible $\mathscr{R}$-minimal vector is $X(0,0,0,M)$ (up to a constant factor, of course). Thus we let

$$Y_{0M} = X(0,0,0,M) . \tag{8}$$

Now, if $N\neq 0$ and $M\neq 0$ then the solutions for $\mathscr{R}$-minimal vectors span at least a two-dimensional vector space. We choose the representatives of the sets $\{Y_{NM}\}$ in the following way:

$$Y_{NM} = \sum_{s<N} c_{s0}^{NM} X(N-s,s,0,M) , \tag{9}$$

with c_{s0}^{NM} satisfying the recurrence relation:

$$\begin{aligned}(N-s)(-2\Gamma_1 - s - N + 1)c_{s0}^{NM} \\ - 2(M+\Gamma_2)(s+1)c_{s+1,0}^{NM} \\ - (s+2)(s+1)c_{s+2,0}^{NM} = 0 . \end{aligned} \tag{10}$$

Thus Y_{NM} have exactly the same form as Y_{N0} except that Γ_2 is replaced by $\Gamma_2 + M$. Hence we will obtain in this manner $\mathscr{R}$-minimal vectors for $\mathscr{L}$ modules.

Other $\mathscr{R}$-minimal vectors are defined inductively using the energy operator as the generator. Thus we obtain the following set of $\mathscr{R}$-minimal vectors:

$$Y_{NM}^0 = Y_{NM} , \quad Y_{NM}^m = (2ik_0)^m Y_{N,M-m}^0 , \tag{11}$$

where $m\leqslant\min(N,M)$. Now,

$$\begin{aligned}\mathrm{AMB} = \{h_+^n Y_{NM}^m = Y_{NM}^{nm} ,\ n,m,N,M\in\mathbb{N}, \\ m\leqslant\min(N,M)\} . \end{aligned}$$

We call the above four parameter basis an energy-cyclic AMB since it is generated by the energy operator k_0. This basis will allow us to obtain the modules $\mathscr{U}_+(\Gamma)$ in a closed form.

V. MODULES $\mathscr{U}_+(\Gamma)$ IN AN ENERGY CYCLIC ANGULAR MOMENTUM BASIS

In this section we shall present the main result of this article. It will consist of the formulas for the modules $\mathscr{U}_+(\Gamma)$ in the basis defined in Sec. IV. They are obtained by the induction method and their derivation involves fairly extensive calculations. One obtains

$$h_+ Y_{NM}^{nm} = Y_{NM}^{n+1,m} , \tag{12}$$

$$h_- Y_{NM}^{nm} = nF_N^n Y_{NM}^{n-1,m} , \tag{13}$$

$$h_3 Y_{NM}^{nm} = G_N^n Y_{NM}^{nm} , \tag{14}$$

$$\begin{aligned}f_+ Y_{NM}^{nm} = \left(\frac{N-m+1}{N+1}\right)Y_{N+1,M}^{nm} + \left(\frac{m}{N+1}\right)Y_{N+1,M}^{n,m-1} + \beta_{NMm} Y_{NM}^{n+1,m} \\ + \beta_{NMm}^{\downarrow} Y_{NM}^{n+1,m-1} + \alpha_{NMm} Y_{N-1,M}^{n+2,m} + \alpha_{NMm}^{\downarrow} Y_{N-1,M}^{n+2,m-1} , \end{aligned} \tag{15}$$

$$f_- Y_{NM}^{nm} = -\left(\frac{n(n-1)(N-m+1)}{N+1}\right)Y_{N+1,M}^{n-2,m} + \left(-\frac{n(n-1)m}{N+1}\right)Y_{N+1,M}^{n-2,m-1} + nF_N^n\beta_{NMm} Y_{NM}^{n-1,m}$$

$$+nF^n_N\beta^{\downarrow}_{NMm}Y^{n-1,m-1}_{NM}-F^n_N(F^n_N+1)\alpha_{NMm}Y^{nm}_{N-1,M}-F^n_N(F^n_N+1)\alpha^{\downarrow}_{NMm}Y^{n,m-1}_{N-1,M}\,, \tag{16}$$

$$f_3Y^{nm}_{NM}=\left(\frac{n(N-m+1)}{N+1}\right)Y^{n-1,m}_{N+1,M}+\left(\frac{nm}{N+1}\right)Y^{n-1,m-1}_{N+1,M}+G^n_N\beta_{NMm}Y^{nm}_{NM}$$
$$+G^n_N\beta^{\downarrow}_{NMm}Y^{nm}_{NM}-F^n_N\alpha_{NMm}Y^{n+1,m}_{N-1,M}-F^n_N\alpha^{\downarrow}_{NMm}Y^{n+1,m-1}_{N-1,M}\,, \tag{17}$$

$$k_0Y^{nm}_{NM}=(-i/2)Y^{n,m+1}_{N,M+1}\,, \tag{18}$$

$$k_3Y^{nm}_{NM}=\left(-\frac{n}{2(N+1)}\right)Y^{n-1,m+1}_{N+1,M+1}+\left(\frac{n}{2(N+1)}\right)Y^{n-1,m}_{N+1,M+1}+\tfrac{1}{2}G^n_N\beta^{\uparrow\uparrow}_{NMm}Y^{n,m+1}_{N,M+1}$$
$$+\tfrac{1}{2}G^n_N\beta^{\uparrow}_{NMm}Y^{nm}_{N,M+1}-\tfrac{1}{2}F^n_N\alpha^{\uparrow\uparrow}_{NMm}Y^{n+1,m+1}_{N-1,M+1}-\tfrac{1}{2}F^n_N\alpha^{\uparrow}_{NMm}Y^{n+1,m}_{N-1,M+1}\,, \tag{19}$$

$$k_+Y^{nm}_{NM}=\left(-\frac{1}{2(N+1)}\right)Y^{n,m+1}_{N+1,M+1}+\left(\frac{1}{2(N+1)}\right)Y^{nm}_{N+1,M+1}+\tfrac{1}{2}\beta^{\uparrow\uparrow}_{NMm}Y^{n+1,m+1}_{N,M+1}$$
$$+\tfrac{1}{2}\beta^{\uparrow}_{NMm}Y^{n+1,m}_{N,M+1}+\tfrac{1}{2}\alpha^{\uparrow\uparrow}_{NMm}Y^{n+2,m+1}_{N-1,M+1}+\tfrac{1}{2}\alpha^{\uparrow}_{NMm}Y^{n+2,m}_{N-1,M+1}\,, \tag{20}$$

$$k_-Y^{nm}_{NM}=\left(\frac{n(n-1)}{2(N+1)}\right)Y^{n-2,m+1}_{N+1,M+1}+\left(-\frac{n(n-1)}{2(N+1)}\right)Y^{n-2,m}_{N+1,M+1}+\tfrac{1}{2}nF^n_N\beta^{\uparrow\uparrow}_{NMm}Y^{n-1,m+1}_{N,M+1}$$
$$+\tfrac{1}{2}nF^n_N\beta^{\uparrow}_{NMm}Y^{n-1,m}_{N,M+1}-\tfrac{1}{2}F^n_N(F^n_N+1)\alpha^{\uparrow\uparrow}_{NMm}Y^{n,m+1}_{N-1,M+1}-\tfrac{1}{2}F^n_N(F^n_N+1)\alpha^{\uparrow}_{NMm}Y^{nm}_{N-1,M+1}\,, \tag{21}$$

where the coefficients $G^n_N, F^n_N, \beta_{NMm}, \beta^{\downarrow}_{NMm}, \beta^{\uparrow}_{NMm}, \beta^{\uparrow\uparrow}_{NMm}, \alpha_{NMm}, \alpha^{\downarrow}_{NMm}, \alpha^{\uparrow}_{NMm}, \alpha^{\uparrow\uparrow}_{NMm}$ are given below as functions of Γ:

$$\alpha_{NMm}=\frac{N(2\Gamma_1+N+m-2)((\Gamma_1+N-1)^2-(\Gamma_2+M-m)^2)}{(\Gamma_1+N-1)^2(2\Gamma_1+2N-3)(2\Gamma_1+2N-1)}\,,$$

$$\alpha^{\uparrow}_{NMm}=\frac{N((\Gamma_1+N-2)-(\Gamma_2+M-m))((\Gamma_1+N-1)-(\Gamma_2+M-m))}{(\Gamma_1+N-1)^2(2\Gamma_1+2N-3)(2\Gamma_1+2N-1)}\,,$$

$$\alpha^{\downarrow}_{NMm}=\frac{mN((\Gamma_1+n-2)-(\Gamma_2+M-m))((\Gamma_1+n-1)-(\Gamma_2+M-m))}{(\Gamma_1+N-1)^2(2\Gamma_1+2N-3)(2\Gamma_1+2N-1)}\,,$$

$$\alpha^{\uparrow\uparrow}_{NMm}=\frac{N((\Gamma_1+N-1)^2-(\Gamma_2+M-m)^2)}{(\Gamma_1+N-1)^2(2\Gamma_1+2N-3)(2\Gamma_1+2N-1)}\,,$$

$$\beta_{NMm}=\frac{(\Gamma_2+M-m)(\Gamma_1+m-1)}{(\Gamma_1+N)(\Gamma_1+N-1)}\,,$$

$$\beta^{\uparrow}_{NMm}=\frac{(\Gamma_1+N-1)-(\Gamma_2+M-m)}{(\Gamma_1+N)(\Gamma_1+N-1)}\,,$$

$$\beta^{\downarrow}_{NMm}=\frac{m((\Gamma_1+N-1)-(\Gamma_2+M-m))}{(\Gamma_1+N)(\Gamma_1+N-1)}\,,$$

$$\beta^{\uparrow\uparrow}_{NMm}=\frac{\Gamma_2+M-m}{(\Gamma_1+N)(\Gamma_1+N-1)}\,,$$

and $G^n_N=\Gamma_1+N+n, F^n_N=-2\Gamma_1-2N-n+1$. The notation that we used in the above formulas is not accidental. Thus the arrows correspond to the action of the operators that increases or decreases values of m and/or M. Those are the indices that are implemented by the Poincaré algebra. Hence their interpretation is given below:

$\downarrow$ - $m\to m-1$,

$\uparrow$ - $M\to M+1$,

$\uparrow\uparrow$ -$m\to m+1$, $\quad M\to M+1$.

One has to add one more condition on Eqs. (12)–(21). Namely, in the right-hand sides of all equations we define

$$Y^{n1}_{0M}=Y^{n0}_{0M}\,. \tag{22}$$

Let us recall that the basis Y^{nm}_{NM} has been defined for $n,m,N,M\in\mathbb{N}$ and $m\leqslant\min(N,M)$. The artificial condition just given can be treated as a "boundary condition" that enables us to write all the equations in a closed form. The way in which the matrix elements in Eqs. (12)–(21) are factored follows from the embedding chain $so(3)\subset so(3,1)\subset iso(3,1)$.

Let us discuss now the most important features of the modules $\mathcal{U}_+(\Gamma)$ exhibited by Eqs. (12)–(21).

(A) Define $\mathcal{U}^m_M=\{Y^{nm}_{NM},\ n,N\geqslant 0,\ N\geqslant m\}$, where $M\geqslant m$. Then each $\mathcal{U}^m_M$ is a Verma $\mathcal{L}$ module that is either an inf-in-mod or an inf-ir-mod depending on the value of Γ. Let us treat $\mathcal{U}^0_0$ as a prototypical $\mathcal{L}$ module. Here $\mathcal{U}^0_0$ is the module that would be obtained if one considered the enveloping algebra of $\mathcal{L}$. It is essentially the same module that was considered[5] in the case of the Lorentz algebra. Then, $\mathcal{U}^m_M$ becomes $\mathcal{U}^0_0$ defined in terms of starred objects:

$$\Gamma^*_1=\Gamma_1+m,\quad \Gamma^*_2=\Gamma_2+M-m,\quad N^*=N-m\,,$$

$$Y^{*nm}_{NM}=\begin{cases}Y^{nm}_{NM}, & \text{if } N=0,\\ \dfrac{1}{(N-m+1)\cdots(N-1)N}\,Y^{nm}_{NM}, & \text{if } N\neq 0.\end{cases}$$

(B) For each $\Gamma\in\mathbb{C}^2$, $\mathscr{U}_+(\Gamma)$ is indecomposable since the value of M never decreases. It is quite clear from Fig. 1 how submodules, quotient modules, and subquotient modules can be defined.

(C) The sum $\mathscr{U}_0^0+\mathscr{U}_1^1+\mathscr{U}_2^2\cdots$ is a quotient module and is essentially the same module as $\Omega_+(\Gamma)$ recently[4] examined. Thus the study conducted previously[4,5] is a special (and very illuminating) case of the results contained in this paper.

(D) Equations (12)–(21) can be extended to all $n\in\mathbb{Z}$. The extended vector spaces and modules will be denoted with a tilde, i.e., $\tilde{\mathscr{U}}_+(\Gamma)$, $\tilde{\mathscr{U}}_M^m$, etc.

(E) If $2\Gamma\in\mathbb{Z}^2$ one may obtain fin-in-mods (or fin-ir-mods when trivial on $\mathscr{K}$) of $\mathscr{P}$ by passing to the quotients. The essential features of the procedure are the same as previously[4] studied. The fin-in-mods obtained on quotients of the composition series discussed in (C) coincide with fin-in-mods therein[4] contained. New fin-in-mods can be obtained as other quotient modules (see Sec. VI).

(F) The module $\mathscr{U}_+(\Gamma)$ goes over into the module $\mathscr{U}_-(-\Gamma)$ under the Lie algebra automorphism a:

$$a(h_3)=-h_3,\quad a(p_3)=-p_3,\quad a(h_+)=h_-,$$
$$a(h_-)=h_+,\quad a(p_+)=p_-,$$
$$a(p_-)=p_+,\quad a(k_+)=k_-,\quad a(k_-)=k_+,$$
$$a(r_-)=-r_+,\quad a(r_+)=-r_-.$$

(G) It can be seen that the indecomposability of $\mathscr{U}_+(\Gamma)$ essentially follows from the fact that $\mathscr{P}$ is a semidirect product of $\mathscr{K}$ and $\mathscr{L}$. Hence the situation is quite different than in the case of $\mathscr{L}$ itself where passing to the quotients gives[5] fin-ir-mods. This observation might be of a more general nature and apply to analogous modules of other non-semisimple Lie algebras that are semidirect products of an Abelian ideal and a Levi subalgebra.

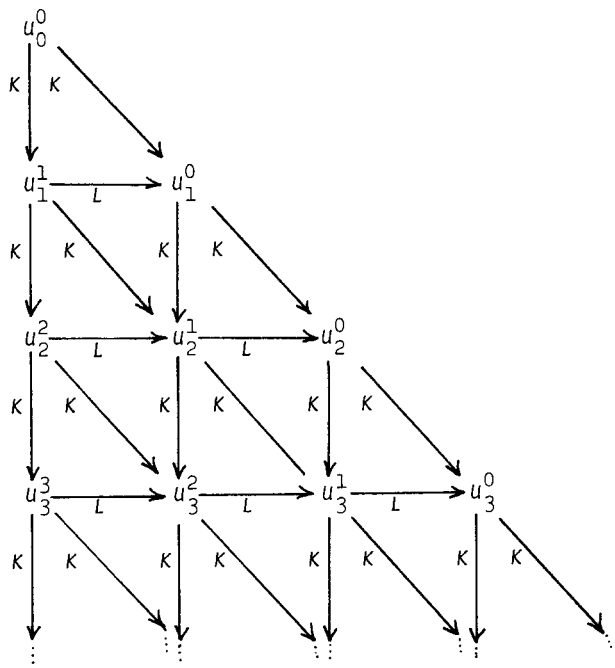

FIG. 1. The module $\mathscr{U}_+(\Gamma)$. The action between subquotient modules $\mathscr{U}_M^m$ is carried by $\mathscr{K}$ or $\mathscr{L}$ as shown above.

VI. FINITE-DIMENSIONAL INDECOMPOSABLE MODULES

In this section we are going to discuss fin-in-mods of $\mathscr{P}$. They are obtained as quotient modules of $\mathscr{U}_+(\Gamma)$ [$\tilde{\mathscr{U}}_+(\Gamma)$]. One obtains infinitely many such modules. We are going to content ourselves with giving only a few examples. A more systematic treatment goes beyond the scope of this paper.

In general, fin-in-mods are obtained when $2\Gamma\in\mathbb{Z}^2$. Then some of the coefficients in Eqs. (12)–(21) vanish giving rise to submodules and hence to quotient modules (finite dimensional). Thus one has to examine the zeros of α, $\alpha^{\uparrow}$, $\alpha^{\downarrow}$, $\alpha^{\uparrow\uparrow}$, β, $\beta^{\uparrow}$, $\beta^{\downarrow}$, $\beta^{\uparrow\uparrow}$, and F_N^n, G_N^n as well as the singularities of the former ones for specific values of Γ. Clearly, singularities cannot occur on the quotient modules if one wants to obtain meaningful results. The nature of separation of submodules from their complements (that is, what is worth mentioning, hidden in the PBW basis) is for each $\mathscr{U}_M^m$ ($\tilde{\mathscr{U}}_M^m$) the same and follows the pattern closely examined[5] in the case of the Lorentz algebra (or, simply, $\mathscr{U}_0^0$). The procedure is somewhat geometrical although it is conceivable to look at Eqs. (12)–(21) from the algebraic point of view. Thus if we put the basis of $\mathscr{U}_M^m$ on a two-dimensional plane with N on the horizontal axis and n on the vertical axis we obtain a subset of a $\mathbb{Z}^2$ lattice. Two kinds of separation lines can be introduced: broken defined by the equation $F_N^n=-2\Gamma_1-2N-n+1$ and horizontal given by $N^+=\Gamma_2+M-m-\Gamma_1+1$, $N^-=-\Gamma_2-M+m-\Gamma_1+1$. One can see that the action of all operators of $\mathscr{P}$ is within each $\mathscr{U}_M^m$ the same and if it amounts to an interaction with an adjacent $\mathscr{U}_{M+1}^{m+1}$, $\mathscr{U}_{M+1}^m$, or $\mathscr{U}_M^{m-1}$ then it is only modified by a translation [in other words, under projection onto the canonical image of $\mathscr{U}_+(\mathscr{L})$ in each $\mathscr{U}_M^m$ it is identical for all $\mathscr{U}_M^m$].

Let us now present several examples of fin-in-mods obtained as quotient modules of $\mathscr{U}_+(\Gamma)$ [$\tilde{\mathscr{U}}_+(\Gamma)$]. We are going to characterize them by giving the value of Γ, number of interacting $\mathscr{L}$ modules (those are $\mathscr{L}$ fin-ir-mods), their $\mathscr{R}\subset\mathscr{L}$ decomposition as well as their bases in the explicit form.

(A) A four-dimensional $\mathscr{P}$-fin-in-mod is induced on a quotient of $\mathscr{U}_0^0\cup\mathscr{U}_1^0$ when $\Gamma=(-\frac{1}{2},-\frac{1}{2})$. The number of interacting $\mathscr{L}$-fin-in-mods is 2 and the $\mathscr{R}\subset\mathscr{L}$ decomposition is $2+2$ (two fin-ir-mods, both two-dimensional). The basis is given by $\{Y_{00}^{00},Y_{00}^{10}\}\cup\{Y_{01}^{00},Y_{01}^{10}\}$ (we separate the bases of interacting modules). The matrix representation is of the block form:

$$\begin{pmatrix}\sigma_1 & 0\\ \xi_{12} & \sigma_2\end{pmatrix},$$

where

$$\sigma_1(h_3)=\begin{pmatrix}-\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{pmatrix},\quad \sigma_1(h_+)=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix},$$
$$\sigma_1(h_-)=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\quad \sigma_1(f_+)=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix},$$

$$\sigma_1(f_-)=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\quad \sigma_1(f_3)=\begin{pmatrix}-\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{pmatrix},$$

$$\sigma_2(h_3)=\begin{pmatrix}-\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{pmatrix},\quad \sigma_2(h_+)=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix},$$

$$\sigma_2(h_-)=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\quad \sigma_2(f_+)=\begin{pmatrix}0 & 0\\ -1 & 0\end{pmatrix},$$

$$\sigma_2(f_-)=\begin{pmatrix}0 & -1\\ 0 & 0\end{pmatrix},\quad \sigma_2(f_3)=\begin{pmatrix}\frac{1}{2} & 0\\ 0 & -\frac{1}{2}\end{pmatrix},$$

$$\xi_{12}(k_+)=\begin{pmatrix}0 & 0\\ -1 & 0\end{pmatrix},\quad \xi_{12}(k_-)=\begin{pmatrix}0 & -1\\ 0 & 0\end{pmatrix},$$

$$\xi_{12}(r_+)=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix},\quad \xi_{12}(r_-)=\begin{pmatrix}0 & 0\\ 0 & -1\end{pmatrix}.$$

(B) A five-dimensional $\mathscr{P}$-fin-in-mod is induced on a quotient of $\mathscr{U}_0^0\cup\mathscr{U}_1^1$ when $\Gamma=(-1,0)$. The number of interacting $\mathscr{L}$-fin-ir-mods is two and the $\mathscr{R}\subset\mathscr{L}$ decomposition is $4+1=(3+1)+1$. The basis is given by $\{Y_{00}^{00},Y_{00}^{10},Y_{00}^{20},Y_{10}^{00}\}\cup\{Y_{11}^{01}\}$. The matrix representation can be obtained in a straightforward manner.

(C) A seven-dimensional $\mathscr{P}$-fin-in-mod is induced on a quotient of $\tilde{\mathscr{U}}_0^0\cup\tilde{\mathscr{U}}_1^1$ when $\Gamma=(1,2)$. The number of interacting $\mathscr{L}$-fin-ir-mods is two and the $\mathscr{R}\subset\mathscr{L}$ decomposition is $4+3=(1+3)+3$. The basis is given by $\{\tilde{Y}_{00}^{-10},\tilde{Y}_{10}^{-10},\tilde{Y}_{10}^{-20},\tilde{Y}_{10}^{-30},\cup\{\tilde{Y}_{11}^{-11},\tilde{Y}_{11}^{-21},\tilde{Y}_{11}^{-31}\}$.

(D) An eight-dimensional $\mathscr{P}$-fin-in-mod is induced on a quotient of $\mathscr{U}_0^0\cup\mathscr{U}_1^1$ when $\Gamma=(-\frac{3}{2},-\frac{1}{2})$. The number of interacting $\mathscr{L}$-fin-ir-mods is 2 and the $\mathscr{R}\subset\mathscr{L}$ decomposition is $6+2=(4+2)+2$. The basis is given by $\{Y_{00}^{00},Y_{00}^{10},Y_{00}^{20},Y_{00}^{30},Y_{10}^{00},Y_{10}^{10}\}\cup\{Y_{11}^{01},Y_{11}^{11}\}$.

(E) An eight dimensional $\mathscr{P}$-fin-in-mod is induced on a quotient of $\mathscr{U}_0^0\cup\mathscr{U}_0^0\cup\mathscr{U}_1^1$ when $\Gamma=(-1,0)$. The number of interacting $\mathscr{L}$-fin-ir-mods is 3 and the $\mathscr{R}\subset\mathscr{L}$ decomposition is $4+3+1=(3+1)+3+1$. The basis is given by $\{Y_{00}^{00},Y_{00}^{10},Y_{00}^{20},Y_{10}^{00}\}\cup\{Y_{01}^{00},Y_{01}^{10},Y_{01}^{20}\}\cup\{Y_{11}^{00}\}$.

The matrix representation of the above $\mathscr{P}$-fin-in-mods is of the following block form:

$$\begin{pmatrix}\sigma_1 & 0\\ \xi_{12} & \sigma_2\end{pmatrix}$$

in the case of two-interacting $\mathscr{L}$-fin-ir-mods (A,B,C,D), or

$$\begin{pmatrix}\sigma_1 & 0 & 0\\ \xi_{12} & \sigma_2 & 0\\ \xi_{13} & 0 & \sigma_3\end{pmatrix}$$

in the case of three interacting $\mathscr{L}$-fin-ir-mods (E). Other $\mathscr{P}$-fin-in-mods can be obtained in the manner outlined above. It is interesting to note that from among all the $\mathscr{P}$-fin-in-mods up to dimension 8, that were classified by Paneitz,[7] all but one have been identified.

VII. CONCLUDING REMARKS

We derived the formulas for the induced modules $\mathscr{U}_+(\Gamma)$ of the Poincaré algebra $\mathscr{P}$ in an energy-cyclic AMB. It turns out that $\mathscr{U}_+(\Gamma)$ is indecomposable for each $\Gamma\in\mathbb{C}^2$. The subquotient modules of $\mathscr{U}_+(\Gamma)$ are Verma $\mathscr{L}$ modules. A large family of new $\mathscr{P}$-fin-in-mods can be obtained on quotients of $\mathscr{U}_+(\Gamma)$ when $2\Gamma\in\mathbb{Z}^2$. Several examples were given.

It would be of interest to determine whether other energy-cyclic AMB's give more suitable bases for physical applications. The corresponding formulas could be easily obtained through a linear transformation. Also, a more systematic approach to $\mathscr{P}$-fin-in-mods can be attempted.

The extensions of $\mathscr{U}_+(\Gamma)$ to the scale-invariant algebra ($\mathscr{P}$ extended by dilations) can be obtained without much effort. However, it would be interesting to study the extension formulas for the conformal algebra.

All of the above matters are currently under investigation.

[1] A. O. Barut and R. Raczka, *Theory of Group Representations and Applications* (Polish Scientific, Warsaw, 1977).
[2] L. Hlavaty and J. Niederle, Czech. J. Phys. B t.**29** (3), 283 (1979).
[3] P. A. M. Dirac, Int. J. Theor. Phys. **23** (8), 677 (1984).
[4] R. Lenczewski and B. Gruber, J. Phys. A **19**, 1 (1986).
[5] B. Gruber and R. Lenczewski, J. Phys. A **16**, 3703 (1983).
[6] D. N. Verma, Bull. Am. Math. Soc. **74**, 160 (1968).
[7] S. M. Paneitz, Ann. Inst. H. Poincaré **40** (1), 35 (1984).
[8] I. M. Gel'fand, R. A. Minlos, and Z. Y. Shapiro, *Representations of the Rotation and Lorentz Groups and Their Applications* (Pergamon, Oxford, 1963).
[9] M. A. Naimark, *Linear Representations of the Lorentz Group* (Pergamon, Oxford, 1964).
[10] J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory* (Springer, New York, 1972).
[11] J. Dixmier, *Enveloping Algebras* (North-Holland, Amsterdam, 1977).

Classes of exactly solvable nonlinear evolution equations for Grassmann variables: The normal form method

Christian Elphick
Laboratoire de Physique Théorique,[a] *Université de Nice, Parc Valrose, 06034 Nice Cedex, France*

(Received 4 March 1986; accepted for publication 4 February 1987)

A systematic procedure is presented to solve analytically differential equations for Grassmann variables with the most general nonlinearity. The method consists in the reduction of the original equation to its simplest form (normal form). The classes of solvable normal forms are determined only by the structure of the linear part of the original equation and are parametrized in terms of the number of critical eigenvalues.

I. INTRODUCTION

Differential equations involving anticommuting (Grassmann) variables (GDE) are often encountered in theoretical physics[1–4] due to the fact that the Grassmann variables θ_i, $[\theta_i,\theta_j]_+ = 0$ are the classical counterparts ($\hbar \to 0$) of quantum fermionic operators a_i, $a_i^\dagger$, $[a_i,a_j]_+ = 0$, $[a_i,a_j^\dagger]_+ = \hbar\delta_{ij}$. For instance, they have been considered by Casalbuoni[4] in the study of the classical mechanics for a Bose–Fermi system; by Berezin and Marinov[1] in the context of the supersymmetric treatment of a classical relativistic free particle which, after quantization, becomes a free Dirac particle; and by Olshanetsky,[5] who considered the supersymmetric version of integrable models by the inverse scattering method. Let us note that GDE also naturally arise when we study a quantum Fermi system via the path integral in the Bargmann–Fock coherent state representation. In this formulation the integral kernel of the evolution operator is given by an integral over a Grassmann algebra whose explicit calculation leads to the evaluation of the action on the classical trajectory satisfying the Grassmann Hamiltonian equations (see Appendix C).

As another example of GDE we mention the widely studied massive Thirring model[6] defined by

$$i\frac{\partial}{\partial x}\phi_1 = m\phi_2 + g\bar{\phi}_2\phi_2\phi_1\,, \tag{1.1}$$

$$i\frac{\partial}{\partial t}\phi_2 = m\phi_1 + g\bar{\phi}_1\phi_1\phi_2\,, \tag{1.2}$$

where ϕ_1, ϕ_2 are Grassmann fields. This model possesses an infinite number of conserved quantities[7] and presents soliton-type behavior. Moreover, Morris[8] has been able to determine a Bäcklund transformation for the anticommuting massive Thirring model by generalizing the prolongation structure method of Wahlquist and Estabrook[9] to Grassmann algebra valued differential forms.

It is the purpose of this paper to study the most general GDE in the sense that they are not necessarily derived from the variation of an action (as in Refs. 1 and 4) and they contain the most general nonlinearity compatible with the Grassmann algebra.

The plan of this paper is as follows: In Sec. II the differential equations to be studied are presented, together with some notation to be used in the remaining sections. For the sake of clarity I will consider here the case of a discrete number of degrees of freedom (eventually a countable infinity of degrees of freedom), with the generalization to the continuous case (Grassmann fields) being straightforward. Section III is devoted to developing the normal form method for the equations of Sec. II, that is, a method that reduces these equations to their simplest form. Finally, in Sec. IV we comment on other classes of normal forms that can be easily integrated and on the relation of Thirring-type equations with our normal form method.

II. NONLINEAR GRASSMANN DIFFERENTIAL EQUATIONS

To begin with let $\mathcal{G}$ be an infinite-dimensional complex Grassmann algebra. Let us consider in $\mathcal{G}$ a family $\mathcal{F}$ of n real odd elements depending on a real parameter t:

$$\mathcal{F} = \{\theta_i(t);\theta_i(t) = \theta_i^*(t),\ [\theta_i(t),\theta_j(t)]_+ = 0,\ \forall t\in[t_0,+\infty),\ i,j = 1,\ldots,n\}\,,$$

where the asterisk denotes the involution operation[10] in $\mathcal{G}$ and $[\ ,\ ]_+$ is the anticommutator.

We assume that the elements of $\mathcal{F}$ vary with t according to

$$\frac{\partial}{\partial t}\boldsymbol{\theta} = L\,\boldsymbol{\theta} + \mathbf{N}(\boldsymbol{\theta})\,, \tag{2.1}$$

$$\boldsymbol{\theta}(t = t_0) = \boldsymbol{\theta}_0\,, \tag{2.2}$$

where $\boldsymbol{\theta}$ is an n-dimensional vector, with $\boldsymbol{\theta} = \theta_i\mathbf{e}^i$ (sum over repeated indices), $\mathbf{e}^i$, $i = 1,\ldots,n$ the canonical basis in a vectorial space that we call $\mathcal{H}_1$. Here, L is a real linear operator acting on $\mathcal{H}_1$ and $\mathbf{N}(\boldsymbol{\theta})$ stands for an arbitrary nonlinearity which preserves the odd character of (2.1), and therefore can be written as

$$\sum_{j=1}^{(1/2)(m-1)} \boldsymbol{\varphi}_i^{(2j+1)}\,\epsilon_{i;i_1i_2\cdots i_{2j+1}}\theta_{i_1}\theta_{i_2}\cdots\theta_{i_{2j+1}}\,, \tag{2.3}$$

where $\boldsymbol{\varphi}_i^{(2j+1)}$, $i = 1,\ldots,n$; $j = 1,\ldots,\frac{1}{2}(m-1)$ belong to $\mathcal{H}_1$, $m = n$ (resp. $n-1$) if n is odd (resp. even), and $\epsilon_{i;i_1i_2\cdots i_{2j+1}}$ is totally antisymmetric in the indices i_k, $k = 1,\ldots,2j+1$. Note here that if $n = 2$, there is no nonlinear term since the only term available, $\theta_1\theta_2$, is an even element of $\mathcal{G}$ which is not in $\mathcal{F}$.

Let us remark here that the Thirring model (1.1) and

[a] Unité Associée au C.N.R.S.

 0022-2488/87/061243-07$02.50

(1.2) can be cast into the form (2.1) by writing $\phi_1 = \varphi_1 + i\varphi_2$, $\phi_2 = \psi_1 + i\psi_2$, where $\varphi_1, \varphi_2, \psi_1, \psi_2$ are new independent real Grassmann fields, and by assuming that these fields only depend on a single variable $\xi = \lambda x - \lambda^{-1}t$, $\lambda \in \mathbb{R}$ (Grassmann solitons).

III. THE NORMAL FORM METHOD

Let us assume that the spectrum S of L is the union of two disjoint parts $S = S_- \cup S_0$, where $\gamma \in S_-$ (resp. S_0) if $\mathrm{Re}\,\gamma < 0$ (resp. $\mathrm{Re}\,\gamma = 0$). We suppose that the linear operator L has n_c eigenvalues belonging to S_0 which we call critical [since they will determine the solvable classes of (2.1) by this method], while the remaining eigenvalues γ_α, $\alpha = 1,...,M = n - n_c$ (M eventually infinite) are all different and have negative real parts. Since L is a real linear operator the critical eigenvalues can be zero on purely imaginary pairs $(i\omega_k, -i\omega_k)$, $k = 1,...,s$. Then $n_c = 2s + l$, where l is the algebraic multiplicity of the zero eigenvalue.

In the following we will consider the class of nonlinear equations parametrized by $s = 1$, $l = 1$ and it will be proved that it can be exactly solved. (Section IV and Appendices B and C deal with other classes that can also be treated by the method developed in this section.)

The method is as follows. According to our assumptions $\mathscr{H}_1$ is the direct sum of two subspaces invariant by L, $\mathscr{H}_1 = \mathscr{H}_1^c \oplus \mathscr{H}_1^s$, where $\mathscr{H}_1^s$ is spanned by the vectors ψ_α, $L\psi_\alpha = \gamma_\alpha \psi_\alpha$ and $\mathscr{H}_1^c$ is the critical space spanned by $\{\phi_1, \phi_2 = \phi, \phi_3 = \bar{\phi}\}$ such that

$$L\phi_i = \mathbb{J}_{ji}\phi_j\,, \quad \mathbb{J} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & i\omega & 0 \\ 0 & 0 & -i\omega \end{pmatrix}. \tag{3.1}$$

Let us introduce into $\mathscr{G}$ a new family of odd variables $\{A_1, A_2 = A, A_3 = A^*, B_\alpha, \alpha = 1,...,M\}$ defined through the nonlinear change of variables

$$\theta(t) = \sum_{i=1}^{3} A_i(t)\phi_i + \sum_{\alpha=1}^{M} B_\alpha(t)\psi_\alpha + \sum_{j=1}^{(m-1)/2} \theta^{[2j+1]}(C_p(t))\,, \tag{3.2}$$

where $m = n$ for odd n, $m = n - 1$ for even n, and $\theta^{[2j+1]}$ is homogeneous of degree $2j + 1$ in the variables $\{C_p\}$, where $C_p = A_p$, $p = 1, 2, 3$, $C_{3+\alpha} = B_\alpha$, $\alpha \geqslant 1$, and therefore is of the form

$$\theta^{[2j+1]} = \sum (C_{i_1})^{r_1}(C_{i_2})^{r_2}\cdots(C_{i_p})^{r_p}\theta^{[2j+1]r_1r_2\cdots r_p}_{i_1i_2\cdots i_p}\,, \tag{3.3}$$

with

$$\sum_{i=1}^{p} r_i = 2j + 1, \quad r_i \in \{0,1\}, \quad i_j \neq i_k, \quad \text{if } j \neq k\,, \tag{3.4}$$

and $\theta^{[2j+1]r_1\cdots r_p}_{i_1\cdots i_p} \in \mathscr{H}_1$ totally antisymmetric in $i_1\cdots i_p$ and time independent.

We look for equations for $\{A_i, B_\alpha\}$ of the form

$$\frac{\partial A_i}{\partial t} = \mathbb{J}_{ij}A_j + f_i^{[3]} + f_i^{[5]} + \cdots + f_i^{[m]}\,,$$

$$\frac{\partial}{\partial t}B_\alpha = \gamma_\alpha B_\alpha + g_\alpha^{[3]} + g_\alpha^{[5]} + \cdots + g_\alpha^{[m]}\,, \quad \text{(no sum over } \alpha\text{)}, \tag{3.5}$$

where $f_i^{[r]}, g_\alpha^{[r]}$ are homogeneous of degree r in $\{C_p\}$.

From (3.2), we have

$$\frac{\partial \theta}{\partial t} = \frac{\partial C_p}{\partial t}\frac{\partial^l}{\partial C_p}\theta\,, \tag{3.6}$$

where the superscript l means that in differentiating with respect to C_p we must displace the variable C_p to the left before dropping it.

By replacing (3.2) in (2.1) and using (3.5) and (3.6), we obtain, after an identification of each order in $\{C_p\}$, a hierarchy of equations for $\theta^{[r]}$, $r = 3,5,...,m$. For $r = 1$, we obtain

$$\mathbb{J}_{ij}A_j\phi_i + \gamma_\beta B_\beta \psi_\beta = L(A_i\phi_i + B_\beta\psi_\beta)\,, \tag{3.7}$$

which is an identity due to (3.1) and $L\psi_\beta = \gamma_\beta\psi_\beta$.

At order $r \geqslant 3$ (odd r), we obtain the following homological equation[11]:

$$\mathscr{L}\theta^{[r]} = (\mathscr{A} + \mathscr{B} - L)\theta^{[r]} = \mathbf{I}^{[r]} - f_i^{[r]}\phi_i - g_\alpha^{[r]}\psi_\alpha \equiv \mathbf{K}^{[r]}\,, \quad r = 3,5,...,m, \tag{3.8}$$

where

$$\mathscr{A} = \mathbb{J}_{ij}A_j\frac{\partial^l}{\partial A_i}\,, \quad \mathscr{B} = \gamma_\alpha B_\alpha \frac{\partial^l}{\partial B_\alpha} \tag{3.9}$$

and

$$\mathbf{I}^{[r]} = \mathbf{N}^{[r]}(\theta) - \sum_{s=3}^{r-2}\left(\frac{\partial}{\partial t}C_p\right)^{[s]}\frac{\partial^l}{\partial C_p}\theta^{[r-s+1]}\,. \tag{3.10}$$

If $r = 3$ the last term in (3.10) is absent.

Let us look now at the structure of the space in which (3.8) is written: It is the tensor product $\mathscr{H} = \mathscr{H}_1 \otimes \mathscr{H}_2 \otimes \mathscr{H}_3$, where $\mathscr{H}_2$ (resp. $\mathscr{H}_3$) is generated by monomials constructed from the variables A_i (resp. B_α). Therefore $\mathscr{H}$ is an orthogonal sum of $\mathscr{L}$-invariant subspaces.

$$\mathscr{H} = \bigoplus_{r=3,5,...,m}\ \bigoplus_{N=0}^{N=r} \mathscr{H}_3^{[r-N]} \otimes \mathscr{H}_2^{[N]} \otimes (\mathscr{H}_1^c \oplus \mathscr{H}_1^s)\,, \tag{3.11}$$

where $\mathscr{H}_2^{[N]}$ is generated by $A_1^{r_1}A_2^{r_2}A_3^{r_3}$, $\Sigma_{i=1}^3 r_i = N \leqslant 3$, $r_i \in \{0,1\}$, if $N > 3$ $\mathscr{H}_2^{[N]}$ is empty, and $\mathscr{H}_3^{[r-N]}$ is generated by $B_1^{r_1}\cdots B_M^{r_M}$, $\Sigma_{i=1}^M r_i = r - N$, $r_i \in \{0,1\}$.

We now introduce a scalar product in $\mathscr{H}_2$, $\mathscr{H}_3$ defined as

$$\langle f_1, f_2\rangle = \int \exp\left(-\sum_{i=p''}^{p'} C_i^* C_i\right) f_1(C_j) f_2^*(C_j) \times dC_{p'}^*\, dC_{p'}\cdots dC_{p''}^*\, dC_{p''}\,, \tag{3.12}$$

where $p' = 3$, $p'' = 1$ for $\mathscr{H}_2$, and $p' = M$, $p'' = 4$ for $\mathscr{H}_3$ [note that (3.12) is a multiple Grassmann integral defined in the Berezin[10] sense]. We define the scalar product in $\mathscr{H}$ as

$$\langle\,,\rangle_{\mathscr{H}} = \langle\,,\rangle_{\mathscr{H}_1}\langle\,,\rangle_{\mathscr{H}_2}\langle\,,\rangle_{\mathscr{H}_3}\,, \tag{3.13}$$

where $\langle\,,\rangle_{\mathscr{H}_1}$ is such that the eigenbasis is orthonormal.

It is easy to see now that $\mathscr{L}$ is a linear noninvertible operator acting on $\mathscr{H}$ and therefore (3.8) will have no solutions for $\theta^{[r]}$ unless $\mathbf{K}^{[r]}$ is orthogonal to the null space of the adjoint of $\mathscr{L}$ in the scalar product (3.13). This solvability condition will determine the unknown functions $f_i^{[r]}, g_\alpha^{[r]}$ in (3.5) in such a way that Eq. (3.8) can be solved for $\theta^{[r]}$. We note that $f_i^{[r]}, g_\alpha^{[r]}$ (resp. $\theta^{[r]}$) are determined modulo terms in Ran $\mathscr{L}$ (resp. Ker $\mathscr{L}$). We choose to make a minimal prescription such that all these gauge terms are taken to be zero. [Obviously the original solution $\theta(t,t_0)$ of (2.1) will not depend on this choice.]

At this point it is worthwhile remarking that the operators $A_i, \partial^l/\partial A_i$ (resp. $B_\alpha, \partial^l/\partial B_\alpha$) are adjoint to each other in the scalar product (3.12) and that they satisfy ($a_i^\dagger = A_i, a_i = \partial^l/\partial A_i, b_\alpha^\dagger = B_\alpha, b_\alpha = \partial^l/\partial B_\alpha$)

$$[a_i, a_j^\dagger]_+ = \delta_{ij}, \quad [b_\alpha, b_\beta^\dagger]_+ = \delta_{\alpha\beta}, \tag{3.14}$$

i.e., they are Fermionic creation and annihilation operators in the Fock space $\mathscr{H}_2$ (resp. $\mathscr{H}_3$) [Eqs. (3.14) show that $\mathscr{L}$ is the Fermionic version of the homological operator introduced by Arnol'd[11]] and, for instance, we can write the generators of $\mathscr{H}_2^{[N]}, \mathscr{H}_3^{[r-N]}$ as

$$\begin{aligned}
&|r_1,r_2,r_3\rangle = (a_1^\dagger)^{r_1}(a_2^\dagger)^{r_2}(a_3^\dagger)^{r_3}|0\rangle, \quad |0\rangle = 1,\\
&\sum_{i=1}^{3} r_i = N, \quad r_i \in \{0,1\},\\
&|r_1,\ldots,r_M\rangle = (b_1^\dagger)^{r_1}\cdots(b_M^\dagger)^{r_M}|0\rangle,\\
&\sum_{i=1}^{M} r_i = r-N, \quad r_i \in \{0,1\},
\end{aligned} \tag{3.15}$$

which are obviously orthonormal.

Therefore the adjoint of $\mathscr{L}$ in the scalar product (3.12) reads as

$$\mathscr{L}^\dagger = \bar{\mathbb{J}}_{ij} A_i \frac{\partial^l}{\partial A_j} + \bar{\gamma}_\alpha B_\alpha \frac{\partial^l}{\partial B_\alpha} - L^\dagger, \tag{3.16}$$

where $L^\dagger$ is the conjugate transpose of L: $L^\dagger = \bar{L}^t$.

It is not difficult to check that the null space of $\mathscr{L}^\dagger$ in $\mathscr{H}$ is generated by (provided the set $\{\gamma_\alpha\}$ satisfies the nonresonant condition $\gamma_\alpha \neq \Sigma_{\alpha' \in \Delta} \sigma_{\alpha'}$, $\forall \Delta \subset \{-2,-1,1,\ldots,M\}$, $\sigma_{\alpha'} = \{\sigma_{-2} = i\omega, \sigma_{-1} = -i\omega, \gamma_{\alpha'}\}$)

$$\{A_1A_2A_3\phi_1, B_\alpha A_2A_3\psi_\alpha, \ \alpha = 1,\ldots,M\}. \tag{3.17}$$

It is worth noting that although vectors of the form $\mathbf{X}_\alpha = B_\alpha A_1A_2A_3\psi_\alpha$ do satisfy $\mathscr{L}^\dagger \mathbf{X}_\alpha = 0$, they are not included in (3.17) since they do not belong to $\mathscr{H}$ (they are even elements in $\mathscr{G}$).

Therefore we have the following solvability conditions:

$$\begin{aligned}
&\langle f_1^{[r]}, A_1A_2A_3\rangle_{\mathscr{H}_2\otimes\mathscr{H}_3} = \langle \mathbf{I}^{[r]}, A_1A_2A_3\phi_1\rangle_{\mathscr{H}},\\
&\langle g_\alpha^{[r]}, B_\alpha A_2A_3\rangle_{\mathscr{H}_2\otimes\mathscr{H}_3} = \langle \mathbf{I}^{[r]}, B_\alpha A_2A_3\psi_\alpha\rangle_{\mathscr{H}}, \quad \alpha = 1,\ldots,M.
\end{aligned} \tag{3.18}$$

Consequently we conclude that we can take $f_2^{[r]} = f_3^{[r]} = 0$, $f_1^{[r]} = g_\alpha^{[r]} = 0$, for $r > 3$, and

$$\begin{aligned}
&f_1^{[3]} = \langle \mathbf{N}^{[3]}(\theta), A_1A_2A_3\phi_1\rangle_{\mathscr{H}} A_1A_2A_3 = kA_1A_2A_3,\\
&g_\alpha^{[3]} = \langle \mathbf{N}^{[3]}(\theta), B_\alpha A_2A_3\psi_\alpha\rangle_{\mathscr{H}} B_\alpha A_2A_3 = k_\alpha B_\alpha A_2A_3.
\end{aligned} \tag{3.19}$$

We remark that k, k_α in (3.19) are totally determined since $\mathbf{N}^{[3]}(\theta)$ only depends on $\theta^{[1]} = A_i\phi_i + B_\alpha\psi_\alpha$.

Having completely determined the functions $f_i^{[r]}, g_\alpha^{[r]}$ in such a way that (3.8) can be solved for $\theta^{[r]}$, and noting that $\mathbf{I}^{[r]}$ only depends on $\theta^{[s]}$ for $s < r$, we see that (3.8) can be solved by recursion in r. [In Appendix A it is shown how to solve Eq. (3.8).]

Thus we have finally derived the normal form equations for $\{A_i, B_\alpha\}$. By using (3.18) and (3.19) in (3.5), we obtain

$$\begin{aligned}
&\frac{\partial A_1}{\partial t} = kA_1AA^*, \quad \frac{\partial A}{\partial t} = i\omega A, \quad \frac{\partial A^*}{\partial t} = -i\omega A^*,\\
&\frac{\partial B_\alpha}{\partial t} = B_\alpha(\gamma_\alpha + k_\alpha AA^*), \quad \alpha = 1,\ldots,M.
\end{aligned} \tag{3.20}$$

[Note that k, k_α are complex numbers since $A_1^* = A_1$, $B_\alpha^* = B_\alpha$, $(AA^*)^* = -AA^*$ (since A, A^* anticommute).]

Once Eq. (3.8) has been solved for $\theta^{[r]}$, the original vector $\theta(t)$ is obtained through (3.2); if one evaluates this expression at $t = t_0$, one obtains a relation between θ_0 and $\{A_i(t_0), B_\alpha(t_0)\}$, which can be solved for $\{A_i(t_0), B_\alpha(t_0)\}$, determining in this way the initial conditions for (3.20) as functions of the original initial conditions. In short, $\{A_i(t_0), B_\alpha(t_0)\}$ can be completely determined from the knowledge of θ_0.

Equations (3.20) are easily solved and we obtain

$$\begin{aligned}
&A(t) = e^{i\omega(t-t_0)}A(t_0), \quad A^*(t) = e^{-i\omega(t-t_0)}A^*(t_0),\\
&A_1(t) = (1 + k(t-t_0)A(t_0)A^*(t_0))A_1(t_0),\\
&B_\alpha(t) = e^{\gamma_\alpha(t-t_0)}(1 + k_\alpha(t-t_0)A(t_0)A^*(t_0))B_\alpha(t_0).
\end{aligned} \tag{3.21}$$

By replacing (3.21) back into (3.2) we obtain the solution of (2.1) and (2.2). Let us remark that one can formally obtain from (3.20) the case when S_0 is empty by putting $A = A^* = A_1 = 0$; then the original problem is reduced to the linear equations,

$$\frac{\partial B_\alpha}{\partial t} = \gamma_\alpha B_\alpha, \quad \alpha = 1,\ldots,M. \tag{3.22}$$

We finally note that the solvability conditions have a simple interpretation in terms of particle physics: Ker $\mathscr{L}^\dagger$ is generated by vectors $\mathbf{X} = X_1\phi_1 + X_2\phi_2 + X_3\phi_3$ and $\mathbf{Y} = B_\alpha Y_\alpha \psi_\alpha$ (no sum over α) such that $\mathbf{X} \in$ Ker $(\mathscr{A}^\dagger - \mathscr{L}^\dagger)$ and $Y \in$ Ker $\mathscr{A}^\dagger$, which means that they satisfy ($\mathscr{A}^\dagger = -\mathscr{A}$)

$$\begin{aligned}
&\tfrac{1}{2}\mathscr{A}X_1 = 0, \quad \tfrac{1}{2}\mathscr{A}X_2 = \tfrac{1}{2}X_2, \quad \tfrac{1}{2}\mathscr{A}X_3 = -\tfrac{1}{2}X_3,\\
&\tfrac{1}{2}\mathscr{A}Y_\alpha = 0, \quad \alpha = 1,\ldots,M.
\end{aligned} \tag{3.23}$$

By writing $\frac{1}{2}\mathscr{A}$ as

$$\tfrac{1}{2} a_i^\dagger (\lambda_3)_{ij} a_j, \tag{3.24}$$

where λ_3 is the diagonal Gell-Mann matrix diag$(0,1,-1)$, we easily see that $\frac{1}{2}\mathscr{A}$ is one of the diagonal generators of the Cartan subalgebra of color SU(3) [$\frac{1}{2}\mathscr{A} = I_3$ is also one of the generators of the isospin subgroup of SU(3)]. Therefore $\mathbf{X}$ can be regarded as a quark triplet with components (s,u,d). It follows that the solvability conditions project Eq. (3.8) on the $\frac{1}{2}$-isospin plane (u,d).

IV. GENERAL COMMENTS

Let us finally mention that besides the class of equations studied here $[s=1,\ l=1]$ and the trivial case when S_0 is empty, there are other classes that can also be treated and solved by the method presented here in a straightforward way. These are the following:

$$[s=0,\ l=1],\quad \mathbb{J}=(0); \tag{4.1}$$

$$[s=0,\ l=2],\quad \mathbb{J}_1=\begin{pmatrix}0&0\\0&0\end{pmatrix}\quad\text{or}\quad \mathbb{J}_2=\begin{pmatrix}0&1\\0&0\end{pmatrix}; \tag{4.2}$$

$$[s=1,\ l=0],\quad \mathbb{J}_3=\begin{pmatrix}i\omega&0\\0&-i\omega\end{pmatrix}; \tag{4.3}$$

$$[s=0,\ l=3],\quad \mathbb{J}=\begin{pmatrix}0&1&0\\0&0&1\\0&0&0\end{pmatrix}$$

$$\text{or}\quad \mathbb{J}=\begin{pmatrix}\mathbb{J}_1&&0\\&&0\\0&0&0\end{pmatrix}\quad\text{or}\quad \mathbb{J}=\begin{pmatrix}\mathbb{J}_2&&0\\&&0\\0&0&0\end{pmatrix}; \tag{4.4}$$

$$[s=1,\ l=2],\quad \mathbb{J}=\begin{pmatrix}\mathbb{J}_3&\mathbb{J}_1\\\mathbb{J}_1&\mathbb{J}_1\end{pmatrix}\quad\text{or}\quad \mathbb{J}=\begin{pmatrix}\mathbb{J}_3&\mathbb{J}_1\\\mathbb{J}_1&\mathbb{J}_2\end{pmatrix}. \tag{4.5}$$

For other classes with more complicated matrices $\mathbb{J}$, although the method does not give directly the solution to (2.1), it allows us to write (2.1) in its normal form, which means in its simplest form (with the least number of nonlinear terms) the solution can be found either by inspection or by developing new methods such as Bäcklund transformations or inverse scattering methods applied to GDE.

It is also interesting to remark that if we consider the case when

$$\mathbb{J}=\begin{pmatrix}\mathbb{J}_3&\mathbb{J}_1\\\mathbb{J}_1&\mathbb{J}_3\end{pmatrix}$$

and we apply the method presented in Sec. III, we obtain a normal form equation for the critical variables, which is exactly the anticommuting massive Thirring model when written in terms of the variable $\xi=\lambda x-\lambda^{-1}t$, i.e., the normal form equations read as

$$\begin{aligned}
i\frac{\partial}{\partial t}\phi_1&=\omega\phi_2+g\,\phi_2^*\phi_2\phi_1,\\
i\frac{\partial}{\partial t}\phi_2&=\omega\phi_1+g\phi_1^*\phi_1\phi_2,\\
\frac{\partial B_\alpha}{\partial t}&=B_\alpha(\gamma_\alpha+\delta_\alpha(\phi_2^*\phi_2+\phi_1^*\phi_1)\\
&\quad+\beta_\alpha\phi_1^*\phi_1\phi_2^*\phi_2),\quad \alpha=1,\ldots,M.
\end{aligned} \tag{4.6}$$

Therefore we conclude that the Thirring model is a universal equation for all the nonlinear Grassmann differential equations parametrized by $[s=2,\ l=0]$, in the sense that for times $t\gg \mathrm{Sup}_\alpha\ \mathrm{Re}|\gamma_\alpha|^{-1}$, all the dynamics is governed by this model (see Appendix B).

Let us note that rigorously speaking the different classes of GDE should also be parametrized by a third integer c[11] (the codimension in the language of bifurcation theory) which counts the minimum number of parameters needed to unfold the critical situation. For instance, the matrix $\mathbb{J}$ studied here has $c=2$. With the last remark in mind, it is clear that in a zero-parameter class of Eq. (2.1) the generic situation to be found is the case when S_0 is empty.

Finally, let us comment on possible generalizations of this work. One possibility is to extend the present method to a supersymmetric classical system described by the action

$$S=\int_{t_i}^{t_f}dt\,d\theta_1\,d\theta_2(-i\Phi_jD_1D_2\Phi_j-U(\Phi)), \tag{4.7}$$

where D_1, D_2 are supersymmetric covariant derivatives:

$$D_1=-\frac{\partial^{\prime}}{\partial\theta_2}+i\theta_2\frac{\partial}{\partial t},\quad D_2=\frac{\partial^{\prime}}{\partial\theta_1}+i\theta_1\frac{\partial}{\partial t}, \tag{4.8}$$

where $U(\Phi)$ is the superpotential and $\Phi_j, j=1,2,\ldots$, is a set of supervariables,

$$\begin{aligned}
&\Phi_j=\varphi_j(t)+i\bar\theta\psi_j(t)+\tfrac{1}{2}\bar\theta\theta\varphi_j^{\prime}(t),\\
&\bar\theta=\theta^{t}\gamma_0,\quad \gamma_0=\begin{pmatrix}0&1\\-1&0\end{pmatrix},
\end{aligned} \tag{4.9}$$

where φ_j, $\varphi_j^{\prime}$ are commuting variables and ψ_j is a two-component Majorana spinor.

The classical mechanics of our system is given by an action obtained from (4.7) after integration over θ_1, θ_2:

$$S^{\prime}=\int_{t_i}^{t_f}dt(-i\varphi_j\ddot\varphi_j+\psi_j^{\prime}\dot\psi_j+U(\varphi_j,\psi_j)), \tag{4.10}$$

where an overdot stands for the time derivative.

Variation of $S^{\prime}$ leads to nonlinear differential equations coupling commuting and anticommuting variables. In particular, by using the method of Sec. III, we arrive at an equation like (3.8), but where the operator $\mathscr{L}$ is replaced by a super homological operator $\hat{\mathscr{L}}$ which contains both bosonic and fermionic creation and annihilation operators. The structure of Ker $\hat{\mathscr{L}}^{\dagger}$ is much more complicated than in the case considered here and therefore the classification of solvable classes or classes reducible to simple normal forms becomes much more complex (details of the normal form method for a supersymmetric classical mechanics will be given elsewhere).

ACKNOWLEDGMENT

It is a pleasure for the author to thank Professor E. Tirapegui for helpful discussions and much encouragement during the course of this work.

APPENDIX A: SOLUTION OF THE HOMOLOGICAL EQUATION

Here we show how to solve Eq. (3.8) for the Grassmann vector $\theta^{[r]}$, $r=3,5,\ldots,m$. For simplicity we consider the case $M=1$ which will suffice for understanding how the solution mechanism works for $M>1$.

We start by decomposing $\mathbf{N}(\theta)$ in Eq. (2.1) as

$$\mathbf{N}(\theta)=N_1(\theta)\phi_1+N_2(\theta)\phi_2+N_3(\theta)\phi_3+N_4(\theta)\psi, \tag{A1}$$

where

$$N_i(\theta)=k_{i;i_1i_2i_3}\theta_{i_1}\theta_{i_2}\theta_{i_3},\quad i=1,\ldots,4, \tag{A2}$$

and $k_{i;i_1i_2i_3}$ is totally antisymmetric in i_1, i_2, i_3.

For $M=1$, Eq. (3.8) reduces to the equation for $\theta^{[3]}$:

$$\left(i\omega\left(A_2\frac{\partial^l}{\partial A_2}-A_3\frac{\partial^l}{\partial A_3}\right)+\gamma B\frac{\partial^l}{\partial B}-L\right)\theta^{[3]} = \mathbf{N}^{[3]}(\theta)-kA_1A_2A_3\phi_1-k_1BA_2A_3\psi\,, \tag{A3}$$

where use has been made of (3.10) and (3.19).

We note that $\mathbf{N}^{[3]}(\theta)$ only depends on $\theta^{[1]}$, which is known by construction [Eq. (3.2)] and is given by (A1) when we replace θ_i by $\rho_i A_i$, $\rho_i=\langle \mathbf{e}_i,\phi_i\rangle$, $i=1,3$, and θ_4 by $\rho_4 B$, $\rho_4=\langle \mathbf{e}_4,\psi\rangle$. We obtain

$$\mathbf{N}^{[3]}(\theta)=N_1^{[3]}\phi_1+N_2^{[3]}\phi_2+N_3^{[3]}\phi_3+N_4^{[3]}\psi\,, \tag{A4}$$

where

$$N_i^{[3]}=a_iA_1A_2A_3+b_iBA_1A_2+c_iBA_1A_3+d_iBA_2A_3 \tag{A5}$$

and the constants $a_i,\ldots,d_i$ are functions of ρ_i, $k_{i;i_1i_2i_3}$ [$a_1=k$, $d_4=k_1$ by the solvability conditions (3.18)].

Now writing $\theta^{[3]}$ as

$$\theta^{[3]}=\theta_1^{[3]}\phi_1+\theta_2^{[3]}\phi_2+\theta_2^{*[3]}\phi_3+\theta_4^{[3]}\psi \tag{A6}$$

and using (A3), we arrive at the following:

$$\mathscr{D}\theta_1^{[3]}=B(b_1A_1A_2+c_1A_1A_3+d_1A_2A_3)\,, \tag{A7}$$

$$(\mathscr{D}-i\omega)\theta_2^{[3]}=N_2^{[3]}\,, \tag{A8}$$

$$(\mathscr{D}-\gamma)\theta_4^{[3]}=a_4A_1A_2A_3+b_4BA_1A_2+c_4BA_1A_3\,, \tag{A9}$$

where

$$\mathscr{D}=i\omega\left(A_2\frac{\partial^l}{\partial A_2}-A_3\frac{\partial^l}{\partial A_3}\right)+\gamma B\frac{\partial^l}{\partial B}.$$

The solutions of Eqs. (A7)–(A9) are obtained by expanding $\theta_1^{[3]}$, $\theta_2^{[3]}$, $\theta_3^{[3]}$ in terms of monomials of degree 3 in (A_1,A_2,A_3,B) and comparing the coefficients of each monomial in both sides of (A7)–(A9). We obtain

$$\theta_1^{[3]}=B\left(\frac{b_1}{\gamma+i\omega}A_1A_2+\frac{c_1}{\gamma-i\omega}A_1A_3+\frac{d_1}{\gamma}A_2A_3\right), \tag{A10}$$

$$\theta_2^{[3]}=-\frac{a_2}{i\omega}A_1A_2A_3 + B\left(\frac{b_2}{\gamma}A_1A_2+\frac{c_2}{\gamma-2i\omega}A_1A_3+\frac{d_2}{\gamma-i\omega}A_2A_3\right), \tag{A11}$$

$$\theta_4^{[3]}=-(a_4/\gamma)A_1A_2A_3+(B/i\omega)A_1(b_4A_2-c_4A_3)\,. \tag{A12}$$

The above expressions completely determine the nonlinear change of variables (3.2) which, together with (3.21), give the complete solution of Eq. (2.1).

APPENDIX B: THE THIRRING MODEL AS THE NORMAL FORM [$s=2$, $l=0$]

Here we prove that the Thirring model is the normal form for the class parametrized by $[s=2,\,l=0]$.

We assume that $t\gg \mathrm{Sup}_\alpha\ \mathrm{Re}|\gamma_\alpha|^{-1}$, which means that we are in the asymptotic regime, where all the variables B_α, $\alpha=1,\ldots,M$, have relaxed to zero and the dynamics is solely described in terms of the critical variables (A_1,A_1^*,A_2,A_2^*).

In this regime, Eq. (3.8) reads as

$$\mathscr{L}\theta^{[r]}=\left(\mathbb{J}_{ij}A_i\frac{\partial^l}{\partial A_j}-\mathbb{J}\right)\theta^{[r]} = I^{[r]}-f_i^{[r]}\phi_i=\mathbf{K}^{[r]},\quad r=3,5,\ldots m. \tag{B1}$$

To satisfy the solvability condition $\mathbf{K}^{[r]}\in\mathrm{Ran}\ \mathscr{L}$, it is sufficient to choose $f_i^{[r]}\phi_i$ in $\mathrm{Ker}\ \mathscr{L}^\dagger\subset\mathscr{H}_1\otimes\mathscr{H}_2^{[r]}$, or equivalently, $\mathbf{F}=\Sigma_{r=3,5,\ldots m}f_i^{[r]}\phi_i$ in $\mathrm{Ker}\ \mathscr{L}^\dagger\subset\mathscr{H}_1\otimes\mathscr{H}_2$:

$$\left((\mathbb{J}^\dagger)_{ij}A_j\frac{\partial^l}{\partial A_i}-\mathbb{J}^\dagger\right)\mathbf{F}=0\,. \tag{B2}$$

By introducing an auxiliary parameter τ, Eq. (B2) can be cast into the form

$$\frac{d}{d\tau}e^{-\mathbb{J}^\dagger\tau}\mathbf{F}(e^{\mathbb{J}^\dagger\tau}\mathbf{A})=0,\quad \mathbf{A}=A_i\phi_i, \tag{B3}$$

from which we deduce that

$$\mathbf{F}(e^{\mathbb{J}^\dagger\tau}\mathbf{A})=e^{\mathbb{J}^\dagger\tau}\mathbf{F}(\mathbf{A})\,. \tag{B4}$$

For the class $[s=2,\,l=0]$, $\mathbb{J}^\dagger$ is given by

$$\mathbb{J}^\dagger=\begin{pmatrix}-i\omega & 0 & 0 & 0\\ 0 & +i\omega & 0 & 0\\ 0 & 0 & -i\omega & 0\\ 0 & 0 & 0 & i\omega\end{pmatrix}, \tag{B5}$$

and therefore (B4) tells us that $\mathbf{F}$ is invariant under two independent rotations by $\omega\tau$ in the planes (A_1,A_1^*), (A_2,A_2^*). Employing the above invariance and assuming that the original system is T invariant [which means that if $\theta(t)$ is a solution of (2.1) then $\theta^*(-t)$ is also a solution of (2.1)], we obtain

$$F_1=igA_2^*A_2A_1,\quad F_2=F_1^*,\qquad F_3=igA_1^*A_1A_2,\quad F_4=F_3^*, \tag{B6}$$

where g is a real constant. Making now the change of variables

$$\varphi_1=A_1+iA_2^*,\quad \varphi_2=A_1-iA_2^*\,, \tag{B7}$$

we arrive at the following normal form for the class $[s=2,\,l=0]$:

$$\frac{1}{i}\frac{\partial\varphi_1}{\partial t}=\omega\varphi_2+\frac{g}{2}\varphi_2^*\varphi_2\varphi_1\,,\qquad \frac{1}{i}\frac{\partial\varphi_1}{\partial t}=\omega\varphi_1+\frac{g}{2}\varphi_1^*\varphi_1\varphi_2\,, \tag{B8}$$

which is nothing but the Thirring model for Grassmann solitons[8] if we identify ω with the soliton mass and t with an appropriate variable defined in a comoving system with the soliton.

APPENDIX C: SOME PHYSICAL APPLICATIONS

We show here the application of the method presented in Sec. III to some physical examples. Let us consider a quantum Fermi system described by a normal ordered Hamiltonian $H(a_j^\dagger,a_j,t)$. By working in the fermionic coherent state representation of Bargmann–Fock, we obtain that the integral kernel of the evolution operator is given by the

Grassmann integral,

$$U(\theta^*_{j(f)},t_f;\theta^*_{j(i)},t_i) = \int \mathscr{D}(\theta^*_j,\theta_j) \times \exp\left(\frac{1}{2}(\theta^*_{j(f)}\theta_{j(f)} + \theta^*_{j(i)}\theta_{j(i)}) + iS\right), \quad \text{(C1)}$$

where

$$S = \int_{t_i}^{t_f} dt \left(\frac{1}{2i}(\theta^*_j\dot{\theta}_j - \dot{\theta}^*_j\theta_j) - h(\theta^*_j,\theta_j,t)\right) \quad \text{(C2)}$$

and h is the classical value of H when we replace the operators $a_j^\dagger$, a_j by the Grassmann variables θ^*_j, θ_j. Explicit evaluation of the functional integral in (C1) leads to the evaluation of the action S on the extremal trajectory, satisfying the classical equations of motion (Hamilton's equations):

$$\frac{1}{i}\dot{\theta}_j = \frac{\partial^l h}{\partial\theta^*_j}, \quad \text{(C3)}$$

$$\frac{1}{i}\dot{\theta}^*_j = \frac{\partial^l h}{\partial\theta_j}. \quad \text{(C4)}$$

Next, we illustrate the normal form method in three examples involving Hamiltonian GDE.

Example 1: Let us consider the class parametrized by $[s=1, l=0]$. Its Jordan matrix reads as

$$\mathbb{J} = \begin{pmatrix} i\omega & \\ & -i\omega \end{pmatrix}. \quad \text{(C5)}$$

For times $t \gg \sup_\alpha \mathrm{Re}|\gamma_\alpha|^{-1}$ the asymptotic dynamics is described only by two variables A_1, A_1^*. Since the only available nonlinearity is even, we readily conclude that the normal form is linear:

$$\dot{A}_1 = i\omega A_1 \quad \text{(C6)}$$

and corresponds to the classical equation of motion of a particle with spin-$\frac{1}{2}$ and magnetic moment μ submitted to the action of a constant magnetic field B_Z ($\omega = \mu B_Z$). The quantum and classical Hamiltonians are

$$H = B_z S_z = \tfrac{1}{2}\mu B_z(a^\dagger a - aa^\dagger), \quad \text{(C7)}$$

$$h = \mu B_z(A_1^* A_1 - \tfrac{1}{2}). \quad \text{(C8)}$$

Example 2: We consider two deuterons with an isospin–isospin type interaction. The quantum Hamiltonian reads ($\hbar = 1$)

$$H = \omega_1(p_1^\dagger p_1 + n_1^\dagger n_1) + \omega_2(p_2^\dagger p_2 + n_2^\dagger n_2) + g\mathbf{I}_1\cdot\mathbf{I}_2, \quad \text{(C9)}$$

where $p_i^\dagger, p_i$ ($n_i^\dagger, n_i$) are Fermionic creation and annihilation operators of proton (neutron) states and the isospin operators $\mathbf{I}_1$, $\mathbf{I}_2$ are given by

$$\mathbf{I}_1 = a_i^\dagger(\boldsymbol{\sigma}_1)_{ij}a_j, \quad \mathbf{I}_2 = b_i^\dagger(\boldsymbol{\sigma}_2)_{ij}b_j \quad \text{(C10)}$$

($\boldsymbol{\sigma}$ are the Pauli matrices and $a_1 = p_1$, $a_2 = n_1$, $b_1 = p_2$, $b_2 = n_2$). Here, H in (C9) commutes with the charge $Q = p_1^\dagger p_1 + p_2^\dagger p_2$ and the baryon number $B = p_1^\dagger p_1 + p_2^\dagger p_2 + n_1^\dagger n_1 + n_2^\dagger n_2$. We will prove that any term added to H that breaks the Q-B conservation laws is an irrelevant or gauge term since its classical counterpart can be completely eliminated from the classical equations of motion (C3) and (C4) and therefore the evolution operator (C1) will remain unaffected by the nonphysical term.

For simplicity we consider the Hamiltonian $H + H'$, where

$$H' = \lambda(p_1^\dagger p_1 p_2 n_2 + n_2^\dagger p_2^\dagger p_1^\dagger n_1^\dagger). \quad \text{(C11)}$$

The classical Hamiltonian equations read as ($p_1 \to \theta_1$, $n_1 \to \theta_2$, $p_2 \to \eta_1$, $n_2 \to \eta_2$)

$$\begin{aligned}
(1/i)\dot{\theta}_1 &= \omega_1\theta_1 + g(2\theta_2\eta_2^*\eta_1 + \theta_1(\eta_1^*\eta_1 - \eta_2^*\eta_2)) \\
&\quad + \lambda\theta_1(\eta_1\eta_2 + \eta_2^*\eta_1^*), \\
(1/i)\dot{\theta}_2 &= \omega_1\theta_2 + g(2\theta_1\eta_1^*\eta_2 + \theta_2(\eta_2^*\eta_2 - \eta_1^*\eta_1)), \quad \text{(C12)} \\
(1/i)\dot{\eta}_1 &= \omega_2\eta_1 + g(2\eta_2\theta_2^*\theta_1 + \eta_1(\theta_1^*\theta_1 - \theta_2^*\theta_2)) \\
&\quad - \lambda\theta_1^*\theta_1\eta_2^*, \\
(1/i)\dot{\eta}_2 &= \omega_2\eta_2 + g(2\eta_1\theta_1^*\theta_2 + \eta_2(\theta_2^*\theta_2 - \theta_1^*\theta_1)) \\
&\quad + \lambda\theta_1^*\theta_1\eta_1^*.
\end{aligned}$$

We note that Eqs. (C12) correspond to the class $[s=4, l=0]$ with Jordan matrix $\mathbb{J} = \mathrm{diag}(\Omega_1,\Omega_1,\Omega_2,\Omega_2)$, with $\Omega_j = \mathrm{diag}(i\omega_j, -i\omega_j)$, $j = 1,2$. Following Sec. III and using the invariance property (B4) we conclude that there exists a nonlinear change of variables $(\theta_i,\eta_i) \to (A_i)$ such that the equations of motion in the new variables contain only nonlinear terms equivariant under the one-parameter Lie group generated by $\mathbb{J}^\dagger$: $e^{-\mathbb{J}^\dagger\tau}$. We easily see from (C12) that the nonlinear terms with g in factor respect this symmetry while those with λ in factor break it and therefore can be eliminated.

Example 3: We consider two interacting fermionic oscillators with Hamiltonian

$$H = \sum_{k=1}^{3}\left(\frac{\omega_1}{2}(a_k^\dagger a_k - a_k a_k^\dagger) + \frac{\omega_2}{2}(b_k^\dagger b_k - b_k b_k^\dagger) + g\sum_{l=1}^{3} a_k^\dagger b_k^\dagger a_l b_l\right). \quad \text{(C13)}$$

The interaction term has been constructed in such a way that H commutes with the SU(3) generators:

$$Q_a = \tfrac{1}{2}(a_k^\dagger(\lambda_a)_{kl}a_l - b_k^\dagger(\lambda_a^*)_{kl}b_l), \quad \text{(C14)}$$

where λa are the Gell-Mann matrices.

The classical Hamiltonian h reads as (up to a constant; $a_i \to \theta_i$, $b_i \to \eta_i$)

$$h = \omega_1\theta_k^*\theta_k + \omega_2\eta_k^*\eta_k + g\theta_k^*\eta_k^*\theta_l\eta_l \quad \text{(C15)}$$

and the equations of motion are

$$\begin{aligned}
(1/i)\dot{\theta}_j &= \omega_1\theta_j + g\eta_j^*\theta_l\eta_l, \\
(1/i)\dot{\eta}_j &= \omega_2\eta_j + g\theta_j^*\theta_l\eta_l.
\end{aligned} \quad \text{(C16)}$$

The associated Jordan matrix reads $\mathbb{J} = \mathrm{diag}(\Omega_1,\Omega_1,\Omega_1,\Omega_2,\Omega_2,\Omega_2)$, with $\Omega_j = (i\omega_j, -i\omega_j)$, $j = 1,2$, and it is easy to see that the Lie group generated by $\mathbb{J}^\dagger$ leaves Eqs. (C16) invariant; therefore they are automatically written in normal form. In other words, we have proved that the quantum SU(3) symmetry induces the classical equivariance under $e^{\mathbb{J}^\dagger\tau}$ implying the normal form of Hamilton's equations.

[1] F. A. Berezin and M. S. Marinov, Ann. Phys. (NY) **104**, 336 (1977).
[2] F. A. Berezin, Theor. Math. Phys. **6**, 194 (1971).

[3]Y. Oknuki and S. Kamefuchi, J. Math. Phys. **21**, 601 (1980).
[4]R. Casalbuoni, Nuovo Cimento A **33**, 389 (1976).
[5]M. A. Olshanetsky, Commun. Math. Phys. **88**, 63 (1983).
[6]S. Coleman, Phys. Rev. D **11**, 2088 (1975).
[7]D. Kulish and E. Nissimov, JETP Lett. **24**, 247 (1976).
[8]H. C. Morris, J. Math. Phys. **19**, 85 (1978).
[9]H. D. Wahlquist and F. B. Estabrook, J. Math. Phys. **16**, 1 (1975).
[10]F. A. Berezin, *The Method of Second Quantization* (Academic, New York, 1966).
[11]V. I. Arnol'd, *Geometrical Methods in the Theory of Ordinary Differential Equations* (Springer, New York, 1977).

Lie transformations, similarity reduction, and solutions for the nonlinear Madelung fluid equations with external potential

G. Baumann and T. F. Nonnenmacher
Department of Mathematical Physics, University of Ulm, 7900 Ulm, West Germany

(Received 20 February 1986; accepted for publication 10 December 1986)

The application of Lie-group methods to a system of coupled nonlinear partial differential equations representing what is usually called a Madelung fluid is shown. The generating operators of the transformation group that depends on five arbitrary group constants will be constructed, and all subclasses of systems of ordinary differential equations derived by similarity reduction will be presented in tabular form. Two subclasses of physical interest are investigated in detail and the similarity solutions are compared with solutions found earlier by the application of inverse scattering transform techniques to the cubic nonlinear Schrödinger equation. Similarity solutions for the Madelung equations with linear external potential $\Gamma(x) = -f_0 x$ are presented.

I. INTRODUCTION

In recent years considerable interest has been focused on nonlinear evolution equations and their methods of solution. Among the most powerful methods for solving nonlinear partial differential equations are the inverse scattering transform technique (IST) and the Lie-group-based similarity method (LSM) originally initiated by Lie[1] in his classical integration theory. The basic idea of Lie's approach is to study the invariance properties of given differential equations under continuous groups of transformations. If the most extended Lie group of transformations of a given system S is known, then it is possible to construct classes of particular solutions, called the similarity solutions of S.

The LSM seems to be applicable to a broader class of nonlinear evolution equations than the IST, which has been applied intensively to nondissipative, nonlinear partial differential equations (NPDE) representing completely integrable Hamiltonian systems.[2–6] Solitons are the most prominent examples of nonlinear solutions constructed by the IST. The class of IST-solvable NPDE possesses a number of common properties such as infinite sequences of conservation laws, Bäcklund transformations with associated geometric and group theoretical properties,[7] and Painlevé transcendental equations characterized by no movable critical points. Relations between solitons and Painlevé-type equations has been pointed out,[8,9] and a unified approach to transformations and elementary solutions of Painlevé equations has been developed by Fokas and Ablowitz.[10]

When discussing dissipative nonlinear evolution equations such as diffusion-type equations[11] or nonlinear kinetic (Boltzmann) equations,[12–19] it is seen that IST-like techniques do not exist thus far. However, in all these cases,[11–16] the LSM has been applied successfully in order to construct similarity solutions. The most prominent examples of these types of solutions are the BKW-mode solutions discovered by Bobylev[18] and Krook and Wu[19] and classified in group-theoretic terms[12,13] by making use of the LSM.

Motivated by a successful application of the LSM on dissipative NPDE, one may ask to which extent the LSM is also applicable to IST-solvable NPDE. Lakshmanan and Kaliappan[20] have already investigated relations between Lie transformations, nonlinear evolution equations, and Painlevé forms of some pertinent examples of NPDE. They applied the LSM and obtained by a similarity reduction of KdV, sine–Gordon, nonlinear Schrödinger equations, etc., a list of nonlinear ordinary differential equations (NODE), which could be classified in part as Painlevé equations.

The central motivation of this article is to apply the LSM to the nonlinear system of Madelung's quantum fluid equations, which represent a nondissipative system of NPDE for the probability density $\rho(x,t) = \psi^*\psi$ and the phase function $S = S(x,t)$. These are related by the Madelung transformation[21]

$$\psi(x,t) = \sqrt{\rho}\,\exp(-iS/\hbar), \tag{1.1}$$

when $h = 2\pi\hbar$ is Planck's constant and $\psi(x,t)$ and $\psi^*(x,t)$ satisfy the cubic nonlinear Schrödinger equations

$$i\hbar\psi_t = -(\hbar^2/2m)\psi_{xx} + V(x)\psi + \kappa|\psi|^2\psi \tag{1.2}$$

and their complex conjugate. Here, $V(x)$ is an external potential and κ is a real-valued constant.

It is well known that the system of NPDE (1.2) is IST integrable.[5,6] We note that (1.2) can be written as a Hamiltonian system by making use of canonical Poisson brackets[6,22,23] and we mention that the canonical transformation (1.1) transforms[22,23] the system of NPDE (1.2) for the complex-valued functions ψ and ψ^* into a system of NPDE for real-valued functions ρ and $S = m\phi$, which represents what is usually called a Madelung fluid[21–24]:

$$\phi_t + \frac{1}{2}\phi_x^2 + \alpha\rho + \Gamma(x) = \beta\frac{1}{\sqrt{\rho}}\frac{\partial^2\sqrt{\rho}}{\partial x^2}, \tag{1.3a}$$

$$\rho_t + \partial_x(\rho\phi_x) = 0, \tag{1.3b}$$

with

$$\alpha = \kappa/m, \quad \beta = \hbar^2/2m^2, \quad \Gamma(x) = (1/m)V(x).$$

We shall apply the LSM to the system (S) of NPDE (1.3) in order to study the invariance of S under continuous groups of transformations depending on one infinitesimal parameter ϵ. The most extended (ϵ) Lie group of transformations, admitted by S, will be shown to depend on five

 0022-2488/87/061250-11$02.50

arbitrary group constants, and the general class of similarity solutions will be seen to separate into different subclasses according to the number of nonzero group constants. It will be analyzed to what extent these subclasses of similarity solutions differ from (or coincide with) classes of solutions already obtained by means of quite different analytical methods.

Most of the previous attention in the literature has been focused on the nonlinear Schrödinger equation (1.2) without an external potential [$V(x)=0$]. Including a potential of type $V(x)=-mf_0x$, Alonso[25] studied the invariance of Eq. (1.2) under the Galilei group from the viewpoint of the IST. In recent times, however, some interest in the Madelung fluid equations (1.3) came from hydrodynamics, in order to describe weakly interacting Bose condensates.[26–28] Analytical solutions of Eq. (1.3) have been found for $\alpha=0$ and $\Gamma(x)=0$ by applying[29] the Lie–Bäcklund transformation, and for $\alpha\neq 0$ [$\Gamma(x)=0$] by transforming the known soliton solutions of the nonlinear Schrödinger equation (1.2) to the fluid dynamical field variables[30,31] $\rho(x,t)$ and $\phi(x,t)$. A systematic application of the IST or the LSM on the system of Madelung fluid equations (1.3) has not been given thus far. Hence we shall apply the LSM on (1.3) in order to construct by similarity reduction the most extended class of ordinary differential equations for the similarity functions. Some of them can be classified as Painlevé-type equations and some exact heretofore undiscovered classes of similarity solutions will be presented.

The present article is organized as follows. First, in Sec. II, we define the symbols and notation for the group generators, similarity forms, etc., by giving a brief reformulation of Lie's basic concept of a one-parameter (ϵ) transformation group in a form that is applicable to the system (1.3) of partial differential equations. In Sec. III we construct the generators of the group, the similarity variables for the most extended class, and some subclasses of physical interest, and derive by reduction the corresponding classes of ordinary differential equations, which are presented in tabular forms. In Sec. IV we construct exact classes of similarity solutions for the density $\rho(x,t)$ and the phase variable $S(x,t)=m\phi(x,t)$, taking into account two choices for the external potential [$\Gamma(x)=0$ and $\Gamma(x)=-f_0x$].

II. LIE GROUP OF TRANSFORMATIONS, GENERATORS OF THE GROUP, AND REDUCTION SCHEME

Consider a system of partial differential equations with two dependent variables ρ and ϕ and two independent variables x and t [as in Eq. (1.3)]:

$$H_i(x,t,\phi,\rho,\phi_x,\rho_x,\phi_t,\rho_t,\phi_{xx},\rho_{xx},\dots)=0,\quad i=1,2, \tag{2.1}$$

where subscripts denote partial differentiations. Consider further a one-parameter (ϵ) Lie group of transformations

$$\begin{aligned} x'&=f(x,t,\rho,\phi;\ \epsilon),\quad t'=g(x,t,\rho,\phi;\ \epsilon),\\ \rho'&=h(x,t,\rho,\phi;\ \epsilon),\quad \phi'=j(x,t,\rho,\phi;\ \epsilon). \end{aligned} \tag{2.2}$$

Let $\rho=\theta(x,t)$ and $\phi=\Xi(x,t)$ be solutions of (2.1). If we replace the variables ρ, ϕ, x, and t in Eq. (2.1) by v, w, and $x'=f(x,t,\theta,\Xi;\epsilon)$, $t'=g(x,t,\theta,\Xi;\epsilon)$, Eq. (2.1) becomes

$$H_i(x',t',v,w,v_{x'},w_{x'},v_{t'},w_{t'},v_{x'x'},w_{x'x'},\dots)=0,\quad i=1,2. \tag{2.3}$$

Then $v=\theta(x',t')$ and $w=\Xi(x',t')$ are solutions of (2.3). We say that the transformations (2.2) leave Eq. (2.1) invariant if $v=h(x,t,\theta,\Xi;\epsilon)$ and $w=j(x,t,\theta,\Xi;\epsilon)$ are solutions to (2.3) whenever $\rho=\theta(x,t)$ and $\phi=\Xi(x,t)$ are solutions to (2.1). This condition implies that if Eqs. (2.1) and (2.3) have a unique solution, then

$$\begin{aligned} \theta(x',t')&=h(x,t,\theta(x,t),\Xi(x,t);\epsilon),\\ \Xi(x',t')&=j(x,t,\theta(x,t),\Xi(x,t);\epsilon). \end{aligned} \tag{2.4}$$

Hence $\theta(x,t)$ and $\Xi(x,t)$ satisfy the one-parameter functional equations

$$\begin{aligned} \theta(f(x,t,\theta,\Xi;\epsilon),g(x,t,\theta,\Xi;\epsilon))&=h(x,t,\theta,\Xi;\epsilon),\\ \Xi(f(x,t,\theta,\Xi;\epsilon),g(x,t,\theta,\Xi;\epsilon))&=j(x,t,\theta,\Xi;\epsilon). \end{aligned} \tag{2.5}$$

Expanding (2.2) about the identity $\epsilon=0$, one can generate the following infinitesimal transformations:

$$\begin{aligned} x'&=x+\epsilon\xi_1(x,t,\rho,\phi)+O(\epsilon^2),\\ t'&=t+\epsilon\xi_2(x,t,\rho,\phi)+O(\epsilon^2),\\ \phi'&=\phi+\epsilon\eta^1(x,t,\rho,\phi)+O(\epsilon^2),\\ \rho'&=\rho+\epsilon\eta^2(x,t,\rho,\phi)+O(\epsilon^2). \end{aligned} \tag{2.6}$$

The functions ξ_1, ξ_2, η^1, and η^2 are the infinitesimals of the transformations for the variables x, t, ϕ, and ρ, respectively. In order to find the infinitesimals we need to extend the group to calculate how derivative terms transform. The transformations (2.6), together with the transformations for the first, second,... derivatives, are called first, second,... extensions. We denote the infinitesimals for ρ_x, ρ_t, ρ_{xx}, ϕ_x, ϕ_t, and ϕ_{xx} by η_1^2, η_2^2, η_{11}^2, η_1^1, η_2^1, and η_{11}^1, respectively. As an example, we give

$$\begin{aligned} \eta_1^1=&\frac{\partial\eta^1}{\partial x}+\left\{\frac{\partial\eta^1}{\partial\phi}\phi_x+\frac{\partial\eta^1}{\partial\rho}\rho_x\right\}-\frac{\partial\xi_1}{\partial x}\phi_x-\frac{\partial\xi_2}{\partial x}\phi_t\\ &-\frac{\partial\xi_1}{\partial\phi}\phi_x\phi_x-\frac{\partial\xi_1}{\partial\rho}\rho_x\phi_x-\frac{\partial\xi_2}{\partial\phi}\phi_x\phi_t-\frac{\partial\xi_2}{\partial\rho}\rho_x\phi_t. \end{aligned} \tag{2.7}$$

Similar explicit expressions for higher extensions can be given.[11] Using these various extensions, the infinitesimal criteria for the invariance of (2.1) under the group (2.2) is given by

$$\hat{X}H_i|_{H_i=0}=0,\quad i=1,2, \tag{2.8}$$

where the tangent vector field $\hat{X}$ is given by

$$\begin{aligned} \hat{X}=\hat{X}_U&+\eta_1^1\frac{\partial}{\partial\phi_x}+\eta_2^1\frac{\partial}{\partial\phi_t}+\eta_1^2\frac{\partial}{\partial\rho_x}+\eta_2^2\frac{\partial}{\partial\rho_t}\\ &+\eta_{11}^1\frac{\partial}{\partial\phi_{xx}}+\eta_{11}^2\frac{\partial}{\partial\rho_{xx}}+\cdots, \end{aligned} \tag{2.9a}$$

$$\hat{X}_U=\xi_1\partial_x+\xi_2\partial_t+\eta^1\frac{\partial}{\partial\phi}+\eta^2\frac{\partial}{\partial\rho}. \tag{2.9b}$$

Here, $\hat{X}_U$ is the unextended operator. Condition (2.8) provides an algorithm for finding ξ_1, ξ_2, η^1, and η^2. For any solutions $\rho=\theta(x,t)$ and $\phi=\Xi(x,t)$ of (2.1), Eq. (2.8) may be treated as a form in derivatives of θ and Ξ whose coeffi-

cients depend on (θ,Ξ,x,t) and the unknowns $(\xi_1,\xi_2,\eta^1,\eta^2)$. Collecting together the coefficients of like derivative terms in θ and Ξ and setting all of them equal to zero we get a system of linear partial differential equations. In practice these equations are solvable and thus ξ_1, ξ_2, η^1, and η^2 are determined.

Our objective now is to find $\theta(x,t)$ and $\Xi(x,t)$ given that (2.2) leaves (2.1) invariant. Expanding (2.2) about the identity $\epsilon = 0$ we generate Eq. (2.6). We now use (2.6) to expand the functional equation (2.5) about $\epsilon = 0$. This leads to the following first-order partial differential equations:

$$\xi_1\phi_x + \xi_2\phi_t = \eta^1, \quad \xi_1\rho_x + \xi_2\rho_t = \eta^2, \tag{2.10}$$

provided that ξ_1, ξ_2, η^1, and η^2 are known functions of x, t, ϕ, and ρ. Equation (2.10) is called the invariant surface condition. The solutions of (2.10) are obtained by solving the following characteristic equations:

$$\frac{dx}{\xi_1} = \frac{dt}{\xi_2} = \frac{d\phi}{\eta^1} = \frac{d\rho}{\eta^2}. \tag{2.11}$$

The general solution of these equations will involve three arbitrary constants, of which one constant takes the role of similarity variable ζ and the other constants, say $F_1(\zeta)$ and $F_2(\zeta)$, play the role of dependent variables (usually called similarity functions). Thus we finally obtain the similarity forms

$$v = E(x,t,F_1(\zeta))$$

and

$$w = G(x,t,F_2(\zeta)). \tag{2.12}$$

Substitution of (2.12) into (2.1) results in a system of ordinary differential equations for $F_1(\zeta)$ and $F_2(\zeta)$. The results mentioned above for two dependent and independent variables can be extended to any number of dependent and independent variables. In the following sections we will give an application of the above procedure to the nonlinear Madelung fluid equations (1.3).

III. APPLICATION TO THE MADELUNG FLUID EQUATIONS

A. The infinitesimal elements

In applying the infinitesimal Lie-group methods, a straightforward calculation yields the following infinitesimal elements of the ϵ-Lie group [for the external potential we take the choice $\Gamma(x) = -f_0 x$]:

$$\begin{aligned}
&\xi_1 = Bx + Ft + H + \tfrac{3}{2}Bf_0t^2,\\
&\xi_2 = 2Bt + C,\\
&\eta^1 = (F + 3Bf_0t)x + G + f_0\left[\tfrac{1}{2}Bf_0t^3 + (F/2)t^2 + Ht\right],\\
&\eta^2 = -2B\rho.
\end{aligned} \tag{3.1}$$

Thus we obtained a one-parameter group of transformations depending on five arbitrary group constants (B,C,F,G,H). For the special case $\Gamma(x) = 0$, i.e., $f_0 = 0$, it follows that

$$\begin{aligned}
&\xi_1 = Bx + Ft + H, \quad \xi_2 = 2Bt + C,\\
&\eta^1 = Fx + G, \quad \eta^2 = -2B\rho.
\end{aligned} \tag{3.2}$$

From the most extended group (3.1) one finds 22 nontrivial subgroups listed in Table I, where taking f_0 equal to zero one obtains the equations for the Madelung fluid without external potential.

The similarity variables ζ as well as the similarity solutions $\theta(x,t)$ and $\Xi(x,t)$ are found by solving the characteristic equations (2.11). The results of these integrations depend crucially on the number of vanishing group constants. Here, we will focus our attention on the classes of similarity solutions characterized by the choices $B = 0$ and by $B = F = 0$. It will be shown that these subclasses of similarity solutions lead to solutions of the Painlevé II type and a classical soliton solution.

B. Lie algebra constructed from the infinitesimal operators

The knowledge of the infinitesimal elements ξ_1, ξ_2, η^1, and η^2 given in Eqs. (3.1) enables us to construct, from the unextended operator $\hat{X}_U$ [Eq. (2.9b)], five operators $\hat{X}_i$ $(i = 1,\ldots,5)$ according to the existence of five group constants (B,C,F,G,H). Taking the group constants equal to 0 one obtains from (2.9b) via (3.1) the null operator $\hat{X}_0 = 0$. The five generating operators $\hat{X}_i$ can be constructed by taking one of the group constants equal to 1 and the remaining four constants equal to zero: for $(B,C,F,G,H) = (1,0,0,0,0)$ one obtains

$$\hat{X}_1 = \left(x + \frac{3}{2}f_0t^2\right)\frac{\partial}{\partial x} + 2t\frac{\partial}{\partial t} + \left(3f_0tx + \frac{f_0^2}{2}t^3\right)\frac{\partial}{\partial\phi} - 2\rho\frac{\partial}{\partial\rho};$$

for (0, 1, 0, 0, 0) one obtains

$$\hat{X}_2 = \frac{\partial}{\partial t};$$

for (0, 0, 1, 0, 0) one obtains

$$\hat{X}_3 = t\frac{\partial}{\partial x} + \left(x + \frac{f_0}{2}t^2\right)\frac{\partial}{\partial\phi};$$

for (0, 0, 0, 1, 0) one obtains

$$\hat{X}_4 = \frac{\partial}{\partial\phi};$$

and for (0, 0, 0, 0, 1) one obtains

$$\hat{X}_5 = \frac{\partial}{\partial x} + f_0t\frac{\partial}{\partial\phi}.$$

These operators must form a Lie algebra. Thus, in general, we have to prove that the commutator of any two operators is a linear combination of these same operators with constant coefficients C_{ijm} (structure constants), i.e., the commutator relation (closure property)

$$[\hat{X}_i,\hat{X}_j] = C_{ijm}\hat{X}_m$$

must hold and it must be shown that the following properties for the commutators are being satisfied: (i) antisymmetry

$$[\hat{X}_i,\hat{X}_j] = -[\hat{X}_j,\hat{X}_i],$$

and (ii) the Jacobi identity

$$[\hat{X}_i,[\hat{X}_j,\hat{X}_k]] + [\hat{X}_j,[\hat{X}_k,\hat{X}_i]] + [\hat{X}_k,[\hat{X}_i,\hat{X}_j]] = 0.$$

Proving property (i) is equivalent to showing that

TABLE I. Similarity variables and ordinary differential equations for the Madelung fluid with $\Gamma(x) = -f_0 x$.

Case	Similarity variables	Ordinary differential equations
$B,C,G,F,H \neq 0$	$\zeta = \dfrac{x-(F/B^2)(Bt+C)+H/B-(f_0/2B^2)(B^2t^2-2BCt-2C^2)}{(2Bt+C)^{1/2}}$ $\phi = (2Bt+C)^{1/2}\left\{\dfrac{F}{B}+f_0t-\dfrac{C}{B}f_0\right\}\zeta$ $+\dfrac{1}{2}\ln(2Bt+C)\left\{\dfrac{G}{B}+\dfrac{CHf_0-FH}{B^2}+\dfrac{C^3f_0^2-2C^2Ff_0+CF^2}{2B^3}\right\}$ $+\dfrac{f_0^2}{3}t^3+t^2\dfrac{f_0}{B}(F-Cf_0)+t\dfrac{1}{2B^2}(F^2-C^2f_0^2-2HBf_0)+F_1(\zeta)$ $\rho = F_2(\zeta)/(2Bt+C)$	$(\ddot{F}_1-2B)F_2+\dot{F}_2(\dot{F}_1-B\zeta)=0$ $F_2\ddot{F}_2-\dfrac{1}{2}\dot{F}_2^2-\dfrac{\alpha}{\beta}F_2^3+F_2^2\left\{\dfrac{B}{\beta}\zeta\dot{F}_1-\dfrac{1}{2\beta}\dot{F}_1^2+\dfrac{FH}{\beta B}+\dfrac{C^2Ff_0}{\beta B^2}-\dfrac{CF^2}{2\beta B^2}-\dfrac{C^3f_0}{2\beta B^2}-\dfrac{f_0CH}{\beta B}-\dfrac{G}{\beta}\right\}=0$
$B=0$	$\zeta = x-(F/2C)t^2-(H/C)t$ $\phi = \dfrac{F^2}{6C^2}t^3+\dfrac{FH}{2C^2}t^2+\dfrac{F}{C}\zeta t+\dfrac{G}{C}t$ $+f_0\left(\dfrac{F}{6C}t^3+\dfrac{H}{2C}t^2\right)+F_1(\zeta)$ $\rho = F_2(\zeta)$	$\ddot{F}_1F_2+\dot{F}_2(\dot{F}_1-H/C)=0$ $F_2\ddot{F}_2-\dfrac{1}{2}\dot{F}_2^2-\dfrac{\alpha}{\beta}F_2^3+F_2^2\left\{\left(\dfrac{f_0}{\beta}-\dfrac{F}{\beta C}\right)\zeta+\dfrac{H}{\beta C}\dot{F}_1-\dfrac{1}{2\beta}\dot{F}_1^2-\dfrac{G}{\beta C}\right\}=0$
$C=0$	$\zeta = [x-(F/B)t+H/B-(f_0/2)t^2]/\sqrt{t}$ $\phi = \sqrt{t}\left\{\dfrac{F}{B}+f_0t\right\}\zeta+\dfrac{1}{2}\ln t\left\{\dfrac{G}{B}-\dfrac{FH}{B^2}\right\}$ $+\dfrac{f_0^2}{3}t^3+\dfrac{F}{B}f_0t^2+t\left\{\dfrac{F^2}{2B^2}-\dfrac{Hf_0}{B}\right\}+F_1(\zeta)$ $\rho = F_2(\zeta)/t$	$(\ddot{F}_1-1)F_2+\dot{F}_2(\dot{F}_1-\tfrac{1}{2}\zeta)=0$ $F_2\ddot{F}_2-\dfrac{1}{2}\dot{F}_2^2-\dfrac{\alpha}{\beta}F_2^3+F_2^2\times\left\{\dfrac{1}{2\beta}\zeta\dot{F}_1-\dfrac{1}{2\beta}\dot{F}_2^2+\dfrac{FH}{2\beta B^2}-\dfrac{G}{2\beta B}\right\}=0$
$G=0$	$\zeta = \dfrac{x-(F/B^2)(Bt+C)+H/B-(f_0/2B^2)(B^2t^2-2BCt-2C^2)}{(2Bt+C)^{1/2}}$ $\phi = (2Bt+C)^{1/2}\left\{\dfrac{F}{B}+f_0t-\dfrac{C}{B}f_0\right\}\zeta$ $+\dfrac{1}{2}\ln(2Bt+C)\left\{\dfrac{CHf_0-FH}{B^2}+\dfrac{C^3f_0-2C^2Ff_0+CF^2}{2B^3}\right\}$ $+\dfrac{f_0^2}{3}t^3+t^2\dfrac{f_0}{B}(F-Cf_0)+t\dfrac{1}{2B^2}(F^2-C^2f_0^2-2HBf_0)+F_1(\zeta)$ $\rho = F_2(\zeta)/(2Bt+C)$	$(\ddot{F}_1-2B)F_2+\dot{F}_2(\dot{F}_1-B\zeta)=0$ $F_2\ddot{F}_2-\dfrac{1}{2}\dot{F}_2^2-\dfrac{\alpha}{\beta}F_2^3+F_2^2\left\{\dfrac{B}{\beta}\zeta\dot{F}_1-\dfrac{1}{2\beta}\dot{F}_1^2+\dfrac{FH}{\beta B}-\dfrac{CF^2}{2\beta B^2}+\dfrac{C^2Ff_0}{\beta B^2}-\dfrac{C^3f_0^2}{2\beta B^2}-\dfrac{HCf_0}{\beta B}\right\}=0$
$F=0$	$\zeta = \dfrac{x+H/B-(f_0/2B^2)(B^2t^2-2BCt-2C^2)}{(2Bt+C)^{1/2}}$ $\phi = (2Bt+C)^{1/2}\left\{f_0t-\dfrac{C}{B}f_0\right\}\zeta$ $+\dfrac{1}{2}\ln(2Bt+C)\left\{\dfrac{G}{B}+\dfrac{CHf_0}{B^2}+\dfrac{C^3f_0^2}{2B^3}\right\}$ $+\dfrac{f_0^2}{3}t^3-\dfrac{Cf_0^2}{B}t^2-t\dfrac{1}{2B^2}(C^2f_0^2+2HBf_0)+F_1(\zeta)$ $\rho = F_2(\zeta)/(2Bt+C)$	$(\ddot{F}_1-2B)F_2+\dot{F}_2(\dot{F}_1-B\zeta)=0$ $F_2\ddot{F}_2-\dfrac{1}{2}\dot{F}_2^2-\dfrac{\alpha}{\beta}F_2^3+F_2^2\times\left\{\dfrac{1}{\beta}\zeta\dot{F}_1-\dfrac{1}{2\beta}\dot{F}_1^2-\dfrac{G}{\beta}-\dfrac{CHf_0}{\beta B}-\dfrac{C^3f_0^2}{2\beta B^2}\right\}=0$
$H=0$	$\zeta = \dfrac{x-(F/B^2)(Bt+C)-(f_0/2B^2)(B^2t^2-2BCt-2C^2)}{(2Bt+C)^{1/2}}$ $\phi = (2Bt+C)^{1/2}\left\{\dfrac{F}{B}+f_0t-\dfrac{C}{B}f_0\right\}\zeta$ $+\dfrac{1}{2}\ln(2Bt+C)\left\{\dfrac{G}{B}+\dfrac{C^3f_0^2-2C^2Ff_0+CF^2}{2B^3}\right\}$ $+\dfrac{f_0^2}{3}t^3+t^2\dfrac{f_0}{B}(F-Cf_0)+t\dfrac{1}{2B^2}(F^2-C^2f_0^2)+F_1(\zeta)$ $\rho = F_2(\zeta)/(2Bt+C)$	$(\ddot{F}_1-2B)F_2+\dot{F}_2(\dot{F}_1-B\zeta)=0$ $F_2\ddot{F}_2-\dfrac{1}{2}\dot{F}_2^2-\dfrac{\alpha}{\beta}F_2^3+F_2^2\left\{\dfrac{B}{\beta}\zeta\dot{F}_1-\dfrac{1}{2\beta}\dot{F}_1^2+\dfrac{C^2Ff_0}{\beta B^2}-\dfrac{CF^2}{2\beta B^2}-\dfrac{C^3f_0}{2\beta B^2}-\dfrac{G}{\beta}\right\}=0$
$C=0$ and $G=0$	$\zeta = [x-(F/B)t+H/B-(f_0/2)t^2]/\sqrt{t}$ $\phi = \sqrt{t}\left\{\dfrac{F}{B}+f_0t\right\}\zeta-\dfrac{1}{2}\dfrac{FH}{B^2}\ln t+\dfrac{f_0^2}{3}t^3+\dfrac{F}{B}f_0t^2$ $+t\left\{\dfrac{F^2}{2B^2}-\dfrac{Hf_0}{B}\right\}+F_1(\zeta)$ $\rho = F_2(\zeta)/t$	$(\ddot{F}_1-1)F_2+\dot{F}_2(\dot{F}_1-\tfrac{1}{2}\zeta)=0$ $F_2\ddot{F}_2-\dfrac{1}{2}\dot{F}_2^2-\dfrac{\alpha}{\beta}F_2^3+F_2^2\times\left\{\dfrac{1}{2\beta}\zeta\dot{F}_1-\dfrac{1}{2\beta}\dot{F}_1^2+\dfrac{FH}{2\beta B^2}\right\}=0$
$G=0$ and $F=0$	$\zeta = \dfrac{x+H/B-(f_0/2B^2)(B^2t^2-2BCt-2C^2)}{(2Bt+C)^{1/2}}$ $\phi = (2Bt+C)^{1/2}\left\{f_0t-\dfrac{C}{B}f_0\right\}\zeta$ $+\dfrac{1}{2}\ln(2Bt+C)\left\{\dfrac{CHf_0}{B^2}+\dfrac{C^3f_0^2}{2B^3}\right\}$ $+\dfrac{f_0^2}{3}t^3-\dfrac{Cf_0^2}{B}t^2-t\dfrac{1}{2B^2}(C^2f_0^2+2HBf_0)+F_1(\zeta)$	$(\ddot{F}_1-2B)F_2+\dot{F}_2(\dot{F}_1-B\zeta)=0$ $F_2\ddot{F}_2-\dfrac{1}{2}\dot{F}_2^2-\dfrac{\alpha}{\beta}F_2^3+F_2^2\times\left\{\dfrac{B}{\beta}\zeta\dot{F}_1-\dfrac{1}{2\beta}\dot{F}_1^2-\dfrac{CHf_0}{\beta B}-\dfrac{C^3f_0^2}{2\beta B^2}\right\}=0$

TABLE I. (*Continued.*)

Case	Similarity variables	Ordinary differential equations
	$\rho = F_2(\zeta)/(2Bt + C)$	
$F = 0$ and $H = 0$	$\zeta = \dfrac{x - (f_0/2B^2)(B^2t^2 - 2BCt - 2C^2)}{(2Bt + C)^{1/2}}$ $\phi = (2Bt + C)^{1/2}\left\{f_0 t - \dfrac{C}{B} f_0\right\}\zeta$ $+ \dfrac{1}{2}\ln(2Bt + C)\left\{\dfrac{G}{B} + \dfrac{C^3 f_0^2}{2B^3}\right\}$ $+ \dfrac{f_0^2}{3} t^3 - \dfrac{Cf_0^2}{B} t^2 - \dfrac{C^2 f_0^2}{2B^2} t + F_1(\zeta)$ $\rho = F_2(\zeta)/(2Bt + C)$	$(\ddot{F}_1 - 2B)F_2 + \dot{F}_2(\dot{F}_1 - B\zeta) = 0$ $F_2\ddot{F}_2 - \dfrac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3 + F_2^2$ $\times\left\{\dfrac{1}{\beta}\zeta\dot{F}_1 - \dfrac{1}{2\beta}\dot{F}_1^2 - \dfrac{G}{\beta} - \dfrac{C^3 f_0^2}{2\beta B^2}\right\} = 0$
$B = 0$ and $G = 0$	$\zeta = x - (F/2C)t^2 + (H/C)t$ $\phi = \dfrac{F^2}{6C} t^3 + \dfrac{FH}{2C^2} t^2 + \dfrac{F}{C}\zeta t + \dfrac{f_0}{C}\left(\dfrac{F}{6} t^3 + \dfrac{H}{2} t^2\right) + F_1(\zeta)$ $\rho = F_2(\zeta)$	$\ddot{F}_1 F_2 + \dot{F}_2(\dot{F}_1 - H/C) = 0$ $F_2\ddot{F}_2 - \dfrac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3 + F_2^2$ $\times\left\{\left(\dfrac{f_0}{\beta} - \dfrac{F}{\beta C}\right)\zeta + \dfrac{H}{\beta C}\dot{F}_1 - \dfrac{1}{2\beta}\dot{F}_1^2\right\} = 0$
$C = 0$ and $F = 0$	$\zeta = [x + H/B - (f_0/2)t^2]/\sqrt{t}$ $\phi = \sqrt{t} f_0 t\zeta + \dfrac{G}{2B}\ln t + \dfrac{f_0^2}{3} t^3 - \dfrac{Hf_0}{B} t + F_1(\zeta)$ $\rho = F_2(\zeta)/t$	$(\ddot{F}_1 - 1)F_2 + \dot{F}_2(\dot{F}_1 - \frac{1}{2}\zeta) = 0$ $F_2\ddot{F}_2 - \dfrac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3 + F_2^2$ $\times\left\{\dfrac{1}{2\beta}\zeta\dot{F}_1 - \dfrac{1}{2\beta}\dot{F}_2^2 - \dfrac{G}{2\beta B}\right\} = 0$
$G = 0$ and $H = 0$	$\zeta = \dfrac{x - (F/B^2)(Bt + C) - (f_0/2B^2)(B^2t^2 - 2BCt - 2C^2)}{(2Bt + C)^{1/2}}$ $\phi = (2Bt + C)^{1/2}\left\{\dfrac{F}{B} + f_0 t - \dfrac{C}{B} f_0\right\}\zeta$ $+ \dfrac{1}{2}\ln(2Bt + C)\left\{\dfrac{C^3 f_0 - 2C^2 F f_0 + CF^2}{2B^3}\right\}$ $+ \dfrac{f_0^2}{3} t^3 + t^2\dfrac{f_0}{B}(F - Cf_0) + t\dfrac{F^2 - C^2 f_0^2}{2B^2} + F_1(\zeta)$ $\rho = F_2(\zeta)/(2Bt + C)$	$(\ddot{F}_1 - 2B)F_2 + \dot{F}_2(\dot{F}_1 - B\zeta) = 0$ $F_2\ddot{F}_2 - \dfrac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3$ $+ F_2^2\left\{\dfrac{B}{\beta}\zeta\dot{F}_1 - \dfrac{1}{2\beta}\dot{F}_1^2\right.$ $\left. - \dfrac{CF^2}{2\beta B^2} + \dfrac{C^2 F f_0}{\beta B^2} - \dfrac{C^3 f_0^2}{2\beta B^2}\right\} = 0$
$B = 0$ and $F = 0$	$\zeta = x - (H/C)t$ $\phi = (G/C)t + (f_0 H/2C)t^2 + F_1(\zeta)$ $\rho = F_2(\zeta)$	$\ddot{F}_1 F_2 + \dot{F}_2(\dot{F}_1 - H/C) = 0$ $F_2\ddot{F}_2 - \dfrac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3 + F_2^2$ $\times\left\{\dfrac{f_0}{\beta}\zeta + \dfrac{H}{\beta C}\dot{F}_1 - \dfrac{1}{2\beta}\dot{F}_1^2 - \dfrac{G}{\beta C}\right\} = 0$
$C = 0$ and $H = 0$	$\zeta = [x - (F/B)t - (f_0/2)t^2]/\sqrt{t}$ $\phi = \sqrt{t}\left\{\dfrac{F}{B} + f_0 t\right\}\zeta + \dfrac{G}{2B}\ln t + \dfrac{f_0^2}{3} t^3$ $+ \dfrac{F}{B} f_0 t^2 + \dfrac{F^2}{2B^2} t + F_1(\zeta)$ $\rho = F_2(\zeta)/t$	$(\ddot{F}_1 - 1)F_2 + \dot{F}_2(\dot{F}_1 - \frac{1}{2}\zeta) = 0$ $F_2\ddot{F}_2 - \dfrac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3 + F_2$ $\times\left\{\dfrac{1}{2\beta}\zeta\dot{F}_1 - \dfrac{1}{2\beta}\dot{F}_1^2 - \dfrac{G}{2\beta B}\right\} = 0$
$B = 0$ and $H = 0$	$\zeta = x - (F/2C)t^2$ $\phi = \dfrac{F^2}{6C^2} t^3 + \dfrac{F}{C}\zeta t + \dfrac{G}{C} t + \dfrac{f_0 F}{6C} t^3 + F_1(\zeta)$ $\rho = F_2(\zeta)$	$\ddot{F}_1 F_2 + \dot{F}_2\dot{F}_1 = 0$ $F_2\ddot{F}_2 - \dfrac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3 + F_2^2$ $\times\left\{\left(\dfrac{f_0}{\beta} - \dfrac{F}{\beta C}\right)\zeta - \dfrac{1}{2\beta}\dot{F}_1^2 - \dfrac{G}{\beta C}\right\} = 0$
$C = 0$, $G = 0$, and $F = 0$	$\zeta = [x + H/B - (f_0/2)t^2]/\sqrt{t}$ $\phi = \sqrt{t} f_0 t\zeta + (f_0^2/3)t^3 - (Hf_0/B)t + F_1(\zeta)$ $\rho = F_2(\zeta)/t$	$(\ddot{F}_1 - 1)F_2 + \dot{F}_2(\dot{F}_1 - \frac{1}{2}\zeta) = 0$ $F_2\ddot{F}_2 - \frac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3 + F_2^2$ $[(1/2\beta)\zeta\dot{F}_1 - (1/2\beta)\dot{F}_1^2] = 0$
$G = 0$, $F = 0$, and $H = 0$	$\zeta = \dfrac{x - (f_0/2B^2)(B^2t^2 - 2BCt - 2C^2)}{(2Bt + C)^{1/2}}$ $\phi = (2Bt + C)^{1/2}\left\{f_0 t - \dfrac{C}{B} f_0\right\}\zeta$ $+ \dfrac{1}{2}\dfrac{C^3 f_0^2}{2B^3}\ln(2Bt + C)$ $+ \dfrac{f_0^2}{3} t^3 - \dfrac{Cf_0^2}{B} t^2 - \dfrac{C^2 f_0^2}{2B^2} t + F_1(\zeta)$ $\rho = F_2(\zeta)/(2Bt + C)$	$(\ddot{F}_1 - 2B)F_2 + \dot{F}_2(\dot{F}_1 - B\zeta) = 0$ $F_2\ddot{F}_2 - \dfrac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3 + F_2^2$ $\times\left\{\dfrac{1}{\beta}\zeta\dot{F}_1 - \dfrac{1}{2\beta}\dot{F}_1^2 - \dfrac{C^3 f_0^2}{2\beta B^2}\right\} = 0$
$B = 0$, $G = 0$, and $F = 0$	$\zeta = x - (H/C)t$ $\phi = (f_0 H/2C)t^2 + F_1(\zeta)$	$\ddot{F}_1 F_2 + \dot{F}_2(\dot{F}_1 - H/C) = 0$ $F_2\ddot{F}_2 - \dfrac{1}{2}\dot{F}_2^2 - \dfrac{\alpha}{\beta}F_2^3 + F_2^2$

TABLE I. (*Continued.*)

Case	Similarity variables	Ordinary differential equations
	$\rho = F_2(\zeta)$	$\times\left\{\frac{f_0}{\beta}\zeta + \frac{H}{\beta C}\dot{F}_1 - \frac{1}{2\beta}\dot{F}_1^2\right\} = 0$
$C=0$, $F=0$, and $H=0$	$\zeta = [x - (f_0/2)t^2]/\sqrt{t}$	$(\ddot{F}_1 - 1)F_2 + \dot{F}_2(\dot{F}_1 - \frac{1}{2}\zeta) = 0$
	$\phi = \sqrt{t} f_0 t\zeta + (G/2B)\ln t + (f_0^3/3)t^3 + F_1(\zeta)$	$F_2\ddot{F}_2 - \frac{1}{2}\dot{F}_2^2 - \frac{\alpha}{\beta}F_2^3 + F_2^2$
	$\rho = F_2(\zeta)/t$	$\times\left\{\frac{1}{2\beta}\zeta\dot{F}_1 - \frac{1}{2\beta}\dot{F}_2^2 - \frac{G}{2\beta B}\right\} = 0$
$C=0$, $G=0$, and $H=0$	$\zeta = [x - (F/B)t - (f_0/2)t^2]/\sqrt{t}$	$(\ddot{F}_1 - 1)F_2 + \dot{F}_2(\dot{F}_1 - \frac{1}{2}\zeta) = 0$
	$\phi = \sqrt{t}\left\{\frac{F}{B} + f_0 t\right\}\zeta + \frac{f_0^2}{3}t^3 + \frac{F}{B}f_0 t^2 + \frac{F^2}{2B^2}t + F_1(\zeta)$	$F_2\ddot{F}_2 - \frac{1}{2}\dot{F}_2^2 - \frac{\alpha}{\beta}F_2^3 + F_2^2$
	$\rho = F_2(\zeta)/t$	$\times\left\{\frac{1}{2\beta}\zeta\dot{F}_1 - \frac{1}{2\beta}\dot{F}_1^2\right\} = 0$
$B=0$, $G=0$, and $H=0$	$\zeta = x - (F/2C)t^2$	$\ddot{F}_1 F_2 + \dot{F}_2 F_1 = 0$
	$\phi = (F^2/6C^2)t^3 + (F/C)\zeta t + f_0(F/6C)t^3 + F_1(\zeta)$	$F_2\ddot{F}_2 - \frac{1}{2}\dot{F}_2^2 - \frac{\alpha}{\beta}F_2^3 + F_2^2$
	$\rho = F_2(\zeta)$	$\times\left\{\left(\frac{f_0}{\beta} - \frac{F}{\beta C}\right)\zeta - \frac{1}{2\beta}\dot{F}_1^2\right\} = 0$
$C=0$, $G=0$, $F=0$, and $H=0$	$\zeta = [x - (f_0/2)t^2]/\sqrt{t}$	$(\ddot{F}_1 - 1)F_2 + \dot{F}_2(\dot{F}_1 - \frac{1}{2}\zeta) = 0$
	$\phi = \sqrt{t} f_0 t\zeta + (f_0^2/3)t^3 + F_1(\zeta)$	$F_2\ddot{F}_2 - \frac{1}{2}\dot{F}_2^2 - \frac{\alpha}{\beta}F_2^3 + F_2^2$
	$\rho = F_2(\zeta)/t$	$\times\left\{\frac{1}{2\beta}\zeta\dot{F}_1 - \frac{1}{2\beta}\dot{F}_1^2\right\} = 0$

$$C_{ijk} = -C_{jik},$$

and the Jacobi identity (ii) is equivalent to showing that

$$C_{ijm}C_{mkn} + C_{jkm}C_{min} + C_{kim}C_{mjn} = 0.$$

All these properties (closure relation, the property of antisymmetry, and the Jacobi identity) for the operators $\hat{X}_i$ follow immediately from Table II (commutator table), which shows, in addition, that $\hat{X}_4$ is a Casimir operator, i.e., it commutes with all the operators $\hat{X}_i$.

Finally we notice that the special choice $f_0 = 0$ in the operators $\hat{X}_i$ leads to the Lie algebra for the Madelung fluid equations without external potential.

C. Similarity variables

(i) The most extended class of group constants (B,C,F,G,H), all unequal to zero, leads to the following similarity variable ζ and to the similarity solutions ϕ and ρ:

$$\zeta = \frac{x - (F/B^2)(Bt + C) + (H/B) - (f_0/2B^2)(B^2t^2 - 2BCt - 2C^2)}{(2Bt + C)^{1/2}}, \tag{3.3a}$$

$$\phi = (2Bt + C)^{1/2}\zeta\left\{f_0 t - \frac{Cf_0}{B} + \frac{F}{B}\right\} + \frac{1}{2}\ln(2Bt + C) \times\left\{\frac{G}{B} + \frac{CHf_0 - HF}{B^2} + \frac{C^3f_0^2 - 2C^2Ff_0 + CF^2}{2B^3}\right\} + \frac{f_0^2}{3}t^3 + \frac{f_0F - f_0^2C}{B}t^2 + \frac{F^2 - C^2f_0^2 - 2BHf_0}{2B^2}t + F_1(\zeta), \tag{3.3b}$$

$$\rho = F_2(\zeta)/(2Bt + C). \tag{3.3c}$$

The similarity functions $F_1(\zeta)$ and $F_2(\zeta)$ are to be determined later. The special choice $\Gamma(x) = -f_0x = 0$, i.e., $f_0 = 0$, yields

$$\zeta = [x - (F/B)t - FC/B^2 + H/B]/[(2Bt + C)^{1/2}], \tag{3.4a}$$

$$\phi = \frac{F}{B}(2Bt + C)^{1/2}\zeta + \frac{1}{2}\ln(2Bt + C) \times\left\{\frac{G}{B} - \frac{FH}{B^2} + \frac{F^2C}{2B^3}\right\} + \frac{F^2}{2B^2}t + F_1(\zeta), \tag{3.4b}$$

$$\rho = F_2(\zeta)/(2Bt + C). \tag{3.4c}$$

(ii) Subclasses

(a) $(B,C,F,G,H) = (0,C,F,G,H)$

$$\zeta = x - (F/2C)t^2 - (H/C)t, \tag{3.5a}$$

TABLE II. Commutator table for the Madelung fluid with external potential $\Gamma(x) = -f_0x$.

$[\,,\,]$	$\hat{X}_0$	$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$
$\hat{X}_0$	0	0	0	0	0	0
$\hat{X}_1$	0	0	$-3f_0\hat{X}_3 - 2\hat{X}_2$	$\hat{X}_3$	0	$-\hat{X}_5$
$\hat{X}_2$	0	$3f_0\hat{X}_3 + 2\hat{X}_2$	0	$\hat{X}_5$	0	$f_0\hat{X}_4$
$\hat{X}_3$	0	$-\hat{X}_3$	$-\hat{X}_5$	0	0	$-\hat{X}_4$
$\hat{X}_4$	0	0	0	0	0	0
$\hat{X}_5$	0	$\hat{X}_5$	$-f_0\hat{X}_4$	$\hat{X}_4$	0	0

$$\phi=\frac{F^2+f_0FC}{6C^2}t^3+\frac{HF+f_0HC}{2C^2}t^2+\frac{F\zeta+G}{C}t+F_1(\zeta), \tag{3.5b}$$

$$\rho=F_2(\zeta). \tag{3.5c}$$

For $\Gamma(x)=0$, this result reduces to

$$\zeta=x-(F/2C)t^2-(H/C)t, \tag{3.6a}$$

$$\phi=\frac{F^2}{6C^2}t^3+\frac{HF}{2C^2}t^2+\frac{F\zeta+G}{C}t+F_1(\zeta), \tag{3.6b}$$

$$\rho=F_2(\zeta). \tag{3.6c}$$

(b) $(B,C,F,G,H)=(0,C,0,G,H)$. The choices $B=0$ and $F=0$ produce the results

$$\zeta=x-(H/C)t, \tag{3.7a}$$

$$\phi=(f_0H/2C)t^2+(G/C)t+F_1(\zeta), \tag{3.7b}$$

$$\rho=F_2(\zeta). \tag{3.7c}$$

For $\Gamma(x)=0$ one obtains

$$\zeta=x-(H/C)t, \tag{3.8a}$$

$$\phi=(G/C)t+F_1(\zeta), \tag{3.8b}$$

$$\rho=F_2(\zeta). \tag{3.8c}$$

We note that the similarity functions $F_1(\zeta)$ and $F_2(\zeta)$ obey ordinary differential equations, which are obtained by insertion of the corresponding expressions for ϕ and ρ into the original partial differential equations (1.3).

D. Reduction to ordinary differential equations

(i) The most extended class [all group constants (B,C,F,G,H) unequal to zero] leads—by making use of the results given in (3.3)—to the following system of ordinary differential equations (ODE) for $F_1(\zeta)$ and $F_2(\zeta)$:

$$(\ddot{F}_1-2B)F_2+\dot{F}_2(\dot{F}_1-B\zeta)=0, \tag{3.9a}$$

$$F_2\ddot{F}_2-\frac{1}{2}\dot{F}_2^2-\frac{\alpha}{\beta}F_2^3+F_2^2\left\{\frac{B}{\beta}\zeta\dot{F}_1-\frac{1}{2\beta}\dot{F}_1^2+\frac{2C^2Ff_0-CF^2-C^3f_0^2}{2\beta B^2}+\frac{FH-f_0CH}{\beta B}-\frac{G}{\beta}\right\}=0. \tag{3.9b}$$

Taking $\Gamma(x)=0$ we obtain

$$(\ddot{F}_1-2B)F_2+\dot{F}_2(\dot{F}_1-B\zeta)=0, \tag{3.10a}$$

$$F_2\ddot{F}_2-\frac{1}{2}\dot{F}_2^2-\frac{\alpha}{\beta}F_2^3+F_2^2\left\{\frac{B}{\beta}\zeta\dot{F}_1-\frac{1}{2\beta}\dot{F}_1^2-\frac{2B}{\beta}\left(\frac{F^2C}{4B^3}+\frac{G}{2B}-\frac{FH}{2B^2}\right)\right\}=0. \tag{3.10b}$$

(ii) Subclasses

(a) $(B,C,F,G,H)=(0,C,F,G,H)$ The results (3.5) lead to

$$\ddot{F}_1F_2+\dot{F}_2(\dot{F}_1-H/C)=0, \tag{3.11a}$$

$$F_2\ddot{F}_2-\frac{1}{2}\dot{F}_2^2-\frac{\alpha}{\beta}F_2^3+F_2^2\left\{\frac{H}{\beta C}\dot{F}_1-\frac{1}{2\beta}\dot{F}_1^2-\left(\frac{1}{\beta}\left(\frac{F}{C}-f_0\right)\zeta+\frac{G}{\beta C}\right)\right\}=0, \tag{3.11b}$$

and for $\Gamma(x)=0$ one obtains

$$\ddot{F}_1F_2+\dot{F}_2(\dot{F}_1-H/C)=0, \tag{3.12a}$$

$$F_2\ddot{F}_2-\frac{1}{2}\dot{F}_2^2-\frac{\alpha}{\beta}F_2^3+F_2^2\left\{\frac{H}{\beta C}\dot{F}_1-\frac{1}{2\beta}\dot{F}_1^2-\frac{1}{\beta}\left(\frac{F}{C}\zeta+\frac{G}{C}\right)\right\}=0. \tag{3.12b}$$

(b) $(B,C,F,G,H)=(0,C,0,G,H)$ The results (3.7) yield

$$\ddot{F}_1F_2+\dot{F}_2(\dot{F}_1-H/C)=0, \tag{3.13a}$$

$$F_2\ddot{F}_2-\frac{1}{2}\dot{F}_2^2-\frac{\alpha}{\beta}F_2^3+F_2^2\left\{\frac{H}{\beta C}\dot{F}_1-\frac{1}{2\beta}\dot{F}_1^2+\frac{f_0}{\beta}\zeta-\frac{G}{\beta C}\right\}=0. \tag{3.13b}$$

Taking $\Gamma(x)=0$ one obtains

$$\ddot{F}_1F_2+\dot{F}_2(\dot{F}_1-H/C)=0, \tag{3.14a}$$

$$F_2\ddot{F}_2-\frac{1}{2}\dot{F}_2^2-\frac{\alpha}{\beta}F_2^3+F_2^2\left\{\frac{H}{\beta C}\dot{F}_1-\frac{1}{2\beta}\dot{F}_1^2-\frac{G}{\beta C}\right\}=0. \tag{3.14b}$$

All other ODE are listed in Table I.

IV. EXACT CLASSES OF SIMILARITY SOLUTIONS

In Sec. III we reduced the Madelung fluid equations represented by a set of nonlinear partial differential equations (NPDE) (1.3) to various systems of nonlinear ordinary differential equations (NODE) for different choices of the set of group constants. Here we will construct analytical solutions of some special classes of NODE.

The first objective in the study of such NODE is to ascertain whether or not a solution can be obtained either explicitly or implicitly in terms of classical functions. The purpose in such a study is to discover a transformation which will reduce the equation to some type that is known to have a solution of the desired kind. Failing this, one seeks a transformation which will reduce the equation to one that is asymptotic to a form solvable by known functions.

A. Solutions of elliptic form

Following along this line we will give, as a first example, the solution of the coupled nonlinear equations:

$$\ddot{F}_1F_2+\dot{F}_2(\dot{F}_1-H/C)=0 \tag{4.1}$$

and

$$F_2\ddot{F}_2-\frac{1}{2}\dot{F}_2^2-\frac{\alpha}{\beta}F_2^3+F_2^2\times\left\{\frac{H}{\beta C}\dot{F}_1-\frac{1}{2\beta}\dot{F}_1^2-\frac{G}{\beta C}\right\}=0, \tag{4.2}$$

where overdots denote differentiation with respect to ζ, the similarity variable. This system is just subclass (ii) of case (b) from Sec. III [Eq. (3.14)] without external potential.

The first of these two equations can be integrated to give

$$F_1(\zeta)=\int_{\zeta_0}^{\zeta}\left(\frac{I_1}{F_2(\zeta')}+\frac{H}{C}\right)d\zeta', \tag{4.3}$$

where I_1 is a constant of integration. Substituting Eq. (4.3)

into the second of the coupled system, Eq. (4.2), it reduces to the form

$$F_2\ddot{F}_2 - \frac{1}{2}\dot{F}_2^2 - \frac{\alpha}{\beta}F_2^3 - F_2^2 \times\left\{\frac{G}{\beta C} - \frac{1}{2\beta}\left(\frac{H}{C}\right)^2\right\} - \frac{I_1^2}{2\beta} = 0, \tag{4.4}$$

which is an equation for F_2 alone. By means of the transformation

$$\chi^2 = F_2, \tag{4.5}$$

Eq. (4.4) takes the form

$$\ddot{\chi} - \frac{\alpha}{\beta}\chi^3 - \frac{1}{2\beta}\omega\chi - \frac{I_1^2}{4\beta}\frac{1}{\chi^3} = 0, \tag{4.6}$$

with

$$\omega = \{G/C - \tfrac{1}{2}(H/C)^2\}.$$

Introducing a "potential" $V(\chi)$ by

$$V(\chi) = -\frac{\alpha}{2\beta}\frac{1}{4}\chi^4 - \frac{1}{2\beta}\omega\frac{1}{2}\chi^2 + \frac{I_1^2}{8\beta}\frac{1}{\chi^2}, \tag{4.7}$$

one can write Eq. (4.6) as

$$\frac{d}{d\zeta}\left(\frac{1}{2}\dot{\chi}^2 + V(\chi)\right) = 0. \tag{4.8}$$

By integrating Eq. (4.8) twice it follows that

$$\int^{\zeta} d\zeta' = \int^{\chi}\frac{d\chi'}{\sqrt{2(E - V(\chi'))}}, \tag{4.9}$$

where E is a constant of first integration. The potential $V(\chi)$ given by (4.7) determines this solution in a characteristic way. To see this the following transformation of both the dependent and the independent variables

$$\tilde{\zeta} = \sqrt{\alpha/\beta}\,\zeta \tag{4.10}$$

and

$$y = \chi^2 + (1/6\alpha)\omega$$

leads to a standard form of Weierstrass elliptic functions

$$p^{-1}(y) = \int\frac{dy}{\sqrt{4y^3 - g_2 y - g_3}}, \tag{4.11}$$

where g_2 and g_3 are given by

$$g_2 = (3/16\alpha^2)\omega^2 - 8(E\beta/\alpha),$$

$$g_3 = \frac{4}{3}\frac{E\beta}{\alpha^2}\omega - \frac{I_1^2}{\alpha\beta} - \frac{1}{166\alpha^3}\omega^3.$$

If we now invert all our previously used transformations applied to ρ and ϕ, and to ζ as well we can write the solution for the density ρ and phase ϕ as

$$\rho(x,t) = p\{\sqrt{\alpha/\beta}\,[x - (H/C)t\,] - I_4\} - (1/6\alpha)\omega \tag{4.12}$$

and

$$\phi(x,t) = \frac{G}{C} + \frac{H}{C}\left(x - \frac{H}{C}t\right) + \int\frac{I_1\,d\zeta}{p(\sqrt{\alpha/\beta}\,\zeta - I_4) - (1/6\alpha)(\omega/6\alpha)}, \tag{4.13}$$

where I_1 and I_4 are arbitrary integration constants and ζ has the form of a "moving wave variable" $\zeta = x - (H/C)t$, i.e., the solutions represent traveling nonlinear waves.

It is well known that there exists a close connection between Weierstrass elliptic functions and Jacobi elliptic functions. To see how this relation appears in this system of equations we turn our attention back to Eq. (4.3). If we set there the arbitrary constant I_1 equal to zero, one can write a first integral of (4.1) as

$$F_1(\zeta) = (H/C)\zeta + \tilde{I}_1, \tag{4.14}$$

where $\tilde{I}_1$ is an arbitrary constant of integration. This form of $F_1(\zeta)$ solves (4.1) for any $F_2(\zeta)$. However, there exists on the other hand a strong coupling between F_1 and F_2 via (4.2). Inserting the solution (4.14) into Eq. (4.2) one can discard this coupling. The result is

$$F_2\ddot{F}_2 - \tfrac{1}{2}\dot{F}_2^2 - (\alpha/\beta)F_2^3 - (\omega/\beta)F_2^2 = 0, \tag{4.15}$$

which is a nonlinear ordinary differential equation for F_2 with constant coefficients. By means of the transformation (4.5), Eq. (4.15) is reduced to the form

$$\ddot{\chi} - (\alpha/2\beta)\chi^3 - (1/2\beta)\omega\chi = 0. \tag{4.16}$$

Introduction of the "potential" $V_1(\chi)$ results in the reduction to

$$\ddot{\chi} = -\frac{\partial V_1(\chi)}{\partial\chi}, \tag{4.17}$$

where $V_1(\chi)$ is a polynomial in χ of fourth order:

$$V_1(\chi) = -\frac{\alpha}{2\beta}\frac{1}{4}\chi^4 - \frac{1}{2\beta}\omega\frac{1}{2}\chi^2. \tag{4.18}$$

Here we dropped an arbitrary constant in V_1. We note that this potential is a special case of V resulting from (4.7) by taking $I_1 = 0$. An implicit solution for χ follows by integration of (4.17). Because $V_1(\chi)$ is a polynomial of fourth order this solution belongs to the class of Jacobi elliptic functions. The formal solution is

$$\zeta = \int^{\chi}\frac{d\chi'}{\sqrt{2(E - V(\chi'))}}, \tag{4.19}$$

where E is an integration parameter which strongly determines the behavior of this solution. For $\alpha < 0$, $\beta > 0$, and $\omega > 0$ integration of (4.19) can be carried out, with the result

$$\chi(\zeta) = b_1\,\mathrm{cn}\left(\sqrt{(|\alpha|/4\beta)(a_1^2 + b_1^2)}\,(\zeta - \zeta_0), m\right), \tag{4.20}$$

where cn is the cnoidal elliptic function. Here, a_1, b_1, and the modulus m are given by

$$a_1^2 = -\omega/|\alpha| + \sqrt{(\omega/|\alpha|)^2 + 8\beta E/|\alpha|}, \qquad b_1^2 = \omega/|\alpha| + \sqrt{(\omega/|\alpha|)^2 + 8\beta E/|\alpha|}, \tag{4.21}$$

$$m = b_1/\sqrt{a_1^2 + b_1^2}. \tag{4.22}$$

The initial phase ζ_0 can be expressed by an initial condition for χ:

$$\zeta_0 = \frac{1}{\sqrt{(|\alpha|/4\beta)(a_1^2 + b_1^2)}}\,\mathrm{cn}^{-1}(\chi(0)/b_1, m),$$

where $\chi(0)$ is the value of χ at $\zeta = 0$.

Elliptic functions such as those given in (4.20) belong to

the class of doubly periodic functions with $2K$, and K is the complete elliptic integral of the first kind. Another interesting property of this cnoidal function is the fact that cn tends to the hyperbolic function sech if the modulus m is equal to unity, i.e., $\mathrm{cn} \to \mathrm{sech}$ if $m \to 1$.

If we suppose that our group constants C,G,H are fixed and α, β are finite-valued constants then the modulus m is a function of E. Setting this parameter equal to zero m is equal to 1. In this way solution (4.20) skips to the hyperbolic function sech, as noted above. In detail, this solution is given by

$$\chi(\zeta) = (2\omega/|\alpha|)^{1/2} \operatorname{sech}(\sqrt{\omega/2\beta}\,(\zeta - \zeta_0)). \tag{4.23}$$

Inverting the previously used transformations we can give an explicit solution for the density and phase of the Schrödinger field ψ:

$$\rho(x,t) = \frac{2\omega}{|\alpha|} \operatorname{sech}^2\left(\sqrt{\frac{\omega}{2\beta}}\left(x - \frac{H}{C}t - \left(x_0 - \frac{H}{C}t_0\right)\right)\right) \tag{4.24}$$

and

$$\phi(x,t) = (H/C)[x - (H/C)t] + G/C + I_3, \tag{4.25a}$$

where

$$\omega = G/C - \tfrac{1}{2}(H/C)^2. \tag{4.25b}$$

A graphical representation of these solutions in space-time is given in Fig. 1 for ρ. One notes that the pulselike solution moves with constant velocity through space.

As stated in the Introduction, the Schrödinger field ψ is related to the density ρ and phase ϕ via the Madelung transformation (1.1), where $S(x,t) = m\phi(x,t)$. Inserting the density and phase into this transformation one finds

$$\psi(x,t) = \sqrt{\frac{2\omega}{|\alpha|}} \operatorname{sech}\left(\sqrt{\frac{\omega}{2\beta}}\left(x - \frac{H}{C}t - \left(x_0 - \frac{H}{C}t_0\right)\right)\right) \times \exp\left(-\frac{i}{\hbar}m\left(\frac{H}{C}\left(x - \frac{H}{C}t\right) + \frac{G}{C} + I_3\right)\right).$$

This representation of ψ has just the same form as Alonso's[25] result constructed by using IST techniques.

B. Connection to the Painlevé II function P_{II}

The second system of coupled nonlinear ordinary differential equations, which we investigate, includes the equations of subclasses (ii)(a) [Eqs. (3.11) and (3.12)] and the first equation of (ii)(b) [Eq. (3.13)] of Sec. III. This system has [in contrast to the previously discussed equations (4.1) and (4.2)] analytical coefficients. It reads as

$$\ddot{F}_1 F_2 + \dot{F}_2(\dot{F}_1 - H/C) = 0 \tag{4.26a}$$

and

$$F_2\ddot{F}_2 - \frac{1}{2}\dot{F}_2^2 - \frac{\alpha}{\beta}F_2^3 + F_2^2 \times\left\{\frac{H}{\beta C}\dot{F}_1 - \frac{1}{2\beta}\dot{F}_1^2 - \frac{1}{\beta}\left\{\gamma\zeta + \frac{G}{C}\right\}\right\} = 0, \tag{4.26b}$$

where γ is given by the relation

$$\gamma = \begin{cases} \gamma_1 = F/C, & \text{for } f_0 = 0 \text{ and } (0,C,F,G,H), \\ \gamma_2 = F/C - f_0, & \text{for } f_0 \neq 0 \text{ and } (0,C,F,G,H), \\ \gamma_3 = -f_0, & \text{for } f_0 \neq 0 \text{ and } (0,C,0,G,H). \end{cases} \tag{4.26c}$$

The independent variable ζ takes the form

$$\begin{aligned} \zeta &= x - (F/2C)t^2 - (H/C)t, \quad \text{for } \gamma = \gamma_1, \\ \zeta &= x - (F/2C)t^2 - (H/C)t, \quad \text{for } \gamma = \gamma_2, \\ \zeta &= x - (H/C)t, \quad \text{for } \gamma = \gamma_3. \end{aligned} \tag{4.26d}$$

Taking γ equal to zero, i.e., either F or f_0 equal to zero, it follows the previously discussed ODE system. The main difference between the two systems is a ζ-dependent term connected with γ. Thus we have to solve a coupled system of equations that possesses a dependence on independent variables. To solve these equations we proceed in the same way as in Sec. IV A. First we integrate (4.26a) and obtain the function F_1 as an implicit solution of F_2:

$$F_1(\zeta) = \int_{\zeta_0}^{\zeta}\left(\frac{I_1}{F_2(\zeta')} + \frac{H}{C}\right)d\zeta'. \tag{4.27}$$

For $I_1 = 0$ it follows that

$$F_1(\zeta) = (H/C)\zeta + I_2. \tag{4.28}$$

Substituting this result in (4.26b) leads to

$$F_2\ddot{F}_2 - \tfrac{1}{2}\dot{F}_2^2 - (\alpha/\beta)F_2^3 - (1/\beta)F_2^2\{\gamma\zeta + \omega\} = 0, \tag{4.29}$$

where ω is defined in (4.25b). Transforming both the dependent and independent variables by

$$z = (2\beta/\gamma)^{1/2}(1/2\beta)\{\gamma\zeta + \omega\} \tag{4.30a}$$

and

$$\chi^2 = (\alpha/4\beta)(2\beta/\gamma)^{2/3}F_2 \tag{4.30b}$$

one obtains

$$\chi'' - 2\chi^3 - z\chi(z) = 0, \tag{4.31}$$

where primes denote differentiation with respect to z. This second-order nonlinear differential equation is a special case of the so-called Painlevé II type equation. As shown by Painlevé,[32] this type of equation is irreducible and possesses neither movable branch points nor essential singularities. On the other hand, this equation defines a new class of transcendental functions. We see (as a by-product) that—following Ablowitz *et al.*[9]—the Madelung fluid is of the IST type, which is not surprising since the corresponding system of nonlinear Schrödinger equations (1.2) is also IST integrable.

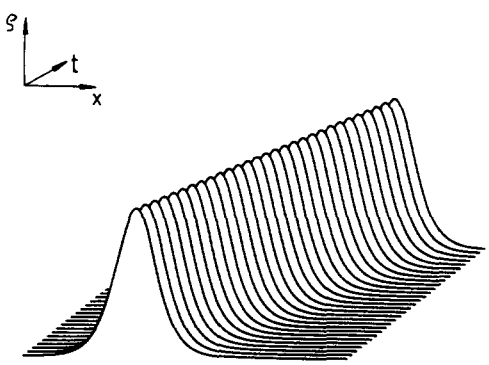

FIG. 1. Classical soliton solution (4.24) of the Madelung fluid for the group constants $(0,C,0,G,H)$ and for $f_0 = 0$.

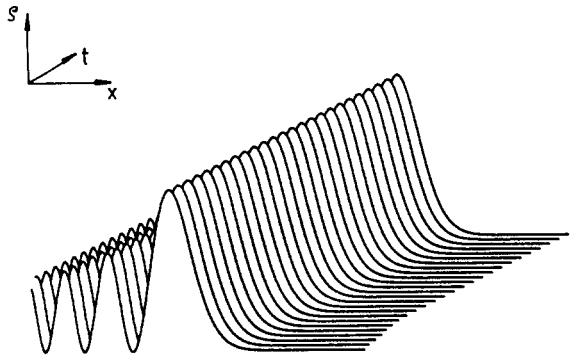

FIG. 2. Representation of the asymptotic solution (4.33) for small ρ with group constants (0,C,0,G,H) and external potential.

To follow our introductory proposals we now approximate (4.31) in such a way that at least an asymptotic analytical solution for this equation can be given. We require that this solution decay rapidly enough as $|x| \to \infty$ (say) that the integral of ρ is defined, i.e., $\rho \to 0$ for $|x| \to \infty$. As stated above, ρ is connected with χ by

$$\rho \propto \chi^2.$$

If we take the positive square root of ρ we can discuss the whole solution in terms of χ. Assuming now for the asymptotic solution that ρ is a small quantity, we can neglect products of χ in Eq. (4.31), which is then reduced to

$$\chi'' - z\chi(z) = 0. \tag{4.32}$$

Equation (4.32) is the definition equation of Airy functions in differential form.[33] The approximate solution for small ρ is therefore given by

$$\chi \propto \mathrm{Ai}(z). \tag{4.33}$$

The asymptotic solutions of Airy functions are[33]

$$\chi \propto (1/2\sqrt{\pi})(1/z^{1/4})\exp(-\tfrac{2}{3}(z)^{3/2}), \quad \text{for } z \to \infty,$$
$$\chi \propto (1/z^{1/4})\sin(\tfrac{2}{3}|z|^{3/2} + \phi_0), \quad \text{for } z \to -\infty.$$

Up to now we have solved a coupled nonlinear system of equations with an arbitrary independent variable ζ. These solutions are

$$\rho \propto P_{II}^2(\zeta),$$

$$\phi = \begin{cases} (3.5b) \\ (3.6b), \\ (3.7b) \end{cases} \quad \text{with } F_1 = (H/C)\zeta + I_2,$$

and ζ as in (4.26d). If we now specify the explicit combination of x and t in ζ we have to distinguish two cases strongly dependent on the values of the group constant F. Taking the choice $B = 0$ and $F = 0$ one gets an ordinary traveling wave with

$$\zeta = x - (H/C)t. \tag{4.34a}$$

The solution for the density $\rho(\zeta) = \rho(x,t)$ is shown in Fig. 2. The choice $F \neq 0$, i.e., regarding the more general subgroup (0,C,F,G,H), one obtains the similarity variable

$$\zeta = x - (F/2C)t^2 - (H/C)t. \tag{4.34b}$$

This is somewhat like an "accelerated wave variable." The corresponding solution $\rho(x,t)$ is plotted in Fig. 3. In contrast to the soliton- or solitary wave-type solutions based on ζ given in (4.34a), the solutions based on the similarity variable (4.34b) are called boomerons in the literature (see Ref. 34). Finally, we note that the soliton-type solution based on (4.34a) is, in addition to its Lie-group properties, Galilei invariant as well, while the boomeron solution based on (4.34b) is not invariant with respect to a Galilei transformation.

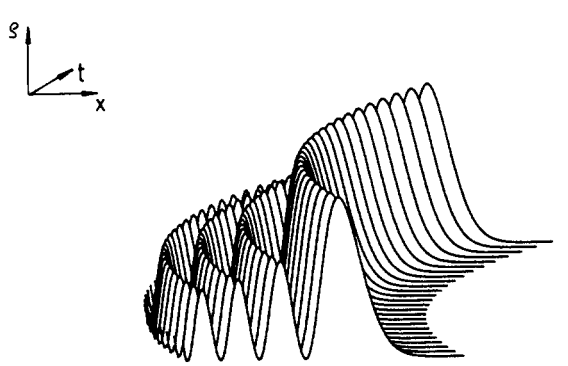

FIG. 3. Asymptotic solution for small ρ with (0,C,F,G,H) and $\gamma = \gamma_1$ or $\gamma = \gamma_2$.

[1] S. Lie, *Theorie der Transformationsgruppen* (Teubner, Leipzig, 1980), 2nd ed. (reprinted by Chelsea, New York, 1970).
[2] C. S. Gardener, J. M. Green, M. D. Kruskal, and R. M. Miura, Commun. Pure Appl. Math. **27**, 97 (1974).
[3] P. D. Lax, Commun. Pure Appl. Math. **21**, 467 (1968).
[4] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **LIII**, 240 (1974).
[5] V. E. Zakharov and A. B. Shabat, Sov. Phys. JETP **34**, 62 (1972).
[6] M. D. Arthur, K. M. Case, J. Math. Phys. **23**, 1771 (1982).
[7] R. K. Bullough and P. J. Caudry, *Solitons* (Springer, New York, 1980).
[8] M. J. Ablowitz, A. Ramani, and H. Segur, J. Math. Phys. **21**, 715 (1980).
[9] M. J. Ablowitz, A. Ramani, and H. Segur, J. Math. Phys. **21**, 1014 (1980).
[10] A. S. Fokas and M. J. Ablowitz, J. Math. Phys. **23**, 2033 (1982).
[11] G. W. Bluman and J. D. Cole, Appl. Math. Sci. **13**, 1 (1974).
[12] G. Tenti and W. H. Hui, J. Math. Phys. **19**, 744 (1978).
[13] T. F. Nonnenmacher, J. Appl. Math. Phys. **35**, 680 (1984).
[14] T. F. Nonnenmacher and G. Dukek, Physica **135A**, 167 (1986).
[15] G. Dukek and T. F. Nonnenmacher, in *Applications of Mathematics in Technology*, edited by V. C. Boffi and H. Neunzert (Teubner, Stuttgart, 1984).
[16] V. C. Boffi and T. F. Nonnenmacher, Nuovo Cimento B **85**, 165 (1985).
[17] G. Spiga, T. F. Nonnenmacher and V. C. Boffi, Physica **131A**, 431 (1985).
[18] A. V. Bobylev, Sov. Phys. Dokl. **20**, 820 (1976).
[19] M. Krook and T. T. Wu, Phys. Rev. Lett. **36**, 1107 (1976).
[20] M. Lakshmanan and P. Kaliappan, J. Math. Phys. **24**, 795 (1983).
[21] E. Madelung, Z. Phys. **40**, 322 (1926).
[22] T. F. Nonnenmacher, Lect. Notes Phys. **253**, 149 (1986).
[23] T. F. Nonnenmacher, G. Dukek and G. Baumann, Lett. Nuovo Cimento **36**, 453 (1983).
[24] F. Guerra and R. Marra, Phys. Rev. D **28**, 1916 (1983).
[25] L. M. Alonso, J. Math. Phys. **23**, 1518 (1982).
[26] L. J. F. Broer, Physica **76**, 364 (1974).
[27] C. A. Jones and P. H. Roberts, J. Phys. A: Math. Gen. **15**, 2599 (1982).
[28] S. J. Putterman and P. H. Roberts, Physica **117A**, 369 (1983).
[29] A. J. Purcell, Phys. Rev. D **30**, 2128 (1984).
[30] T. F. Nonnenmacher and J. D. F. Nonnenmacher, Lett. Nuovo Cimento **37**, 241 (1983).

[31]T. A. Minelli and A. Pascolini, Nuovo Cimento **85B**, 1 (1985).
[32]C. R. Painlevé, Acta Math. **25**, 13 (1902).
[33]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).
[34]A. Degasperis, in *Nonlinear Evolution Equations Solvable by the Spectral Transform*, edited by F. Calogero (Pitman, London, 1978).

Some aspects of the isogroup of the self-dual Yang–Mills system

C. J. Papachristou and B. Kent Harrison
Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602

(Received 8 October 1986; accepted for publication 4 February 1987)

A generalized isovector formalism is used to derive the isovectors and isogroup of the self-dual Yang–Mills (SDYM) equation in the so-called J formulation. In particular, the infinitesimal "hidden symmetry" transformation, a linear system, and a well-known Bäcklund transformation of the SDYM equation are derived in the process. Thus symmetry and integrability aspects of the SDYM system appear in natural relationship to each other within the framework of the isovector approach.

I. INTRODUCTION

In a recent paper[1] the authors discussed the application of isovector techniques[2,3] to systems of partial differential equations corresponding to exterior equations for vector-valued (and, in particular, matrix-valued) differential forms. It was seen that the application of the Lie derivative operator on vector-valued one-forms presents some technical difficulties, and for this reason an *internal exterior derivative* (i.e., an exterior derivative that acts on the fields but not on the variables of the solution manifold) was introduced by the formula

$$\bar{d}F(x^\mu,\psi^i) \equiv dF - \partial_\mu F\, dx^\mu, \tag{1.1}$$

where F is any function of the scalar variables x^μ of the solution manifold and the vector-valued fields ψ^i. If the system of partial differential equations (PDE's) is of order 2 or higher, the variables ψ^i will comprise the dependent variables u^a of the PDE's and the derivatives, up to a certain degree, of the u^a with respect to the x^μ. Given that, in the absence of specific restrictions on the exterior differential forms that represent the system, the variables ψ^i are considered independent of each other (and of the x^μ), we conclude that the problem can be naturally formulated on a jet space with "mixed" (i.e., both scalar- and vector-valued) coordinates.

In the present paper the formalism developed in Ref. 1 is applied to the self-dual Yang–Mills (SDYM) equation in the so-called J formulation.[4] It is seen that the isovector method provides a natural framework for the unification of such distinct concepts as symmetry and integrability. The independence of the coordinates of the underlying jetlike space is important in this context, as the reader will realize. In Sec. II we calculate the isovectors of the SDYM system. These vector fields can be used to construct infinitesimal symmetries (both geometrical and internal) of the system, as discussed in Ref. 1. The above-mentioned independence of coordinates is used in Sec. III to rewrite certain symmetries in a form equivalent to the parametric infinitesimal transformation introduced in Ref. 5. (This transformation is related to the so-called hidden symmetry of the SDYM field.[6]) Remarkably, the process also yields a pair of linear "inverse scattering" equations, the integrability of which is equivalent to the SDYM equation, and the parameter of which is identical to that of the infinitesimal transformation mentioned above. Finally, the results of Secs. I–III are used in Sec. IV to derive Bäcklund transformations for the SDYM system. In particular, the process gives the parametric Bäcklund transformation proposed in Ref. 7.

II. ISOVECTORS OF THE SDYM SYSTEM

The SDYM equation in the J formulation is written as[4–7]

$$\partial_{\bar{y}}(J^{-1}\partial_y J) + \partial_{\bar{z}}(J^{-1}\partial_z J) = 0. \tag{2.1}$$

The complex coordinates y, z, $\bar{y}$, and $\bar{z}$ are related to the coordinates x_1, x_2, x_3, and x_4 of complexified Euclidean space by

$$\begin{aligned} &2^{1/2}y = x_1 + ix_2, \quad 2^{1/2}z = x_3 - ix_4, \\ &2^{1/2}\bar{y} = x_1 - ix_2, \quad 2^{1/2}\bar{z} = x_3 + ix_4. \end{aligned} \tag{2.2}$$

[Note that the pairs $(y,\bar{y})$ and $(z,\bar{z})$ involve elements that are complex-conjugately related in *real* Euclidean space.] For our purposes, J is assumed to be a nonsingular element of the algebra $\mathrm{gl}(N,C)$ in its defining representation.

Equation (2.1) can be rewritten as a set of first-order PDE's:

$$B^1_{\bar{y}} + B^2_{\bar{z}} = 0, \quad B^1 = J^{-1}J_y, \quad B^2 = J^{-1}J_z, \tag{2.3}$$

where a standard notation for partial derivatives has been used. We are thus led, in the spirit of Ref. 1, to define the following set of four-forms in seven variables:

$$\begin{aligned} \gamma_1 &= dy\, dz\, dB^1\, d\bar{z} + dy\, dz\, d\bar{y}\, dB^2, \\ \gamma_2 &= dJ\, dz\, d\bar{y}\, d\bar{z} - JB^1\, dy\, dz\, d\bar{y}\, d\bar{z}, \\ \gamma_3 &= dy\, dJ\, d\bar{y}\, d\bar{z} - JB^2\, dy\, dz\, d\bar{y}\, d\bar{z}. \end{aligned} \tag{2.4}$$

It is easily seen that the $d\gamma_k$ are in the ideal of the γ_k; thus this ideal is closed.

We now proceed to find the isovectors of the system. For this purpose we must expand the Lie derivative of each γ_i into a "linear" combination of all three γ_k. The expansion must be made consistently with the requirement that the Lie derivative preserve the tensorial character of each γ_i separately.

Now, from Eqs. (2.4) it can be seen that the four-forms γ_k have values in $\mathrm{gl}(N,C)$, which is closed under both addition and multiplication. This observation suggests the following expansion:

 0022-2488/87/061261-04$02.50

$$\underset{V}{\pounds}\gamma_i = b_i^k\gamma_k + \Lambda_i^k\gamma_k + \gamma_k M_i^k, \tag{2.5}$$

where the b_i^k are scalars, whereas the zero-forms Λ_i^k and M_i^k have values in $\mathrm{gl}(N,C)$.

The vector field V is defined on a jetlike space with "coordinates" $y, z, \bar{y}, \bar{z}, J, B^1$, and B^2. As argued in Ref. 1, V will have a formal representation,

$$V = D^1\frac{\partial}{\partial y} + D^2\frac{\partial}{\partial z} + D^3\frac{\partial}{\partial \bar{y}} + D^4\frac{\partial}{\partial \bar{z}} + G\frac{\partial}{\partial J} + A^1\frac{\partial}{\partial B^1} + A^2\frac{\partial}{\partial B^2}, \tag{2.6}$$

where the $D^1,\ldots,D^4$ are assumed to be scalar functions of $y, z, \bar{y}, \bar{z}$, while the G, A^1, A^2 are $\mathrm{gl}(N,C)$-valued functions of the above four variables and J, B^1, and B^2. As in Ref. 1, we seek vector fields V for which the coefficients of expansion in Eq. (2.5) depend only on $y, z, \bar{y}$, and $\bar{z}$.

Substituting Eqs. (2.4) and (2.6) into Eq. (2.5), and using Eq. (1.1) to write

$$\pounds\, dJ = dG = G_{,\mu}\, dy^\mu + \bar{d}G,$$
$$\pounds\, dB^k = dA^k = A^k_{,\mu}\, dy^\mu + \bar{d}A^k \quad (k = 1,2),$$

where the y^μ ($\mu = 1,\ldots,4$) denote the $y,\ldots,\bar{z}$, we obtain a set of three exterior equations for four-forms. By equating the coefficients of $dy\, dz\, d\bar{y}\, d\bar{z}$ on both sides of each exterior equation, the following set of PDE's is derived:

$$\begin{aligned}
&A^1_{\bar{y}} + A^2_{\bar{z}} = -(b_1^2 + \Lambda_1^2)JB^1 - (b_1^3 + \Lambda_1^3)JB^2 \\
&\qquad - JB^1M_1^2 - JB^2M_1^3, \\
&G_y - GB^1 - JA^1 - D^\mu_{,\mu}JB^1 \\
&\quad = -(b_2^2 + \Lambda_2^2)JB^1 - (b_2^3 + \Lambda_2^3)JB^2 \\
&\qquad - JB^1M_2^2 - JB^2M_2^3, \\
&G_z - GB^2 - JA^2 - D^\mu_{,\mu}JB^2 \\
&\quad = -(b_3^2 + \Lambda_3^2)JB^1 - (b_3^3 + \Lambda_3^3)JB^2 \\
&\qquad - JB^1M_3^2 - JB^2M_3^3,
\end{aligned} \tag{2.7}$$

where $D^\mu \equiv D^1,\ldots,D^4$.

We now put

$$\begin{aligned}
A^i &= \alpha^{ik}(y^\mu)B^k + \beta^i(y^\mu)J + \overline{A}^i(y^\mu, B^k, J), \\
G &= \delta^k(y^\mu)B^k + \epsilon(y^\mu)J + \overline{G}(y^\mu, B^k, J),
\end{aligned} \tag{2.8}$$

where the α^{ik}, β^i, δ^k, and ϵ are scalars. Then

$$\bar{d}A^i = \alpha^{ik}\, dB^k + \beta^i\, dJ + \overline{dA}^i,$$
$$\bar{d}G = \delta^k\, dB^k + \epsilon\, dJ + \overline{dG}.$$

We substitute these expressions into the expansion of Eq. (2.5) and equate coefficients of terms that are *scalar* multiples of similar $\mathrm{gl}(N,C)$-valued basis four-forms. There are 12 such basis four-forms; therefore we obtain a set of 36 equations (eight of which are trivial identities). These results can be summarized as follows:

$$\begin{aligned}
&\beta^1 = \beta^2 = \delta^1 = \delta^2 = 0; \quad \alpha^{12} = D^3_{\bar{z}}, \quad \alpha^{21} = D^4_{\bar{y}}; \\
&D^1_{\bar{y}} = D^1_{\bar{z}} = 0, \quad D^2_{\bar{y}} = D^2_{\bar{z}} = 0, \\
&D^3_y = D^3_z = 0, \quad D^4_y = D^4_z = 0; \\
&b_1^1 = D^1_y + D^2_z + D^3_{\bar{y}} + \alpha^{22} = D^1_y + D^2_z + D^4_{\bar{z}} + \alpha^{11}, \\
&b_2^2 = D^2_z + D^3_{\bar{y}} + D^4_{\bar{z}} + \epsilon, \quad b_3^3 = D^1_y + D^3_{\bar{y}} + D^4_{\bar{z}} + \epsilon, \\
&b_2^3 = -D^2_y, \quad b_3^2 = -D^1_z, \quad b_1^2 = b_1^3 = b_2^1 = b_3^1 = 0.
\end{aligned}$$

We notice, in particular, that the D^1 and D^2 depend only on y and z, while the D^3 and D^4 depend only on $\bar{y}$ and $\bar{z}$.

The remaining terms in the expansion of Eq. (2.5) are those that cannot be expressed as *scalar* multiples of basis four-forms [in the sense that the coefficients in these terms do not commute with the $\mathrm{gl}(N,C)$-valued basis four-forms]. Terms of this type can be divided into four kinds according to their dependence on the basis three-forms $dy\, dz\, d\bar{y}$, $dy\, dz\, d\bar{z}$, $dy\, d\bar{y}\, d\bar{z}$, or $dz\, d\bar{y}\, d\bar{z}$. The $\mathrm{gl}(N,C)$-valued coefficients of each of these basis three-forms must be equated in each of the three exterior equations; this process yields a set of 12 equations which can be divided into two general types:

$$\Lambda_i^k\, dY + (dY)M_i^k = 0, \quad i \neq k, \tag{2.9}$$

and

$$\overline{dH} = \Lambda_k^k\, dY + (dY)M_k^k, \tag{2.10}$$

where $Y \equiv B^1, B^2, J$ and $H \equiv A^1, A^2, G$. The variable Y, by assumption, does not commute with Λ_i^k and M_i^k. Thus Eq. (2.9) is satisfied only if $\Lambda_i^k = M_i^k = 0, i \neq k$. Also, given that, by definition of the internal exterior derivative and by assumption about the Λ_i^k and M_i^k, $dY = \bar{d}Y$, $\bar{d}\Lambda_k^k(y^\mu) = 0$, $\bar{d}M_k^k(y^\mu) = 0$, Eq. (2.10) can be integrated immediately:

$$\overline{H} = \Lambda_k^k Y + YM_k^k + h(y^\mu),$$

where $h(y^\mu)$ is an arbitrary function. Our results are explicitly stated as follows:

$$\begin{aligned}
&\Lambda_1^1 \equiv \Lambda^1(y^\mu), \quad M_1^1 \equiv M^1(y^\mu), \\
&\Lambda_2^2 = \Lambda_3^3 \equiv \Lambda^2(y^\mu), \quad M_2^2 = M_3^3 \equiv M^2(y^\mu), \\
&\Lambda_i^k = M_i^k = 0, \quad \text{for } i \neq k; \\
&\overline{A}^1 = \Lambda^1B^1 + B^1M^1 + h^1(y^\mu), \\
&\overline{A}^2 = \Lambda^1B^2 + B^2M^1 + h^2(y^\mu), \\
&\overline{G} = \Lambda^2J + JM^2 + g(y^\mu),
\end{aligned}$$

where the h^1, h^2, and g are arbitrary $\mathrm{gl}(N,C)$-valued functions.

Appropriate substitutions into Eqs. (2.8) will now give expressions for A^i and G, which can be substituted back into Eqs. (2.7). By using previous results, the coefficients b_i^k can be eliminated in favor of other quantities, while certain replacements can also be made with regard to the Λ_i^k and M_i^k. The result is a set of equalities between some kind of generalized "polynomial" expressions in the variables B^1, B^2, and J, with y^μ-dependent coefficients. The "constant" term in such a "polynomial" is a matrix function $F(y^\mu)$, while the other terms are of the following kinds: qB^k, qJ, qJB^k, QB^k, B^kQ, QJ, JQ, QJB^k, JQB^k, and JB^kQ, where $q(y^\mu)$ is a scalar function and $Q(y^\mu)$ is a $\mathrm{gl}(N,C)$-valued function. Equating coefficients of similar terms we obtain a set of partial differential and algebraic equations, which are not hard to solve. In particular, we find

$$-\Lambda^1 = M^1 = M^2 \equiv M(y,z), \quad \Lambda^2 \equiv \Lambda(\bar{y},\bar{z}),$$
$$h^1(y,z) = M_y, \quad h^2(y,z) = M_z, \quad g(y^\mu) = 0.$$

Equations (2.11) give the complete solution for the components of the isovector field V:

$$
\begin{aligned}
&D^1 = c_1 y + k_1 z + \alpha_1, \quad D^2 = k_2 y + c_2 z + \alpha_2,\\
&D^3 = (c_2 - c)\bar{y} - k_2 \bar{z} + \alpha_3,\\
&D^4 = -k_1 \bar{y} + (c_1 - c)\bar{z} + \alpha_4,\\
&A^1 = -c_1 B^1 - k_2 B^2 - [M(y,z), B^1] + M_y,\\
&A^2 = -k_1 B^1 - c_2 B^2 - [M(y,z), B^2] + M_z,\\
&G = \epsilon(\bar{y},\bar{z})J + \Lambda(\bar{y},\bar{z})J + JM(y,z),
\end{aligned}
\tag{2.11}
$$

where $c_1, c_2, k_1, k_2, c, \alpha_1, \ldots, \alpha_4$ are nine complex parameters, $\epsilon(\bar{y},\bar{z})$ is a scalar function, and $M(y,z)$ and $\Lambda(\bar{y},\bar{z})$ are $\mathrm{gl}(N,C)$-valued functions. From Eqs. (2.11) we can read off the infinitesimal operators P_k corresponding to the nine complex parameters (cf. Ref. 1) and we can show that they form the basis of a Lie algebra. In particular, the operators

$$P_{\alpha_\mu} = \frac{\partial}{\partial y^\mu}$$

and

$$P_{c_1} + P_{c_2} = y^\mu \frac{\partial}{\partial y^\mu} - B^k \frac{\partial}{\partial B^k}$$

represent translations and dilatations, respectively.

Following the discussion in Ref. 1, from Eq. (2.11) we can construct the following infinitesimal internal symmetry transformations:

$$
\begin{aligned}
&B^{1\prime} \simeq B^1 + [M(y,z), B^1] - M_y,\\
&B^{2\prime} \simeq B^2 + [M(y,z), B^2] - M_z,\\
&J' \simeq J - \epsilon(\bar{y},\bar{z})J - \Lambda(\bar{y},\bar{z})J - JM(y,z),
\end{aligned}
\tag{2.12}
$$

where the ϵ, M, and Λ are infinitesimal. The corresponding finite transformations are

$$
\begin{aligned}
&B^{1\prime} = UB^1U^{-1} + U\partial_y U^{-1},\\
&B^{2\prime} = UB^2U^{-1} + U\partial_z U^{-1},\\
&J' = \beta\bar{U} J U,
\end{aligned}
\tag{2.13}
$$

where

$$U(y,z) = \exp\{-M(y,z)\},$$
$$\bar{U}(\bar{y},\bar{z}) = \exp\{-\Lambda(\bar{y},\bar{z})\},$$

and

$$\beta(\bar{y},\bar{z}) = \exp\{-\epsilon(\bar{y},\bar{z})\}.$$

These are, of course, familiar symmetries of the SDYM system.

III. PARAMETRIC INFINITESIMAL TRANSFORMATION AND LINEAR SYSTEM

If we define a new function

$$\xi(y,z,\bar{y},\bar{z}) \equiv M(y,z) + \epsilon(\bar{y},\bar{z})1_N, \tag{3.1}$$

where 1_N denotes the N-dimensional unit matrix, then the infinitesimal transformations of Eq. (2.12) with $\Lambda(\bar{y},\bar{z}) = 0$ can be rewritten as

$$
\begin{aligned}
&\delta B^1 = [\xi(y^\mu), B^1] - \xi_y,\\
&\delta B^2 = [\xi(y^\mu), B^2] - \xi_z, \quad \delta J = -J\xi(y^\mu),
\end{aligned}
\tag{3.2}
$$

where $\delta B^k \simeq B^{k\prime} - B^k$ and $\delta J \simeq J' - J$. We wish to rewrite these symmetries without the restriction (3.1). It turns out that this is possible due to the independence of the coordinates of the underlying jetlike space. Of course, there is a price to be paid for such an adjustment. But this "price" is a most welcome one: Restriction (3.1) is replaced by a set of *linear* PDE's which, in the case of actual SDYM fields, lead to a linear system for the SDYM equation.

From Eq. (3.1) it is seen that $\xi(y^\mu)$ satisfies the PDE,

$$[\xi_{\bar{y}}, B^1] + [\xi_{\bar{z}}, B^2] - \xi_{y\bar{y}} - \xi_{z\bar{z}} = 0.$$

Given the independence of the y^μ and the B^k (this is the case as long as no restriction on the solution manifold is imposed), the above equation may be written as

$$\partial_{\bar{y}}([\xi, B^1] - \xi_y) + \partial_{\bar{z}}([\xi, B^2] - \xi_z) = 0. \tag{3.3}$$

This is satisfied if there exists a "potential" $\psi(y^\mu, B^k)$ such that

$$[\xi, B^1] - \xi_y = \lambda\psi_{\bar{z}}, \quad [\xi, B^2] - \xi_z = -\lambda\psi_{\bar{y}}, \tag{3.4}$$

where λ is an arbitrary complex parameter. We thus replace the system of Eqs. (3.1) and (3.2) by the following alternate one:

$$\delta B^1 = \lambda\psi_{\bar{z}}, \quad \delta B^2 = -\lambda\psi_{\bar{y}}, \quad \delta J = -J\xi(y^\mu), \tag{3.5}$$

where ψ and ξ satisfy the linear system (3.4). Note that Eqs. (3.4) and (3.5) become independent of Eqs. (3.1) and (3.2) upon restriction to the solution manifold, i.e., for actual SDYM fields.

Let us explore further the significance of Eqs. (3.4) for *actual* SDYM fields (in which case the B^k are *dependent upon* the y^μ). In particular, let us examine the ansatz $\psi(y^\mu) = \xi(y^\mu)$, all y^μ:

$$[\xi, B^1] - \xi_y = \lambda\xi_{\bar{z}}, \quad [\xi, B^2] - \xi_z = -\lambda\xi_{\bar{y}}. \tag{3.6}$$

The integrability criterion $\xi_{y\bar{z}} - \xi_{\bar{z}y} = 0$ yields Eq. (3.3), which, in combination with Eq. (3.6), gives

$$[\xi, B^2_y - B^1_z + [B^1, B^2] + \lambda(B^1_{\bar{y}} + B^2_{\bar{z}})] = 0.$$

We seek conditions for B^1 and B^2 in order that the above equality holds for all λ and independently of ξ. The following pair of PDE's must therefore be satisfied:

$$\partial_y B^2 - \partial_z B^1 + [B^1, B^2] = 0, \tag{3.7}$$

$$\partial_{\bar{y}} B^1 + \partial_{\bar{z}} B^2 = 0. \tag{3.8}$$

Equation (3.7) is a condition for zero curvature and implies that the B^1 and B^2 are pure gauges:

$$B^1 = J^{-1}\partial_y J, \quad B^2 = J^{-1}\partial_z J, \tag{3.9}$$

where J is a nonsingular $\mathrm{gl}(N,C)$ matrix. Then Eq. (3.8) becomes identical to the SDYM equation (2.1), of which Eq. (3.6) is seen to be a linear system.

We remark that our results are in agreement with those of Ref. 5 (although they are given in a slightly different form). The thing to notice is that these results were actually *derived* here, in a rather straightforward manner, by using the isovector technique.

IV. CONNECTION WITH BÄCKLUND TRANSFORMATIONS

By using the original definitions of B^1 and B^2 as given in Eqs. (2.3), the infinitesimal transformations of these quanti-

ties may be written, according to Eqs. (3.5) as

$$J'^{-1}J'_y - J^{-1}J_y = \lambda\psi_{\bar{z}}, \tag{4.1a}$$

$$J'^{-1}J'_z - J^{-1}J_z = -\lambda\psi_{\bar{y}}. \tag{4.1b}$$

Clearly, as J' approaches J, the $\psi_{\bar{y}}$ and $\psi_{\bar{z}}$ must approach zero. One way to achieve this is to put

$$\psi = \xi = 1 - J^{-1}J'. \tag{4.2}$$

Now, if the left-hand sides of Eqs. (4.1a) and (4.1b) are considered as *finite*, rather than infinitesimal differences, then Eqs. (4.1) and (4.2) constitute one possible form of the Bäcklund transformation (BT) proposed in Ref. 7. Alternatively, the infinitesimal parametric transformation (4.1) and (4.2) is also an infinitesimal BT. This was observed in Ref. 5, but we include it in the present discussion due to its direct (and quite interesting) relevance to the isovector method.

Incidentally, the transformation (3.1) and (3.2) is also an infinitesimal BT, with Eq. (3.1) being a sort of algebraic constraint. Indeed, putting $\xi = 1 - J^{-1}J'$ and introducing an arbitrary complex parameter μ, we write

$$J'^{-1}J'_y - J^{-1}J_y = \mu\{[J^{-1}J', J^{-1}J_y] - \partial_y(J^{-1}J')\}, \tag{4.3a}$$

$$J'^{-1}J'_z - J^{-1}J_z = \mu\{[J^{-1}J', J^{-1}J_z] - \partial_z(J^{-1}J')\}, \tag{4.3b}$$

$$J^{-1}J' = M(y,z) + \epsilon(\bar{y},\bar{z})1_N, \tag{4.3c}$$

where $M(y,z)$ is $\mathrm{gl}(N,C)$ valued and $\epsilon(\bar{y},\bar{z})$ is a scalar. Taking $\partial_{\bar{y}}$(4.3a) + $\partial_{\bar{z}}$(4.3b) and using (4.3c), we find

$$\{\partial_{\bar{y}}(J'^{-1}J'_y) + \partial_{\bar{z}}(J'^{-1}J'_z)\} - \{\partial_{\bar{y}}(J^{-1}J_y) + \partial_{\bar{z}}(J^{-1}J_z)\} = \mu[J^{-1}J', \partial_{\bar{y}}(J^{-1}J_y) + \partial_{\bar{z}}(J^{-1}J_z)],$$

according to which J' satisfies the SDYM equation (2.1) if J does. Note that the BT was constructed so as to yield the trivial solution $J' = J$ as a particular solution [this corresponds to $M = 0$ and $\epsilon = 1$ in the algebraic constraint (4.3c)].

ACKNOWLEDGMENT

We are indebted to Professor Yong-Shi Wu for very enlightening discussions and for providing us with useful references.

[1] C. J. Papachristou and B. K. Harrison, "Isogroups of differential ideals of vector-valued differential forms: Application to partial differential equations," in *Proceedings of the XV International Colloquium on Group Theoretical Methods in Physics*, edited by R. Gilmore (World Scientific, Singapore, 1987).

[2] B. K. Harrison and F. B. Estabrook, J. Math. Phys. **12**, 653 (1971).

[3] D. G. B. Edelen, *Applied Exterior Calculus* (Wiley, New York, 1985), Chap. 6.

[4] C. N. Yang, Phys. Rev. Lett. **38**, 1377 (1977); Y. Brihaye, D. B. Fairlie, J. Nuyts, and R. G. Yates, J. Math. Phys. **19**, 2528 (1978).

[5] L.-L. Chau, M. L. Ge, and Y. S. Wu, Phys. Rev. D **25**, 1086 (1982).

[6] L.-L. Chau and Y. S. Wu, Phys. Rev. D **26**, 3581 (1982); L.-L. Chau, M. L. Ge, A. Sinha, and Y. S. Wu, Phys. Lett. **121B**, 391 (1983).

[7] M. K. Prasad, A. Sinha, and L.-L. Chau Wang, Phys. Rev. Lett. **43**, 750 (1979).

The invariant density for a class of discrete-time maps involving an arbitrary monotonic function operator and an integer parameter

C. C. Grosjean
Seminarie voor Wiskundige Natuurkunde, Rijksuniversiteit-Gent, Krijgslaan 281, Gebouw S9, B-9000 Gent, Belgium

(Received 29 April 1986; accepted for publication 17 December 1986)

When x_t and x_{t+1} represent two random variables, each belonging to a real interval [0,1] and being related by a first-order difference equation $x_{t+1} = F(x_t)$, called a discrete-time map, the probability density distribution connected with x_t can be translated into that associated with x_{t+1}. This yields an evolution equation by means of which one can construct an infinite sequence $\{w_t(x) | t \in \mathbb{N}, x \in [0,1]\}$ starting from an integrable function $w_0(x)$ normalized to unity on [0,1]. The question of the convergence of the sequence toward a so-called invariant density function $w(x)$ as $t \to +\infty$ and the problem of finding this limit were examined by a number of authors, mostly studying isolated cases. The present paper solves the problem for a class of discrete-time maps characterized by $x_{t+1} = f(|\mathrm{sn}[l \,\mathrm{sn}^{-1} f^{-1}(x_t)]|)$, $l \in \{2,3,\ldots\}$, whereby f is a real, continuous, monotonically increasing function mapping [0,1] onto itself and sn is the usual symbol for the sinelike Jacobian elliptic function with modulus $k \in [0,1]$ (including the sine function). Convergence is proven under very general conditions on $w_0(x)$ and an explicit formula to calculate $w(x)$ is established. Some properties of $w(x)$ are discussed. A necessary and sufficient condition for the symmetry of $w(x)$ about $x = \frac{1}{2}$ is obtained and attention is also devoted to the inverse problem, leading to a reformulation of the discrete-time map of the type cited above which corresponds to a given invariant density. The examples of practical application considered here cover almost all special cases which were treated in the literature thus far, as well as new cases.

I. INTRODUCTION

In a number of articles,[1-9] most of them of recent date, the evolution of a normalized single-valued real function (usually a probability density distribution) defined on a finite real interval toward an invariant limit function under a given discrete-time mapping has been studied. Let x_t and x_{t+1} be two continuously varying random variables, each belonging, for instance, to [0,1] and related to one another by a first-order difference equation

$$x_{t+1} = F(x_t). \tag{1.1}$$

If $w_t(x_t)$ represents the probability density distribution associated with x_t, so that $w_t(x_t)|dx_t|$ is the elementary probability that the random variable takes a value between x_t and $x_t + dx_t$ (whereby dx_t may be either positive or negative), one can ask for the corresponding elementary probability $w_{t+1}(x_{t+1})|dx_{t+1}|$ associated with the variable x_{t+1} and its differential. The expression of w_{t+1} in terms of w_t and F ultimately leads to a transformation by means of which one can construct iteratively an infinite sequence of functions $\{w_t(x) | t \in \mathbb{N}, x \in [0,1]\}$ starting from a given initial function $w_0(x)$, with all functions normalized to unity when

$$\int_0^1 w_0(x)dx = 1. \tag{1.2}$$

Thus $w_0(x)$ is assumed to stem from a single-valued real function defined on [0,1], being continuous or maybe only piecewise continuous, but satisfying conditions such that its definite integral from 0 to 1 (possibly an improper integral) be convergent. Then, using a suitable proportionality factor, normalization to unity may be achieved, giving rise to $w_0(x)$. The question is, does the infinite sequence of w functions tend to a limit distribution when $t \to +\infty$ and if so, find

$$\lim_{t \to +\infty} w_t(x), \quad 0 \leqslant x \leqslant 1.$$

In Ref. 1, one finds the following particular example:

$$x_{t+1} = F(x_t) = \begin{cases} 2x_t, & 0 \leqslant x_t \leqslant \frac{1}{2}, \\ 2(1-x_t), & \frac{1}{2} \leqslant x_t \leqslant 1, \end{cases} \tag{1.3}$$

in which case the limit distribution is

$$w(x) = \lim_{t \to +\infty} w_t(x) = 1, \quad 0 \leqslant x \leqslant 1. \tag{1.4}$$

In Ref. 5, it is shown that to the cusp-shaped return map

$$x_{t+1} = 1 - 2\sqrt{|x_t|}, \quad -1 \leqslant x_t \leqslant 1, \tag{1.5}$$

by means of which the interval $[-1,1]$ is mapped onto itself, there corresponds as a limit distribution

$$w(x) = \tfrac{1}{2}(1-x), \quad -1 \leqslant x \leqslant 1. \tag{1.6}$$

We shall reconsider these examples in Sec. III.

Ulam and von Neumann's paper[2] comprises still another example which has received considerable attention as of late. Under the discrete-time quadratic map (a special case of the so-called logistic map),

$$x_{t+1} = 4x_t(1-x_t), \tag{1.7}$$

of the interval [0,1] onto itself, any normalizable initial function $w_0(x)$ leads to the invariant density

$$w(x) = 1/\pi[x(1-x)]^{1/2}, \quad 0 \leqslant x \leqslant 1. \tag{1.8}$$

Indeed, Falk[6] has calculated the general element $w_t(x)$ of the sequence $\{w_t(x) | t \in \mathbb{N}, x \in [0,1]\}$ generated under (1.7) starting from the uniform distribution $w_0(x) = 1$ and obtained (1.8) by direct transition to the limit. Falk notes that

 0022-2488/87/061265-10$02.50

the symmetry of $w_0(x)$ about $x=\frac{1}{2}$ is broken at $t=1,2,3,...$ and returns only in the limit $t=+\infty$. Nandakumaran[7] has extended Falk's results to all probability density distributions of the form

$$w_0(x)=[x^n(1-x)^n]/[B(n+1,n+1)], \quad n\in\mathbb{N}, \quad 0\leqslant x\leqslant 1,$$

in particular confirming (1.8). Very recently, Grosjean[8,9] generalized Nandakumaran's work to a broad class of initial functions, normalized to unity but not necessarily positive-semidefinite, first under the quadratic mapping (1.7) and later under the polynomial discrete-time maps of any degree which generalize (1.7). Such mappings are first-order difference equations of the form

$$x_{t+1}=p_l(x_t), \quad l\in\{3,4,...\}, \tag{1.9}$$

in which p_l represents a real polynomial of degree l and are uniquely defined by the following requirements.

(1) For every $x_t\in[0,1]$, the corresponding x_{t+1} also belongs to $[0,1]$ where, in particular, $x_t=0$ entails $x_{t+1}=0$.

(2) To every $x_{t+1}\in]0,1[$, there correspond l distinct real values of x_t belonging to $]0,1[$.

In this way, the explicit expression of (1.9), with (1.7) included, is

$$x_{t+1}=x_t U_{l-1}^2\left(\sqrt{1-x_t}\right), \quad l\in\{2,3,...\}, \tag{1.10}$$

where U symbolizes a Chebyshev polynomial of the second kind, or in parametric form,

$$x_t=(\sin\theta)^2, \quad x_{t+1}=(\sin l\theta)^2, \quad 0\leqslant\theta\leqslant\pi/2, \quad l\in\{2,3,...\}. \tag{1.11}$$

Thinking of initial functions which can possibly be only piecewise continuous by the appearance of a number of finite jumps in $[0,1]$, Grosjean[8,9] proposed to represent $w_0(x)$ by a convergent series of the type

$$w_0(x)=\frac{a_0}{2}+\sum_{n=1}^{+\infty}[a_n T_{2n}(1-2x) + 2b_n x^{1/2}(1-x)^{1/2}U_{2n-1}(1-2x)], \tag{1.12}$$

deduced from the Fourier series

$$\frac{a_0}{2}+\sum_{n=1}^{+\infty}(a_n\cos 4n\theta+b_n\sin 4n\theta), \quad 0\leqslant\theta\leqslant\pi/2, \tag{1.13}$$

by means of the substitution $x=(\sin\theta)^2$ and normalized to unity [see (1.2)] when

$$\frac{a_0}{2}-\sum_{n=1}^{+\infty}\frac{a_n}{4n^2-1}=1. \tag{1.14}$$

Grosjean[8,9] established the following (provisional) theorem.

If in $[0,1]$ the initial function $w_0(x)$ is the sum of a convergent series of the type (1.12) whereby the a coefficients satisfy (1.14) as well as the condition

$$\frac{|a_0|}{2}+\sum_{n=1}^{+\infty}|a_n|: \text{ convergent}, \tag{1.15}$$

then under any polynomial discrete-time map comprised in (1.10), the sequence $\{w_t(x)|t\in\mathbb{N}, x\in[0,1]\}$ converges toward the limit function

$$w(x)=1/\pi[x(1-x)]^{1/2}, \quad 0\leqslant x\leqslant 1.$$

As in Refs. 6 and 7, the method used consisted of calculating $w_t(x)$ explicitly for any $t\in\mathbb{N}_0$ and letting t approach infinity. But in the final step of the proof, in the case that the "a part" in (1.12) involves infinitely many terms, there is need for uniform convergence with respect to t and thus (1.15) is required as a *sufficient* condition. Unfortunately, when (1.15) holds, it deprives the part of $w_0(x)$ which is symmetric about $x=\frac{1}{2}$ of the possibility of exhibiting finite jumps between 0 and 1, on account of the criterion of Weierstrass for the absolute and uniform convergence of the "a part" in a series expansion such as (1.13) and *a fortiori* (1.12). In a subsequent note, Barbour[10] succeeded in eliminating this drawback.

The purpose of the present paper is twofold: We wish to make the conditions imposed upon the initial function $w_0(x)$ much less restrictive and extend the theory to a much wider class of discrete-time maps [rational and irrational functions F in (1.1)]. The proof of the convergence of the sequence $\{w_t(x)|t\in\mathbb{N}, x\in[0,1]\}$ toward a limit density $w(x)$ as $t\to+\infty$ and finding this density will be based upon a procedure differing entirely from the methods applied in the various articles cited in this Introduction, including my own articles.[8,9]

II. THEORETICAL DEVELOPMENT

The parametric transformation $x_t=(\sin\theta)^2$ applied to the logistic map (1.7) is implicit in most studies of the map. As was shown in Sec. I, it has led me to the polynomial generalization of arbitrary degree (1.10) via (1.11). As a next step, inspired by the parametric representation (1.11), one can consider the much broader generalization

$$x_t=f(\sin\theta), \quad x_{t+1}=f(|\sin l\theta|), \quad 0\leqslant\theta\leqslant\pi/2, \quad l\in\{2,3,...\}, \tag{2.1}$$

where $f(y)$ is a single-valued, real, continuous function of y defined on $[0,1]$, having a continuous or piecewise continuous derivative in that interval and increasing monotonically (*sensu stricto*) from 0 to 1 as y increases from 0 to 1. The inverse function, which as a function of $y\in[0,1]$ we denote by $f^{-1}(y)$, is endowed with the same properties and therefore the discrete-time map corresponding to the propounded parametric representation (2.1) is

$$x_{t+1}=f(|\sin[l\arcsin f^{-1}(x_t)]|), \quad 0\leqslant x_t\leqslant 1, \quad l\in\{2,3,...\}. \tag{2.2}$$

But, the sine function is known to be a special case comprised in the Jacobian elliptic sn function. Although one encounters the Jacobian elliptic functions much less frequently than the circular functions in practice, another degree of generality is gained when (2.1) is replaced by

$$x_t=f(\mathrm{sn}(u,k)), \quad x_{t+1}=f(|\mathrm{sn}(lu,k)|), \quad 0\leqslant u\leqslant K(k), \quad l\in\{2,3,...\}, \tag{2.3}$$

where the modulus $k\in[0,1[$ and $K(k)$ is the complete elliptic

integral of the first kind, i.e.,

$$K(k)=\int_0^1 \frac{dx}{(1-x^2)^{1/2}(1-k^2x^2)^{1/2}},$$

whose value is real, finite, and greater than $\pi/2$ for $k\in]0,1[$. In the special case $k=0$, the Jacobian elliptic functions reduce to circular functions. From here onward, the modulus will not be written explicitly as an argument of the sn function and the complete elliptic integral except where it is desirable to include it in the notation. The counterpart of (2.2) is

$$x_{t+1}=f(|\mathrm{sn}[l\,\mathrm{sn}^{-1}f^{-1}(x_t)]|),\quad 0\leqslant x_t\leqslant 1,\quad l\in\{2,3,\ldots\},\tag{2.4}$$

with sn^{-1} the usual symbol for the inverse of the Jacobian sn function, defined on [0,1] by

$$\mathrm{sn}^{-1}v=\int_0^v \frac{dx}{(1-x^2)^{1/2}(1-k^2x^2)^{1/2}},\quad 0\leqslant v\leqslant 1.\tag{2.5}$$

Here, (2.4) is the definitive form of the discrete-time mappings upon which the remaining calculations in the present article will be based. It is sufficiently general to include a variety of interesting special cases, but it is not yet *the* most general form which can be considered.

Note that the formulas (2.2) and (2.4), each one connecting x_t and x_{t+1}, are closely related to the relevant concept of conjugate functions, i.e.,

$$b(x)=h\{a[h^{-1}(x)]\},$$

which has played an important role in a number of articles. One finds it, for instance, in articles by Halmos[11] and by Ulam,[12] and it was discussed by Grossmann and Thomae as well.[13] Also, Györgyi and Szépfalusy[14] made extensive use of the concept of conjugation to determine invariant measures for various one-dimensional maps. In the important special case of $l=2$, the absolute value stripes may be deleted in (2.2) and (2.4) under the assumptions made for the function $f(y)$ in (2.1). If a conjugating function $h(y)$ is such that $h[f(y)]$ is still a single-valued, real, continuous function of y on [0,1], having a continuous or piecewise continuous derivative in that interval and increasing monotonically (*sensu stricto*) from 0 to 1 as y increases from 0 to 1, then one can associate with (2.2), where one puts $l=2$, the discrete-time map

$$x_{t+1}=hf(\sin[2\arcsin f^{-1}h^{-1}(x_t)]);\tag{2.2'}$$

the analysis which will follow is also applicable to this map, with the function operator f simply replaced by hf. As is well known, the invariant densities corresponding to the two conjugate dynamical laws (2.2) with $l=2$ and (2.2') are closely related, which will be confirmed by the final result for $w(x)$. The same remark holds for the two conjugate dynamical laws (2.4) with $l=2$ and

$$x_{t+1}=hf(\mathrm{sn}[2\,\mathrm{sn}^{-1}f^{-1}h^{-1}(x_t)]).\tag{2.4'}$$

To every $x_t\in[0,1]$, there corresponds one real value of x_{t+1} and it also belongs to [0,1]. But, conversely, to every $x_{t+1}\in]0,1[$, there correspond l distinct real values of x_t in the interval [0,1], i.e.,

$$\begin{aligned}
x_t^{(1)}&=r_1(x_{t+1})\\
&=f[\mathrm{sn}((1/l)\mathrm{sn}^{-1}f^{-1}(x_{t+1}))],\\
x_t^{(2)}&=r_2(x_{t+1})\\
&=f[\mathrm{sn}((2K/l)-(1/l)\mathrm{sn}^{-1}f^{-1}(x_{t+1}))],\\
x_t^{(3)}&=r_3(x_{t+1})\\
&=f[\mathrm{sn}((2K/l)+(1/l)\mathrm{sn}^{-1}f^{-1}(x_{t+1}))],\\
&\vdots\\
x_t^{(l)}&=r_l(x_{t+1})=f[\mathrm{sn}((2[l/2]K/l)\\
&\quad-(-1)^l(1/l)\mathrm{sn}^{-1}f^{-1}(x_{t+1}))],
\end{aligned}\tag{2.6}$$

with $[l/2]$ the largest integer smaller than or equal to $l/2$. These expressions remain valid as $x_{t+1}\to 0$ and $x_{t+1}\to 1$ from inside [0,1]. From them, one easily deduces that $x_{t+1}=0$ gives rise to $x_t^{(1)}=0$, $x_t^{(2)}=x_t^{(3)}$, $x_t^{(4)}=x_t^{(5)}$, etc., whereas $x_{t+1}=1$ yields $x_t^{(1)}=x_t^{(2)}$, $x_t^{(3)}=x_t^{(4)}$, and so on. In these cases, the number of distinct real x_t values is reduced to either $[l/2]$ or $[l/2]+1$. As stated in the Introduction, the transformation leading from $w_t(x)$ to $w_{t+1}(x)$ stems from probability theory. If $w_t(x_t)|dx_t|$ represents the elementary probability that the random variable at the integer time instant t take on a value between x_t and x_t+dx_t (where dx_t can be either positive or negative), then we have, for the analogous elementary probability regarding the random variable at the time instant $t+1$,

$$\begin{aligned}
w_{t+1}(x_{t+1})|dx_{t+1}|&=\sum_{m=1}^{l} w_t(x_t^{(m)})|dx_t^{(m)}|\\
&=\sum_{m=1}^{l} w_t(r_m(x_{t+1}))|dr_m(x_{t+1})|.
\end{aligned}$$

Dividing by $|dx_{t+1}|$ and dropping the subscript in x_{t+1}, we obtain the evolution equation yielding $w_{t+1}(x)$ in terms of $w_t(x)$:

$$\begin{aligned}
w_{t+1}(x)=&\frac{df^{-1}(x)/dx}{l[1-(f^{-1}(x))^2]^{1/2}[1-k^2(f^{-1}(x))^2]^{1/2}}\left\{\mathrm{cn}\frac{v}{l}\,\mathrm{dn}\frac{v}{l}f'\left(\mathrm{sn}\frac{v}{l}\right)w_t\left[f\left(\mathrm{sn}\frac{v}{l}\right)\right]\right.\\
&+\mathrm{cn}\frac{2K-v}{l}\,\mathrm{dn}\frac{2K-v}{l}f'\left(\mathrm{sn}\frac{2K-v}{l}\right)w_t\left[f\left(\mathrm{sn}\frac{2K-v}{l}\right)\right]\\
&+\mathrm{cn}\frac{2K+v}{l}\,\mathrm{dn}\frac{2K+v}{l}f'\left(\mathrm{sn}\frac{2K+v}{l}\right)w_t\left[f\left(\mathrm{sn}\frac{2K+v}{l}\right)\right]+\cdots\\
&\left.+\mathrm{cn}\frac{2l'K-(-1)^lv}{l}\,\mathrm{dn}\frac{2l'K-(-1)^lv}{l}f'\left(\mathrm{sn}\frac{2l'K-(-1)^lv}{l}\right)w_t\left[f\left(\mathrm{sn}\frac{2l'K-(-1)^lv}{l}\right)\right]\right\},\quad \forall t\in\mathbb{N},
\end{aligned}\tag{2.7}$$

in which $v = \mathrm{sn}^{-1} f^{-1}(x), l' = [l/2]$, and cn, dn are the two well-known cosinelike Jacobian elliptic functions associated with the sn function. This being by definition the evolution equation connecting any two neighboring functions in the sequence $\{w_t(x) | t \in \mathbb{N}, x \in [0,1]\}$, it is not required after all that the w functions on which (2.7) is applied be positive semidefinite on $[0,1]$.

When (2.7) is used to transform a given initial function $w_0(x)$ into $w_1(x)$, we obtain

$$w_1(x) = \frac{df^{-1}(x)/dx}{l[1-(f^{-1}(x))^2]^{1/2}[1-k^2(f^{-1}(x))^2]^{1/2}} \times \sum_u \mathrm{cn}\, u\, \mathrm{dn}\, u f'(\mathrm{sn}\, u) w_0[f(\mathrm{sn}\, u)], \quad (2.8)$$

in which the summation with respect to u runs over the values

$$\frac{v}{l}, \frac{2K-v}{l}, \frac{2K+v}{l}, \ldots, \frac{2l'K-(-1)^l v}{l}, \quad (2.9)$$

where $v = \mathrm{sn}^{-1} f^{-1}(x)$. These are l distinct real values when $x \in]0,1[$, one in each of the open real intervals

$$]m(K/l),(m+1)(K/l)[, \quad \forall m \in \{0,1,\ldots,l-1\}.$$

When $x \to 0$ or $x \to 1$ from inside $[0,1]$, adjacent pairs of values in (2.9) converge to each other since $v \to 0$ or $v \to K$, respectively. The peculiarity of the evolution equation (2.7) is that when it is applied a second time in order to express $w_2(x)$ in terms of $w_0(x)$, one obtains

$$w_2(x) = \frac{df^{-1}(x)/dx}{l^2[1-(f^{-1}(x))^2]^{1/2}[1-k^2(f^{-1}(x))^2]^{1/2}} \times \sum_u \mathrm{cn}\, u\, \mathrm{dn}\, u f'(\mathrm{sn}\, u) w_0[f(\mathrm{sn}\, u)], \quad (2.10)$$

where the summation with respect to u now runs over the values

$$\frac{v}{l^2}, \frac{2K-v}{l^2}, \frac{2K+v}{l^2}, \ldots, \frac{2l'K-(-1)^l v}{l^2}, \frac{2(l-l')K+(-1)^l v}{l^2}, \ldots, \frac{2l^2-1}{2l^2}K + (-1)^l \times \left(\frac{K}{2l^2} - \frac{v}{l^2}\right), \quad \text{with} \quad v = \mathrm{sn}^{-1} f^{-1}(x), \quad (2.11)$$

which result from the replacement of v successively in each member of the sequence (2.9) by the entire set of l values contained in (2.9). When $x \in]0,1[$, this procedure yields l^2 distinct real values, one belonging to each of the open real intervals

$$]m(K/l^2), (m+1)(K/l^2)[, \quad \forall m \in \{0,1,\ldots,l^2-1\}.$$

The complete verification of the results (2.10) and (2.11) would require a considerable amount of space; therefore, it is left to the reader. Only as an illustrative example of how the combinations of arguments simplify, let us calculate the argument with which the function w_0 appears in the first term of the sum in (2.10). To do this, (2.7) is applied for $t = 1$; hence we find in the first term, among other factors,

$$w_1[f(\mathrm{sn}(\mathrm{sn}^{-1} f^{-1}(x)/l))],$$

and according to (2.8), the argument of this w_1 function should replace x in

$$w_0[f(\mathrm{sn}(\mathrm{sn}^{-1} f^{-1}(x)/l))].$$

This yields

$$w_0\{f[\mathrm{sn}(1/l)(\mathrm{sn}^{-1} f^{-1}[f(\mathrm{sn}(\mathrm{sn}^{-1} f^{-1}(x)/l))])]\},$$

which clearly reduces to

$$w_0[f(\mathrm{sn}(\mathrm{sn}^{-1} f^{-1}(x)/l^2))].$$

In the same manner, after t applications of Eq. (2.7), it follows that

$$w_t(x) = \frac{df^{-1}(x)/dx}{l^t[1-(f^{-1}(x))^2]^{1/2}[1-k^2(f^{-1}(x))^2]^{1/2}} \times \sum_u \mathrm{cn}\, u\, \mathrm{dn}\, u f'(\mathrm{sn}\, u) w_0[f(\mathrm{sn}\, u)], \quad (2.12)$$

in which, when $x \in]0,1[$, the summation with respect to u runs over l^t distinct real values resulting from repeated application of the rule described above:

$$\frac{v}{l^t}, \frac{2K-v}{l^t}, \frac{2K+v}{l^t}, \ldots, \frac{2l^t-1}{2l^t}K + (-1)^l\left(\frac{K}{2l^t} - \frac{v}{l^t}\right),$$

still with $v = \mathrm{sn}^{-1} f^{-1}(x)$. When the interval $[0,K]$ is subdivided into l^t equal subintervals, one finds inside each of these subintervals one u value taking part in the summation appearing in (2.12). Consequently, when $x \in]0,1[$, the expression

$$\sum_u \mathrm{cn}(u)\mathrm{dn}(u) f'(\mathrm{sn}\, u) w_0[f(\mathrm{sn}\, u)] \times \frac{K}{l^t} \quad (2.13)$$

recalls the way in which a Riemann definite integral is defined. Indeed, the real interval $[0,K]$ over which u can vary is subdivided into l^t equal subintervals playing the role of Δu, and inside each of these intervals is located one abscissa at which the value of the integrand is taken. With regard to $w_0(x)$, the simplest possible case is that this normalized initial function is continuous in $[0,1]$. Then, it is immediately clear that for $x \in]0,1[$ the sequence $\{w_t(x) | t \in \mathbb{N}\}$ converges toward the following limit as $t \to +\infty$:

$$\frac{df^{-1}(x)/dx}{K[1-(f^{-1}(x))^2]^{1/2}[1-k^2(f^{-1}(x))^2]^{1/2}} \times \int_0^K \mathrm{cn}\, u\, \mathrm{dn}\, u f'(\mathrm{sn}\, u) w_0[f(\mathrm{sn}\, u)]\, du. \quad (2.14)$$

But,

$$\int_0^K \mathrm{cn}\, u\, \mathrm{dn}\, u f'(\mathrm{sn}\, u) w_0[f(\mathrm{sn}\, u)]\, du = \int_0^1 f'(y) w_0[f(y)]\, dy = \int_0^1 w_0(x)\, dx = 1$$

in virtue of the normalization of $w_0(x)$, and hence $\{w_t(x) | t \in \mathbb{N}\}$ approaches

$$w(x) = \frac{df^{-1}(x)/dx}{K[1-(f^{-1}(x))^2]^{1/2}[1-k^2(f^{-1}(x))^2]^{1/2}}. \quad (2.15)$$

When $x=0$, the u values over which the summation in (2.12) runs are zero and all integer multiples of $2K/l'$ smaller than or equal to K, because, as one lets $x\to+0$, all or almost all pairs of adjacent u values from the second u value onward tend to a common limit. When l is even, one can subdivide $[0,K]$ as follows:

$[0,K/l']$ associated with $u=0$,

$[(2m-1)(K/l'),(2m+1)(K/l')]$
associated with $u=2mK/l'$,
$\forall m\in\{1,2,\ldots,(l'-2)/2\}$,

$[(1-1/l')K,K]$ associated with $u=K$.

Here, every u value between 0 and K is located in the middle of a segment of length $2K/l'$. The sum appearing in (2.12) becomes

$$2\left\{\frac{1}{2}f'(0)w_0(0)+\sum_{m=1}^{(l'-2)/2}\operatorname{cn}\frac{2mK}{l'}\operatorname{dn}\frac{2mK}{l'}\times f'\left(\operatorname{sn}\frac{2mK}{l'}\right)w_0\left[f\left(\operatorname{sn}\frac{2mK}{l'}\right)\right]+\frac{1}{2}(1-k^2)^{1/2}[\operatorname{cn}uf'(\operatorname{sn}u)]_{u=K}w_0(1)\right\},$$

and with $w_0(x)$ still assumed to be continuous in $[0,1]$, this expression, when multiplied by K/l', again converges toward the Riemann definite integral appearing in (2.14) when $t\to+\infty$. Hence, when l is even, (2.15) remains valid in the limit $x=0$. A similar reasoning can be formulated when l is odd, and after that also for $x=1$. Therefore, (2.15) holds good for all $x\in[0,1]$ under the assumption that the initial function $w_0(x)$ is continuous in the interval $[0,1]$.

Our argument can easily be extended to other cases, where the given initial function $w_0(x)$ involves discontinuities in $[0,1]$. When $w_0(x)$ is piecewise continuous, having a limited number of finite jumps in $[0,1]$, there corresponds to every abscissa where such a jump occurs a u value belonging to $[0,K]$, obtained by the transformation $u=\operatorname{sn}^{-1}f^{-1}(x)$. These u values subdivide $[0,K]$ into a finite number of subintervals in each of which $w_0[f(\operatorname{sn}u)]$ is continuous, and for sufficiently large t, when $x\in]0,1[$, every one of these subintervals is itself subdivided into a number of segments of equal length K/l', except perhaps near the boundary points. Letting $t\to+\infty$, the above reasoning can be applied to each of the subintervals; thus one obtains, instead of one Riemann integral like the one in (2.14), a finite sum of Riemann integrals whose intervals of integration are strung together to form $[0,K]$. This sum of Riemann integrals defines the value of

$$\int_0^1 w_0(x)dx \tag{2.16}$$

in the considered case and is therefore equal to unity in virtue of the normalization. Once more, this argument can be extended without difficulty to $x=0$ and $x=1$. Consequently, (2.15) remains valid when $w_0(x)$ is piecewise continuous, exhibiting a limited number of finite jumps in $[0,1]$. Analogous considerations also hold in cases where there occur stronger singularities of $w_0(x)$ in $[0,1]$, making (2.16) an improper integral on the condition, however, that the integral be convergent so that $w_0(x)$ is normalizable in the strict sense. An example of such a case is

$$w_0(x)=1/(2|1-2x|^{1/2}),\quad 0\leqslant x\leqslant 1. \tag{2.17}$$

Here, use can be made of suitable cutoffs and one can show that $w_t(x)$, as given by (2.12), converges toward the right-hand side of (2.15), multiplied by the definition of (2.16) as a convergent improper integral; thus, e.g., in the case of (2.17),

$$\int_0^1 w_0(x)dx=\lim_{\epsilon\to+0}\frac{1}{2}\int_0^{\frac{1}{2}-\epsilon}\frac{dx}{(1-2x)^{1/2}}+\lim_{\epsilon'\to+0}\frac{1}{2}\int_{\frac{1}{2}+\epsilon'}^1\frac{dx}{(2x-1)^{1/2}}=1.$$

In conclusion, the following theorem can be stated.

Under a discrete-time mapping of the type (2.4) with $k\in[0,1[$, subjecting iteratively to the associated transformation (2.7), a single-valued, real, continuous, or piecewise continuous function $w_0(x)$, defined on $[0,1]$ and normalized to unity, where

$$\int_0^1 w_0(x)dx$$

may be an improper, yet convergent integral on account of the singularities of $w_0(x)$ in $[0,1]$, leads to an infinite sequence of functions $\{w_t(x)\,|\,t\in\mathbb{N},\ x\in[0,1]\}$ which converges toward a limit function $w(x)$ when $t\to+\infty$, with

$$w(x)=\frac{df^{-1}(x)/dx}{K(k)[1-(f^{-1}(x))^2]^{1/2}[1-k^2(f^{-1}(x))^2]^{1/2}},\quad 0\leqslant x\leqslant 1. \tag{2.18}$$

Note that the limit function is independent of the integer parameter l appearing in (2.4), but fully determined by the modulus k and the inverse of the function f appearing in the parametric representation (2.3). For $k=0$, which means using (2.1) and (2.2), (2.18) reduces to

$$w(x)=\frac{2\,df^{-1}(x)/dx}{\pi[1-(f^{-1}(x))^2]^{1/2}},\quad 0\leqslant x\leqslant 1. \tag{2.19}$$

Since (2.7) conserves normalization, $w(x)$ given by (2.18) or by (2.19) is also normalized to unity, as can be verified by inspection.

In regard to the preceding theorem, let me emphasize that the initial function $w_0(x)$ must no longer be continuous on $[0,1]$ in order to ensure convergence of the sequence $\{w_t(x)\,|\,t\in\mathbb{N},x\in[0,1]\}$ toward the limit (2.18), in contrast to the continuity imposed on $w_0(x)$ by the sufficient condition (1.15) which was part of the theorems comprised in my earlier articles.[8,9] In those theorems, leaving out the condition (1.15) would not exclude the possibility of convergence toward the limit function $\pi^{-1}[x(1-x)]^{-1/2}$, but convergence would be unproved when $w_0(x)$ involves singularities in $[0,1]$ *within the framework of my earlier proofs*. The present analysis clearly removes this drawback. It shows that in virtue of the more powerful method used in the present article, condition (1.15) can be deleted after all, in agreement with Barbour's result[10] in the case of finite jumps of $w_0(x)$ in $[0,1]$.

Reconsidering the two conjugate dynamical laws (2.2)

with $l=2$ and (2.2') where the conjugating function h is such that both f and hf have the properties required in the beginning of this section, we clearly have as a relation between the corresponding invariant densities $w(x)$ and $w(x;h)$,

$$w(x;h)=w[h^{-1}(x)]\frac{dh^{-1}(x)}{dx},$$

a relation which remains valid for (2.4) with $l=2$ and (2.4'). The examples given in Ref. 12 (cf. Sec. IV and Figs. 2 and 3) are comprised in the present theory. Indeed, the different conjugating functions $h(\theta)$ which are coupled with the special case called $N_2(\theta)$ have the properties specified in the beginning of our Sec. II; therefore this is also the case with $h[f(y)]$. Here, $N_2(\theta)$ is such that

$$x_{t+1}=N_2(x_t)$$

is nothing but the dynamical law (1.3). As is shown in example (5) of Sec. III, this map is included in (2.2), where one puts $l=2$ and the functions which I called $f(y)$ and $f^{-1}(y)$ are

$$f(y)=(2/\pi)\arcsin y,\quad f^{-1}(y)=\sin(\pi y/2),\quad 0\leqslant y\leqslant 1.$$

In certain applications, as we shall see in Sec. III, $w(x)$ turns out to be symmetric about $x=\frac{1}{2}$. This is not necessarily a consequence of symmetry about $x_t=\frac{1}{2}$ in the right-hand side of a discrete-time map written in the form (1.1). Let us, for instance, recall the case of the quadratic map (1.7) where, as was pointed out by Falk,[6] $w_0(x)=1$ gives rise to the asymmetric functions $w_t(x)$, $t=1,2,\ldots$, despite the symmetry of the right-hand side in (1.7). However, the symmetry about $x=\frac{1}{2}$ is restored in the limit function $w(x)$ given by (1.8). In the case of $l=3$ in (1.10), being the cubic map

$$x_{t+1}=x_t(3-4x_t)^2,\quad 0\leqslant x\leqslant 1, \tag{2.20}$$

which does not have the symmetry about $x_t=\frac{1}{2}$, starting from either a symmetric or an asymmetric initial function $w_0(x)$ leads, at any rate, to (1.8), having the symmetry. On the contrary, when one maps linearly $[-1,1]$ onto $[0,1]$ both for x_t and x_{t+1} in (1.5), so as to obtain

$$x_{t+1}=1-|1-2x_t|^{1/2},\quad 0\leqslant x_t\leqslant 1, \tag{2.21}$$

which is, of course, also cusp shaped with the singular point at $x_t=\frac{1}{2}$, the discrete-time map is symmetric about that point. The evolution corresponding to (2.21) is

$$w_{t+1}(x)=(1-x)[w_t(x-x^2/2)+w_t(1-x+x^2/2)],$$
$$0\leqslant x\leqslant 1.$$

Starting from the normalized uniform distribution $w_0(x)=1$ which also has the symmetry, one finds, strangely enough,

$$w_t(x)=2(1-x),\quad \forall t\in\mathbb{N}_0,\quad 0\leqslant x\leqslant 1,$$

and hence

$$w(x)=\lim_{t\to+\infty} w_t(x)=2(1-x), \tag{2.22}$$

which is *the* invariant density, no matter from which initial distribution one starts since it is, under the requirement of normalization to unity, the unique solution of

$$w(x)=(1-x)[w(x-x^2/2)+w(1-x+x^2/2)],$$
$$0\leqslant x\leqslant 1,$$

according to Hemmer.[5] Yet, it does not possess the symmetry about $x=\frac{1}{2}$. Irrespective of these examples, the explicit form of $w(x)$, namely (2.18), associated with a discrete-time map of the type (2.4), enables us to find a necessary and sufficient condition that $w(x)$ be symmetric about $x=\frac{1}{2}$. Indeed, if $w(x)=w(1-x)$, $\forall x\in[0,1]$, then

$$\int_0^y w(x)dx=\int_{1-y}^1 w(x')dx',\quad \forall y\in[0,1],$$

or

$$\mathrm{sn}^{-1}f^{-1}(y)=K(k)-\mathrm{sn}^{-1}f^{-1}(1-y),\quad \forall y\in[0,1].$$

If we put

$$\mathrm{sn}^{-1}f^{-1}(y)=u,\quad 0\leqslant u\leqslant K(k),$$

then

$$\mathrm{sn}\,u=f^{-1}(y),\quad \mathrm{sn}[K(k)-u]=f^{-1}(1-y).$$

But

$$\mathrm{sn}[K(k)-u]=(\mathrm{cn}\,u)/(\mathrm{dn}\,u),$$

and therefore, we find as a necessary and sufficient condition for symmetry,

$$f^{-1}(1-x)=\left(\frac{1-(f^{-1}(x))^2}{1-k^2(f^{-1}(x))^2}\right)^{1/2},\quad \forall x\in[0,1], \tag{2.23}$$

in which we returned to x as the independent variable. If one keeps in mind that $f^{-1}(x)$ increases steadily from 0 to 1 as x increases from 0 to 1, (2.23) can be rewritten rationally as

$$(f^{-1}(1-x))^2=[1-(f^{-1}(x))^2]/[1-k^2(f^{-1}(x))^2]. \tag{2.24}$$

For $k=0$, which corresponds to (2.2), the condition is

$$(f^{-1}(x))^2+(f^{-1}(1-x))^2=1. \tag{2.25}$$

Expressed in terms of f itself, (2.23) becomes

$$f(x)+f(\sqrt{(1-x^2)/(1-k^2x^2)})=1,$$
$$\forall x\in[0,1],\quad k\in[0,1[, \tag{2.26}$$

with, as a special case for $k=0$,

$$f(x)+f(\sqrt{1-x^2})=1,\quad \forall x\in[0,1]. \tag{2.27}$$

Examples of solutions of Eq. (2.27) which increase steadily from 0 to 1 as x increases from 0 to 1, are

$$f(x)=x^2\quad\text{and}\quad f(x)=(2/\pi)\arcsin x. \tag{2.28}$$

They will be brought in relation with certain special cases comprised in Sec. III. Note that in regard to possible symmetry of $w(x)$ about $x=\frac{1}{2}$, the integer parameter l in (2.2) or (2.4) is totally unimportant.

The explicit formula (2.18) enables us to discuss some characteristic features of the limit function $w(x)$. In virtue of the properties $f(x)$ is assumed to have for $x\in[0,1]$, we have

$$\frac{df^{-1}(x)}{dx}\geqslant 0,$$

entailing that

$$w(x)\geqslant 0,\quad 0\leqslant x\leqslant 1, \tag{2.29}$$

irrespective of whether the initial function $w_0(x)$ is positive

semidefinite or not. Furthermore, if

$$\left(\frac{df^{-1}(x)}{dx}\right)_{x=1} > 0,$$

which means $0 \leqslant f'(1) < +\infty$, then one certainly has $w(1) = +\infty$. If at the same time $f^{-1}(x)$ satisfies (2.24) or (2.25), one obviously also has $w(0) = +\infty$ in virtue of the symmetry of $w(x)$ about $x = \frac{1}{2}$. When $x \in [0,1[$, the denominators in (2.18) and (2.19) are finite and positive. To any abscissa $\alpha \in [0,1[$, where $f'(\alpha)$ happens to be zero, there corresponds $w(\beta) = +\infty$ on account of $(df^{-1}(x)/dx)_{x=\beta} = +\infty$, with $\beta = f(\alpha) \in [0,1[$. This can solely occur at isolated abscissas β. Furthermore, also for $x \in [0,1[$, the equality sign in (2.29) can only hold at points where $df^{-1}(x)/dx = 0$, which is associated with $f'(x)$ becoming infinite. Thus if $f(x)$ is a function with a continuous derivative in $[0,1[$, then $w(x) > 0$, $x \in [0,1[$. It could happen, however, that $f'(x)$ is only piecewise continuous in $[0,1[$. Then, for an abscissa $\kappa \in [0,1[$, where $f'(\kappa) = +\infty$, there comes $w(\lambda) = 0$ on account of $(df^{-1}(x)/dx)_{x=\lambda} = 0$, with $\lambda = f(\kappa)$. This can occur solely at isolated abscissas λ because $f(x)$ is assumed to be single valued. It could also happen, still when $f'(x)$ is only piecewise continuous in $[0,1[$, that $f'(x)$ is double valued at one or several abscissas μ, which would mean that $f(x)$ has a left derivative and a right derivative, unequal at such an abscissa. Then $df^{-1}(x)/dx$ and therefore $w(x)$ are also double valued at $x = \nu$, $\nu = f(\mu)$, with the discontinuity a finite or an infinite jump.

The particular form of the right-hand side in (2.18) makes it possible to establish a way to solve the inverse problem: Given a positive-semidefinite invariant density, find the discrete-time maps of the type (2.4) or (2.2) from which it originates. Indeed, integrating on both sides between 0 and y in (2.18), one obtains

$$\int_0^y w(x)dx = \frac{1}{K(k)} \operatorname{sn}^{-1} f^{-1}(y), \quad 0 \leqslant y \leqslant 1,$$

in virtue of (2.5), and therefore

$$f^{-1}(y) = \operatorname{sn}\left(K(k) \int_0^y w(x)dx\right), \quad 0 \leqslant y \leqslant 1. \tag{2.30}$$

Then f clearly follows from the inversion of f^{-1}. This procedure will be applied in two of the examples considered in Sec. III. Equation (2.30) even permits the discrete-time map (2.4), which yields $w(x)$, to be rewritten solely in terms of $w(x)$:

$$\operatorname{sn}\left(K(k) \int_0^{x_{t+1}} w(x)dx\right) = \left|\operatorname{sn}\left(lK(k) \int_0^{x_t} w(x)dx\right)\right|,$$
$$0 \leqslant x_t \leqslant 1, \quad 0 \leqslant x_{t+1} \leqslant 1, \quad l \in \{2,3,...\}, \quad k \in [0,1]. \tag{2.31}$$

In the special case $k = 0$, the corresponding formulas are

$$f^{-1}(y) = \sin\left(\frac{\pi}{2} \int_0^y w(x)dx\right), \tag{2.32}$$

$$\sin\left(\frac{\pi}{2} \int_0^{x_{t+1}} w(x)dx\right) = \left|\sin\left(\frac{l\pi}{2} \int_0^{x_t} w(x)dx\right)\right|. \tag{2.33}$$

Expressed in this manner, a discrete-time map is in an implicit form, i.e., solved neither with respect to x_{t+1} nor with respect to x_t.

III. SOME EXAMPLES OF PRACTICAL APPLICATION

(1) In the case of a polynomial discrete-time map as described by (1.10) or (1.11), we have $k = 0$ and $f(x) = x^2$ with $x \in [0,1]$. Hence, $K = \pi/2$ and $f^{-1}(x) = \sqrt{x}$, and consequently there is convergence of $\{w_t(x) | t \in \mathbb{N}\}$ toward

$$w(x) = \frac{2\,d\sqrt{x}/dx}{\pi(1-x)^{1/2}} = \frac{1}{\pi[x(1-x)]^{1/2}}, \quad 0 \leqslant x \leqslant 1, \tag{3.1}$$

which confirms the result (1.8) obtained by several authors in the special case of the quadratic map. On the basis of Sec. II, one understands why all the polynomial discrete-time mappings comprised in (1.10) give rise to the same invariant density. Since $f'(1) = 2$, thus positive, $w(1) = +\infty$. Since $f'(0) = 0$, we have $(df^{-1}(x)/dx)_{x=0} = +\infty$ and consequently $w(0) = +\infty$. Because of the continuity of $f'(x)$ in $[0,1]$, $w(x) > 0$ for all $x \in [0,1]$. The function $f(x) = x^2$ is a solution of (2.27), which explains the symmetry of (3.1) about $x = \frac{1}{2}$.

(2) The most direct generalization of (1.11) is

$$x_t = (\operatorname{sn}(u,k))^2, \quad x_{t+1} = (\operatorname{sn}(lu,k))^2, \quad 0 \leqslant u \leqslant K(k),$$
$$l \in \{2,3,...\}, \quad 0 < k < 1. \tag{3.2}$$

In this case, the discrete-time map is rational. When $l = 2$, it is

$$x_{t+1} = [4x_t(1-x_t)(1-k^2x_t)]/(1-k^2x_t^2)^2, \tag{3.3}$$

being a generalization of the quadratic map (1.7). As expected, its inverse is expressed by two formulas:

$$x_t^{(1)} = r_1(x_{t+1}) = [1-(1-x_{t+1})^{1/2}] \times [1+(1-k^2x_{t+1})^{1/2}]^{-1},$$
$$x_t^{(2)} = r_2(x_{t+1}) = [1+(1-x_{t+1})^{1/2}] \times [1+(1-k^2x_{t+1})^{1/2}]^{-1}.$$

Here again, we have $f(x) = x^2$, $f^{-1}(x) = \sqrt{x}$, and consequently, irrespective of the value of $l \in \{2,3,...\}$, the invariant density associated with (3.2) is

$$w(x) = \{2K(k)[x(1-x)(1-k^2x)]^{1/2}\}^{-1}, \quad 0 < k < 1, \tag{3.4}$$

which generalizes (3.1). It is not symmetric about $x = \frac{1}{2}$ because x^2 is not a solution of (2.26).

(3) With the discrete-time mapping

$$x_{t+1} = x_t \left| U_{l-1}(\sqrt{1-x_t^2}) \right| \tag{3.5}$$

stemming from

$$x_t = \sin\theta, \quad x_{t+1} = |\sin l\theta|, \quad 0 \leqslant \theta \leqslant \pi/2,$$

there is associated as invariant density,

$$w(x) = 2/[\pi(1-x^2)^{1/2}], \quad 0 \leqslant x \leqslant 1, \tag{3.6}$$

since $f(x) = x$ and $f^{-1}(x) = x$. As x is not a solution of (2.27), there is no symmetry about $x = \frac{1}{2}$, but on account of $f'(x) = 1$, which entails $df^{-1}(x)/dx = 1$, $\forall x \in [0,1]$, $w(x) > 0$, and $w(1) = +\infty$.

(4) Intermediate between $f(x) = x^2$ and $f(x) = x$, we

can consider

$$f(x) = cx^2 + (1-c)x, \quad 0 \leqslant x \leqslant 1, \tag{3.7}$$

increasing steadily from 0 to 1 as x increases from 0 to 1 when $-1 \leqslant c \leqslant 1$. For $f^{-1}(x)$, we find

$$f^{-1}(x) = 2x/\{1-c+[(1-c)^2+4cx]^{1/2}\}.$$

Putting $k=0$, the invariant density corresponding to the discrete-time mappings

$$x_t = c(\sin\theta)^2 + (1-c)\sin\theta,$$
$$x_{t+1} = c(\sin l\theta)^2 + (1-c)|\sin l\theta|,$$
$$0 \leqslant \theta \leqslant \pi/2, \quad l \in \{2,3,\ldots\},$$

or

$$x_{t+1} = c(\sin\theta U_{l-1}(\cos\theta))^2 + (1-c)\sin\theta\,|U_{l-1}(\cos\theta)|, \tag{3.8}$$

in which

$$\sin\theta = 2x_t/\{1-c+[(1-c)^2+4cx_t]^{1/2}\},$$
$$\cos\theta = \left\{1 - \frac{4x_t^2}{\{1-c+[(1-c)^2+4cx_t]^{1/2}\}^2}\right\}^{1/2},$$

is

$$w(x) = 2\left\{\pi[(1-c)^2+4cx]^{1/2}\left[1 - \frac{4x^2}{\{1-c+[(1-c)^2+4cx]^{1/2}\}^2}\right]^{1/2}\right\}^{-1}. \tag{3.9}$$

(5) In Ref. 4, it was stated that the discrete-time mapping

$$x_{t+1} = \begin{cases} 2x_t, & 0 \leqslant x_t \leqslant \frac{1}{2}, \\ 2(1-x_t), & \frac{1}{2} \leqslant x_t \leqslant 1 \end{cases} \tag{3.10}$$

(rewritten in our notation) gives rise to the constant limit density

$$w(x) = 1, \quad 0 \leqslant x \leqslant 1, \tag{3.11}$$

according to Kac.[1] It is amusing to verify this result, proceeding in the backward direction starting from $w(x)=1$. Let us assume that the function called f throughout this article exists and let us at first put the modulus k equal to zero, for simplicity. Then, according to (2.19), we must have

$$\frac{2\,df^{-1}(x)/dx}{\pi[1-(f^{-1}(x))^2]^{1/2}} = 1, \quad 0 \leqslant x \leqslant 1.$$

Integration from 0 to y on both sides, where $y \in [0,1]$, yields

$$(2/\pi)\arcsin f^{-1}(y) = y,$$

and so

$$f^{-1}(y) = \sin(\pi y/2), \quad 0 \leqslant y \leqslant 1. \tag{3.12}$$

In turn, this gives

$$f(x) = (2/\pi)\arcsin x, \quad 0 \leqslant x \leqslant 1, \tag{3.13}$$

which satisfies the condition (2.27) in virtue of the symmetry of $w(x)=1$ about $x=\frac{1}{2}$. Since $f(x)$ has all the necessary properties to be inserted into (2.1), we know at once that for every $l \in \{2,3,\ldots\}$,

$$x_t = (2/\pi)\arcsin(\sin\theta) = (2/\pi)\theta,$$
$$x_{t+1} = (2/\pi)\arcsin(|\sin l\theta|), \quad 0 \leqslant \theta \leqslant \pi/2,$$

is the parametric representation of a discrete-time mapping to which there corresponds $w(x)=1$. Eliminating θ, the result is

$$x_{t+1} = (2/\pi)\arcsin(|\sin(l\pi x_t/2)|), \tag{3.14}$$

or

$$x_{t+1} = \begin{cases} lx_t, & 0 \leqslant x_t \leqslant 1/l, \\ 2-lx_t, & 1/l \leqslant x_t \leqslant 2/l, \\ -2+lx_t, & 2/l \leqslant x_t \leqslant 3/l, \\ \quad\vdots & \\ (-1)^l(2[l/2]-lx_t), & (l-1)/l \leqslant x_t \leqslant 1. \end{cases} \tag{3.15}$$

The case (3.10) is simply that of $l=2$. It is somewhat astonishing that repeating the calculations with $0<k<1$, hence with elliptic functions, does not yield a different result compared to what we just obtained. The finiteness of $w(x)$ at $x=1$ implies that $(df^{-1}(x)/dx)_{x=1}$ be equal to zero or, equivalently, that $f'(1) = +\infty$. According to (3.12) and (3.13), such is indeed the case.

(6) Let us return to the subject of Ref. 5, namely, the cusp-shaped map (1.5) and the associated limit distribution (1.6), which, after linear transformation so that both x_t and x_{t+1} belong to $[0,1]$, give rise to (2.21) and (2.22), respectively. Again, if we assume the existence of f and its inverse f^{-1}, as well as $k=0$, we can integrate from 0 to y in (2.19) with $w(x) = 2(1-x)$. The result is

$$2y - y^2 = (2/\pi)\arcsin f^{-1}(y)$$

and so

$$f^{-1}(y) = \sin(\pi y - (\pi/2)y^2), \quad 0 \leqslant y \leqslant 1.$$

This, in turn, yields

$$f(x) = 1 - (1 - (2/\pi)\arcsin x)^{1/2}, \quad 0 \leqslant x \leqslant 1. \tag{3.16}$$

Here, $f(x)$ and its inverse satisfy all requirements to be utilized for the construction of a discrete-time map of the type (2.1) and (2.2):

$$x_t = 1 - [1 - (2/\pi)\arcsin(\sin\theta)]^{1/2} = 1 - (1-2\theta/\pi)^{1/2}, \tag{3.17}$$
$$x_{t+1} = 1 - [1 - (2/\pi)\arcsin(|\sin l\theta|)]^{1/2}, \quad 0 \leqslant \theta \leqslant \pi/2,$$

or

$$x_{t+1} = 1 - \{1 - (2/\pi)\arcsin[\,|\sin l\pi(x_t - x_t^2/2)|\,]\}^{1/2},$$
$$0 \leqslant x_t \leqslant 1, \quad l \in \{2,3,...\}. \tag{3.18}$$

We know that solving this discrete-time map with respect to x_t results in l real expressions [see (2.6) with $k = 0$]. Hence, in order to compare our result to (2.21), namely,

$$x_{t+1} = 1 - |1 - 2x_t|^{1/2}, \quad 0 \leqslant x_t \leqslant 1, \tag{3.19}$$

we have to put $l = 2$. In this case, we find

$$x_t = 1 - (1 - 2\theta/\pi)^{1/2},$$
$$x_{t+1} = 1 - [1 - (2/\pi)\arcsin(\sin 2\theta)]^{1/2}$$
$$= 1 - |1 - 4\theta/\pi|^{1/2}, \quad 0 \leqslant \theta \leqslant \pi/2, \tag{3.20}$$

or

$$x_{t+1} = 1 - |1 - 4x_t + 2x_t^2|^{1/2}, \quad 0 \leqslant x_t \leqslant 1. \tag{3.21}$$

This is a cusp-shaped return map, but it is asymmetric with respect to $x = \frac{1}{2}$, and therefore it differs quantitatively from (3.19). The Cartesian graph of (3.21) starts at the origin with slope 2 and increases steadily toward the point $(1 - (\sqrt{2}/2), 1)$, where the slope is $+\infty$. The slope jumps to $-\infty$ as the curve starts to decrease, tending toward (1,0), where the slope becomes 0. Both under (3.19) and (3.21) there is convergence of $\{w_t(x) | t \in \mathbb{N},\ x \in [0,1]\}$ toward $2(1-x)$ as $t \to +\infty$, but this is not paradoxical. Different discrete-time maps may give rise to the same invariant density, as is, for instance, the case with the maps comprised in (2.2) or (2.4) for $l = 2,3,...$. But here, the difference lies deeper: (3.19) is simply not a discrete-time map of the kind (2.2). Parametrically, (3.19) can be represented as

$$x_t = 2\theta/\pi, \quad x_{t+1} = 1 - |1 - 4\theta/\pi|^{1/2}, \quad 0 \leqslant \theta \leqslant \pi/2.$$

Still making use of $f(x)$ as defined by (3.16), this parametric representation of (3.19) becomes

$$x_t = 2f(\sin\theta) - [f(\sin\theta)]^2,$$
$$x_{t+1} = f(\sin 2\theta), \quad 0 \leqslant \theta \leqslant \pi/2, \tag{3.22}$$

confirming our statement made near the beginning of Sec. II, namely, that the type of discrete-time maps represented by (2.4) does not comprise all possible maps. It is practically certain that to all discrete-time maps with parametric form

$$x_t = 2f(\sin\theta) - [f(\sin\theta)]^2, \quad x_{t+1} = f(|\sin l\theta|),$$
$$0 \leqslant \theta \leqslant \pi/2, \quad l \in \{2,3,...\}, \tag{3.23}$$

or even more generally,

$$x_t = 2f(\mathrm{sn}(u,k)) - [f(\mathrm{sn}(u,k))]^2,$$
$$x_{t+1} = f(|\mathrm{sn}(lu,k)|), \quad 0 \leqslant u \leqslant K(k), \quad l \in \{2,3,...\}, \tag{3.24}$$

there corresponds a formula to calculate the limit function $w(x)$, being the counterpart of (2.19) or (2.18), respectively. In the special case of (3.16), this formula, where $k = 0$, should yield the same result as (2.19), i.e., $w(x) = 2(1-x)$.

(7) In Ref. 13, one finds two examples of piecewise linear mappings of [0,1] onto itself, symmetric with respect to $x = \frac{1}{2}$ and giving rise to limit functions which are piecewise constant, with a finite jump at $x = \frac{1}{2}$ (cf. Fig. 2 of Ref. 14). It appears of interest to calculate with which discrete-time maps of the type (2.2), where $l = 2$, these limit functions are associated. Györgyi and Szépfalusy[14] consider

$$w(x) = \begin{cases} 1 + c, & 0 \leqslant x < \frac{1}{2}, \\ 1 - c, & \frac{1}{2} < x \leqslant 1, \end{cases} \quad c \in]-1,1[,$$

and their first numerical example corresponds to $c = -0.6$, in which case they obtain the symmetric map

$$x_{t+1} = \begin{cases} 5x_t, & 0 \leqslant x_t \leqslant \frac{1}{10}, \\ \frac{5}{4}x_t + \frac{3}{8}, & \frac{1}{10} \leqslant x_t \leqslant \frac{1}{2}, \\ \frac{13}{8} - \frac{5}{4}x_t, & \frac{1}{2} \leqslant x_t \leqslant \frac{9}{10}, \\ 5(1 - x_t), & \frac{9}{10} \leqslant x_t \leqslant 1. \end{cases}$$

This discrete-time map is not comprised in (2.2) with $l = 2$, since the present formalism yields another dynamical law, nonsymmetric with respect to $x_t = \frac{1}{2}$. Indeed, inserting

$$w(x) = \begin{cases} 0.4, & 0 \leqslant x < \frac{1}{2}, \\ 1.6, & \frac{1}{2} < x \leqslant 1, \end{cases}$$

into (2.19) and integrating from 0 to y, we obtain

$$(2/\pi)\arcsin f^{-1}(y) = \begin{cases} 2y/5, & 0 \leqslant y \leqslant \frac{1}{2}, \\ 8y/5 - \frac{3}{5}, & \frac{1}{2} \leqslant y \leqslant 1, \end{cases}$$

and consequently,

$$f^{-1}(y) = \begin{cases} \sin(\pi y/5), & 0 \leqslant y \leqslant \frac{1}{2}, \\ \sin(4\pi y/5 - 3\pi/10), & \frac{1}{2} \leqslant y \leqslant 1. \end{cases}$$

In turn, this leads to

$$f(y) = \begin{cases} (5/\pi)\arcsin y, & 0 \leqslant y \leqslant \sin(\pi/10) = (\sqrt{5}-1)/4, \\ \frac{3}{8} + (5/4\pi)\arcsin y, & (\sqrt{5}-1)/4 \leqslant y \leqslant 1. \end{cases}$$

Equation (2.2) with $l = 2$ yields

$$x_{t+1} = \begin{cases} f(\sin(2\pi x_t/5)), & 0 \leqslant x_t \leqslant \frac{1}{2}, \\ f(\sin\{[(8x_t - 3)\pi]/5\}), & \frac{1}{2} \leqslant x_t \leqslant 1. \end{cases}$$

Taking into account the various intervals of validity, we obtain as final result:

$$x_{t+1} = \begin{cases} 2x_t, & 0 \leqslant x_t \leqslant \frac{1}{4}, \\ x_t/2 + \frac{3}{8}, & \frac{1}{4} \leqslant x_t \leqslant \frac{1}{2}, \\ 2x_t - \frac{3}{8}, & \frac{1}{2} \leqslant x_t \leqslant \frac{11}{16}, \\ \frac{19}{8} - 2x_t, & \frac{11}{16} \leqslant x_t \leqslant \frac{15}{16}, \\ 8(1 - x_t), & \frac{15}{16} \leqslant x_t \leqslant 1, \end{cases}$$

which is a piecewise linear, nonsymmetric discrete-time map. The absence of symmetry with respect to $x_t = \frac{1}{2}$ could be foreseen since f^{-1} does not satisfy (2.25) in this case. The situation is similar to that of our example (6).

In the case $c = 0.6$, Györgyi and Szépfalusy obtain

$$x_{t+1} = \begin{cases} \frac{5}{4}x_t, & 0 \leqslant x_t \leqslant \frac{2}{5}, \\ 5x_t - \frac{3}{2}, & \frac{2}{5} \leqslant x_t \leqslant \frac{1}{2}, \\ \frac{7}{2} - 5x_t, & \frac{1}{2} \leqslant x_t \leqslant \frac{3}{5}, \\ \frac{5}{4}(1 - x_t), & \frac{3}{5} \leqslant x_t \leqslant 1, \end{cases}$$

whereas analogous calculations as we carried out above yield

$$x_{t+1} = \begin{cases} 2x_t, & 0 \leqslant x_t \leqslant \frac{1}{4}, \\ 8x_t - \frac{3}{2}, & \frac{1}{4} \leqslant x_t \leqslant \frac{5}{16}, \\ \frac{7}{2} - 8x_t, & \frac{5}{16} \leqslant x_t \leqslant \frac{3}{8}, \\ \frac{5}{4} - 2x_t, & \frac{3}{8} \leqslant x_t \leqslant \frac{1}{2}, \\ \frac{1}{2}(1 - x_t), & \frac{1}{2} \leqslant x_t \leqslant 1. \end{cases}$$

The calculations may also be worked out for arbitrary $c \in]-1,1[$ and the result compared to Eq. (3.9) of Ref. 13, but several subcases have to be distinguished.

(8) Finally, let $f(x)$ be of the form

$$f(x) = a + (b + b'x^2)^{1/2}, \quad 0 \leqslant x \leqslant 1.$$

This should be a single-valued, real, continuous function of x on $[0,1]$, increasing steadily from 0 to 1 as x increases from 0 to 1. Such is the case when

$$a = -b^{1/2}, \quad b' = 1 + 2b^{1/2}, \quad b \geqslant 0.$$

Putting $b^{1/2} = c$, we obtain

$$f(x) = -c + [c^2 + (2c+1)x^2]^{1/2}, \quad 0 \leqslant x \leqslant 1, \quad c \geqslant 0, \tag{3.25}$$

satisfying all requirements and having its derivative function $f'(x)$ continuous in $[0,1]$. It constitutes a way of generalizing $f(x) = x$, which is the special case $c = 0$. When $c > 0$, $f'(0) = 0$ and $f'(1) = (2c+1)/(c+1) > 0$, and therefore, $w(0) = w(1) = +\infty$ can be expected for any $k \in [0,1[$. Using the Jacobian sn function, an acceptable discrete-time map is obtained in parametric form:

$$\begin{aligned} x_t &= -c + [c^2 + (2c+1)(\mathrm{sn}(u,k))^2]^{1/2}, \\ x_{t+1} &= -c + [c^2 + (2c+1)(\mathrm{sn}(lu,k))^2]^{1/2}, \\ & \qquad 0 \leqslant u \leqslant K(k), \quad c \geqslant 0, \end{aligned} \tag{3.26}$$

for every value of l belonging to $\{2,3,\ldots\}$. This example is given in order to show that a function f which is still relatively simple may give rise to discrete-time maps of high degree of complexity, especially as the integer value of l increases. The simplest map contained in (3.26) is that corresponding to $l = 2$. Its explicit form is

$$x_{t+1} = -c + \left\{ c^2 + \frac{4(2c+1)^2 x_t (1-x_t)(2c+x_t)(2c+1+x_t)[2c+1-k^2 x_t (2c+x_t)]}{[(2c+1)^2 - k^2 x_t^2 (2c+x_t)^2]^2} \right\}^{1/2}.$$

Even for $l = 3$, the formula relating x_{t+1} to x_t is already considerably more complicated and yet, for any $l \in \{2,3,\ldots\}$, the sequence $\{w_t(x) \mid t \in \mathbb{N}, x \in [0,1]\}$ starting from an arbitrary normalized initial $w_0(x)$ converges toward

$$w(x) = \frac{(2c+1)^{1/2}(c+x)}{K(k)\{x(1-x)(2c+x)(2c+1+x)[2c+1-k^2 x(2c+x)]\}^{1/2}}.$$

For $c > 0$, we indeed have $w(0) = w(1) = +\infty$. For $c = 0$, wherefore $f(x) = x, f'(x) = 1$, the singularity at $x = 0$ disappears as $w(x)$ reduces to

$$w(x) = \{K(k)[(1-x^2)(1-k^2x^2)]^{1/2}\}^{-1},$$

generalizing (3.6).

ACKNOWLEDGMENT

I wish to thank Dr. H. De Meyer, Senior research associate N.F.W.O. (Belgium) for having drawn my attention to a number of articles concerning this subject and also for stimulating discussions resulting from his keen interest in my research work.

[1] M. Kac, Ann. Math. **47**, 33 (1946).
[2] S. M. Ulam and J. von Neumann, Bull. Am. Math. Soc. **53**, 1120 (1947).
[3] A. Lasota and J. A. Yorke, Trans. Am. Math. Soc. **186**, 481 (1973).
[4] J. B. McGuire and C. J. Thompson, Bull. Aus. Math. Soc. **22**, 133 (1980).
[5] P. C. Hemmer, J. Phys. A: Math. Gen. **17**, L247 (1984).
[6] H. Falk, Phys. Lett. A **105**, (3), 101 (1984).
[7] V. M. Nandakumaran, J. Phys. A: Math. Gen. **18**, L1021 (1985).
[8] C. C. Grosjean, J. Phys. A: Math. Gen. **19**, 3535 (1986).
[9] C. C. Grosjean, Acad. Analecta (Proc. R. Ac. Sc., Litt., Fine Arts, Brussels, Belgium) **48**(5), 95 (1986).
[10] A. D. Barbour, J. Phys. A: Math. Gen. **19**, 3921 (1986).
[11] P. R. Halmos, *Lectures on Ergodic Theory* (Chelsea, New York, 1956).
[12] S. M. Ulam, Intersci. Tracts Pure Appl. Math. **8**, 73 (1960).
[13] S. Grossmann and S. Thomae, Z. Naturforsch. **32a**, 1353 (1977).
[14] G. Györgyi and P. Szépfalusy, Z. Phys. B **55**, 179 (1984).

On unitary SU(*N*) ordered exponentials in a strong coupling limit

H. M. Fried[a)]
Physique Théorique, Université de Nice,[b)] *06034 Nice Cedex, France*

(Received 1 December 1986; accepted for publication 4 February 1987)

A construction is given of the leading, averaged output dependence of a unitary ordered exponential in the strong coupling limit of rapidly fluctuating input. While valid for any SU(N), the method does not provide "fine structure" corrections to the leading output behavior. Numerical illustrations are given using a simple SU(3) example.

I. INTRODUCTION

This paper is meant as a first, and partial generalization to SU(N) of the techniques and results of two previous papers[1,2] dealing with the approximation of ordered SU(2) exponentials in the stochastic limit. Such ordered exponentials (OE's) are found in almost every field of mathematical physics which deals with more than one degree of freedom. Typically, one finds that analytic, continuum calculations for quantities of physical interest are rendered impossible by the presence of an OE, with perturbation expansions remaining the only straightforward method of approach. For strong coupling (SC) problems, however, that avenue is blocked; with the exception of machine estimates written for specific problems, it has not been possible to proceed in any systematic way.

The goal of the present approach, for which this paper is, hopefully, a useful first step, is to exhibit for arbitary N the functional dependence of an OE on its input dependence, in the sense that the expected, rapid fluctuations of the output in the stochastic regime are suppressed, and only the "average," or relatively slowly varying, amplitudes are reproduced. (This slowly varying dependence is, however, quite dependent upon the frequency of the rapid fluctuations.) Certain technical limitations of the present work restrict the validity of this analysis, rendering incomplete the "fine structure" (FS) analysis of the previous papers; nevertheless, the construction given below should provide a description of the leading behavior of an OE, as a functional of the defining input dependence, in the stochastic or rapidly fluctuating limit.

The general OE of interest may be described as the solution to the differential equation

$$\frac{\partial U}{\partial t}(t;E) = i\lambda_a E_a(t) U(t;E) , \tag{1.1}$$

with the initial condition $U(0;E) = 1$; that is,

$$U(t;E) = \left(\exp\left(i\int_0^t dt'\,\boldsymbol{\lambda}\cdot\mathbf{E}(t')\right)\right)_+ , \tag{1.2}$$

with λ_a the $N^2 - 1$ independent, Hermitian, fundamental representation matrices (the Gell–Mann matrices) of SU(N), satisfying the properties

$$\mathrm{tr}[\lambda_a] = 0, \quad \mathrm{tr}[\lambda_a\lambda_b] = 2\delta_{ab} ,$$
$$\tfrac{1}{2}\cdot[\lambda_a,\lambda_b] = if_{abc}\lambda_c ,$$
$$\tfrac{1}{2}\cdot\{\lambda_a,\lambda_b\} = (2/N)\delta_{ab} + d_{abc}\lambda_c .$$

The arguments of the OE $U(t;E)$ have been written to emphasize its functional dependence on the input vector $E_a(t')$, as well as on the specific t dependence of its output form. Frequently, the OE's which appear in various problems are intended to be used as part of the integrand of a subsequent functional integration, weighted with an appropriate probability distribution. For example, if $E_a(t) = g^* v_\mu(t)^* \lambda_a {}^* A_\mu{}^a(x - \int_0^t dt'\, v(t'))$, where A_μ^a is a gauge field and the four-velocity $v_\mu(t)$ is represented by $dx_\mu(t)/dt$, the trace of the corresponding OE will define—before quantum fluctuations are attempted—a QCD Wilson loop.[3] If, on the other hand, E_a is proportional to the symmetric, spatial gradients of a fluid velocity, one can formulate[4] the "vortex stretching" term of three-dimensional fluid dynamics. When the probability weightings used are, or are close to, white-noise Gaussian, one may expect very rapid fluctuations of the input E vectors; and this behavior has previously been termed "stochastic." Because this paper deals only with the "deterministic" behavior of $U(t;E)$ under rapid, nonrandom fluctuations of the unit vector $\hat{E} = \mathbf{E}/E$, $E = (\mathbf{E}^2)^{1/2}$, we will henceforth use the phrase "rapidly fluctuating input" (RFI) to denote that portion of the SC regime studied here.

The weak coupling regime, defined by $\int_0^t dt'\, E(t') \ll 1$, is essentially perturbative and poses no problem. The SC regime, for which $\int_0^t dt'\, E(t') > 1$, has two natural and opposite limits defined most simply in terms of the dimensionless ratio $\rho = |d\,\mathbf{E}/dt\,|/E$. For $\rho = 0$, that is for a constant unit vector $\hat{E}$, the OE of (1.2) reduces to an ordinary exponential, if one merely rotates the coordinate axes such that any one of them points in the $\hat{E}$ direction. Hence, for $\rho \ll 1$, one may expect a set of "adiabatic" approximations to U, defined as corrections to the $\rho = 0$ limit; this has been discussed in some detail for SU(2) in Ref. 1, and can be extended to SU(N), although it will not be treated here. The other, opposite limit $\rho \gg 1$ defines the much more interesting RFI situation, and is the subject of this paper.

As in the SU(2) case, one finds that the complex coefficient functions F_0, F_a in the representation

$$U = F_0 + \sum_a \lambda_a F_a \tag{1.3}$$

are given as rapid fluctuations superimposed on a slowly

[a)] Permanent address: Physics Department, Brown University, Providence, Rhode Island 02912.
[b)] Unité Associée 767 au Centre National de la Recherche Scientifique, Physique Théorique, Parc Valrose, 06034 Nice Cedex, France.

 0022-2488/87/061275-08$02.50

varying, or averaged background; and the time variations of the latter depend in a nontrivial way on the rapid fluctuations of the input. In contrast to the SU(2) case, where the four coefficient functions F_0, $-i^*F_a$ are real, the corresponding functions of SU(N) are complex, and one must pay special attention to the unitarity property of the U, following from (1.1) or (1.2),

$$U^{\dagger *}U = U^*U^{\dagger} = 1, \tag{1.4}$$

which places certain restrictions on the coefficient functions.

Because of the need to maintain unitarity, the theoretical work in this paper will make use of the representation

$$U = \exp\left(i\sum_a \lambda_a G_a\right), \tag{1.5}$$

with $\hat{G} = \mathbf{G}/G$, $G = (\mathbf{G}^2)^{1/2}$, and G_a a real vector with $L = N^2 - 1$ components. Approximate forms will be found for G and $\hat{G}$ in the RFI limit, and a simple conversion made to the coefficient functions of (1.3). For that conversion, as well as for use in other steps of the analysis, a version of an ancient representation due to Lagrange and Sylvester[5] will be used, and will be denoted by the phrase "normal form,"

$$\mathcal{F}(\boldsymbol{\lambda}\cdot\mathbf{v}) = \sum_l \mathcal{F}(\xi_l)\left[\frac{1}{N} + \frac{1}{2}\lambda_a\frac{\partial\xi_l}{\partial v_a}\right], \tag{1.6}$$

where the ξ_l denote the N eigenvalues of $\boldsymbol{\lambda}\cdot\mathbf{v}$, with v_a an arbitrary, $L = N^2 - 1$ component vector. The derivation of (1.6) is elementary, holds for any $\mathcal{F}(z)$ whose Fourier transform exists, and has been relegated to the Appendix.

Because it is appropriate to compare the results of the RFI approximation with the "correct," numerically integrated coefficient functions, the latter have been computed using an algorithm[6] that guarantees unitarity. There, for a fixed time interval Δt one first replaces (1.1) by the difference equation,

$$U(t+\Delta t) = U(t) + i\boldsymbol{\lambda}\cdot\mathbf{E}(t)\Delta t U(t)\,; \tag{1.7}$$

and then, in order to insure the unitarity of the numerically integrated solutions, one replaces (1.7) by the equation

$$U(t+\Delta t) = U(t) + (i/2)\boldsymbol{\lambda}\cdot\mathbf{E}(t)\Delta t\,[U(t+\Delta t) + U(t)]\,, \tag{1.8}$$

which is the same as (1.7) and (1.1) only in the limit $\Delta t = 0$. The "solution" to (1.8) may be written as

$$U(t+\Delta t) = \left(\frac{1 + (i/2)\boldsymbol{\lambda}\cdot\mathbf{E}(t)\Delta t}{1 - (i/2)\boldsymbol{\lambda}\cdot\mathbf{E}(t)\Delta t}\right)U(t)\,, \tag{1.9}$$

and is rigorously unitary, by inspection; that is, if $U(t)$ is unitary, then $U(t+\Delta t)$ is also. Equation (1.9) has been used to compute the "exact" amplitudes with which our approximate forms are compared.

As in Ref. 1, the essential idea is to extract, in the large-ρ limit, the slowly varying envelope upon which the rapid fluctuations ride, for in any physical context it is surely only the slowly varying behavior that is of interest. Simple (unitarity) considerations would, as for the SU(2) case, suggest the improvement of the leading, "averaged" dependence, by including FS corrections expressible in powers of $1/\rho$. However, for general SU(N), and in particular as N increases, there will appear a certain difficulty in calculating FS terms, and for reasons quite distinct from the "technical" difficulties alluded to above. Very simply, most of the coefficient functionals F_a will be expressed as increasingly nonlinear functions of the $\hat{G}_b$; and if one is interested in the slowly varying, or averaged behavior of the F_a, one will have to perform averages over the corresponding combinations of the $\hat{G}_b$. Our analysis, however, is really simple only in its extraction of average values of the $\hat{G}_b$; and since the average of products is not equal to the product of averages—in particular, some slowly varying, FS dependence is always missed—one may expect that, for certain F_a, the FS dependence estimated in the manner of Ref. 1 is bound to be incomplete. For this reason, as well as for the "technical" reasons described below, the results of this paper are restricted to the leading-order estimates of F_0 and the F_a, which are constructed from those parts of the averaged $\hat{G}_b$ independent of ρ (and, as in Ref. 1, from the ρ-dependent G).

Finally, a word must be said about the level of rigor—or its absence—in this paper. When the analysis becomes too difficult to permit a brute force extraction of a desired result, an appeal will be made to intuition, to an argument labeled by the phrase "...what else can it be?" And it may well be that some subtlety not forseen by the author will answer this question in a manner different from that found here. On the basis of the numerical comparisons noted below this would not appear too likely; but it must be understood that, without honest mathematical rigor, such possibilities exist. Nevertheless, the predictions made in this paper are, at the very least, interesting; and they may even be true.

II. ESTIMATING $\hat{G}$

A. The differential equation

One begins with the representation (1.5), differentiating U and substituting into the original Eq. (1.1), to obtain the nonlinear relation

$$\boldsymbol{\lambda}\cdot\mathbf{E} = \int_0^1 d\mu\, e^{i\mu\boldsymbol{\lambda}\cdot\mathbf{G}}\boldsymbol{\lambda}\cdot\frac{d\,\mathbf{G}}{dt}e^{-i\mu\boldsymbol{\lambda}\cdot\mathbf{G}},$$

or

$$E_a = \frac{1}{2}\int_0^1 d\mu\,\mathrm{tr}\left[\lambda_a e^{i\mu\boldsymbol{\lambda}\cdot\mathbf{G}}\boldsymbol{\lambda}\cdot\frac{d\,\mathbf{G}}{dt}e^{-i\mu\boldsymbol{\lambda}\cdot\mathbf{G}}\right]. \tag{2.1}$$

In all of the following we will assume that the time variation of the components $\hat{E}_a$ follows the rule

$$\frac{d\hat{E}_a}{dt} = f_{abc}\rho_b E_c\,, \tag{2.2}$$

where the quantities ρ_b are constants. It then follows that $\boldsymbol{\rho}$ and $\hat{E}$ are orthogonal, and that the norm of $\hat{E}$ is unchanged. For simplicity, we will also assume that the magnitude E does not depend on time. From the experience of Ref. 1, validated by the SU(3) example of the next section, one may expect that the form of our leading dependence will be essentially unchanged even if $\boldsymbol{\rho}$ and $\mathbf{E}$ are time dependent, as long as the condition $\rho \gg 1$ is maintained.

Separating $\mathbf{G}$ into the product $\hat{G}G$, and calculating the trace of all the multicommutators of (2.1) (which are obtained by expanding and resumming in powers of μ), one builds

$$E_a = \hat{G}_a(\mathbf{E}\cdot\hat{G}) + \frac{i}{2}\sum_b \frac{d\hat{G}_b}{dt}\left[(1 - e^{2i\mathbf{A}\cdot\hat{G}G})\cdot\frac{1}{\mathbf{A}\cdot\hat{G}}\right]_{ba}, \tag{2.3}$$

where the $(A^i)_{jk} = i\cdot f_{ijk}$ form the adjoint representation matrices of SU(N). Solving for $d\hat{G}/dt$, with the aid of the relation

$$(1 - e^{2iz})^{-1} = \tfrac{1}{2}[1 + i\cot(z)],$$

one can rewrite (2.3) in the form

$$\frac{d\hat{G}_a}{dt} = f_{abc}\hat{G}_b E_c + [\mathbf{A}\cdot\hat{G}\cot(\mathbf{A}\cdot\hat{G}G)]_{ab}(E_b - \hat{G}_b(\mathbf{E}\cdot\hat{G})). \tag{2.4}$$

As in the SU(2) case, the magnitude G is given in terms of $\hat{G}$ in the following way. Multiplication of (2.1) by $\Sigma_a \hat{G}_a$ yields the relation

$$\hat{G}\cdot\mathbf{E} = \frac{dG}{dt},$$

which, together with the initial condition $G(0) = 1$, provides

$$G(t) = \int_0^t dt'\,\mathbf{E}(t')\cdot\hat{G}(t'), \tag{2.5}$$

showing that G is completely specified by knowledge of $\hat{G}$. It will be convenient to introduce a new variable, τ, defined by $d\tau = dt * E(t)$, $\tau = \int_0^t dt' * E(t')$, in order to rewrite (2.5) and (2.4) as

$$G(\tau) = \int_0^\tau d\tau'\,\hat{E}(\tau')\cdot\hat{G}(\tau') \tag{2.6}$$

and

$$\frac{d\hat{G}_a}{d\tau} = -f_{abc}\hat{E}_b\hat{G}_c + [\mathbf{A}\cdot\hat{G}\cot(\mathbf{A}\cdot\hat{G}G)]_{ab}(\hat{E}_b - \hat{G}_b(\hat{E}\cdot\hat{G})). \tag{2.7}$$

For SU(2), where the $f_{abc} = \epsilon_{abc}$, it is easy to see that

$$[(\mathbf{A}\cdot\hat{G})^2]_{ab} = \delta_{ab} - \hat{G}_a\hat{G}_b, \tag{2.8}$$

and that (2.7) reduces to

$$\frac{d\hat{G}}{d\tau} = -\hat{E}\times\hat{G} + (\hat{E} - \hat{G}(\hat{E}\cdot\hat{G}))\cot(G), \tag{2.9}$$

as quoted in Ref. 2. For SU(N), with $N > 2$, there is no reason to believe that the equation corresponding to (2.9) will depend only on the "group invariant" $\mathbf{G}^2$, for there are other invariants, depending on N, such as $d_{abc}G_aG_bG_c$, which could appear in the corresponding equation. It is here that the first stumbling block to a straightforward analysis arises, for the averaging method to be used requires a certain closure property, which is difficult to see directly from (2.6). What we shall do, instead, is to consider first a model problem, in which one pretends that the internal structure of (2.7) is analogous to that of SU(2). Then, on the basis of that model result, we guess the form of the more realistic $\hat{G}$ dependence in the large-ρ limit. Finally, with the aid of the exact (2.7), we can estimate the size of the coefficients of the various terms, and find, for large ρ at any finite τ, that the leading $\hat{G}$ dependence is just that of the model calculation. Needless to say, one would prefer a more straightforward argument; but indirect as the present estimates are, they do agree with numerical computations in at least one simple, but decidedly nontrivial case of SU(3), as shown in the following section.

B. A simplified model

For $N > 2$, the relation generalizing (2.8) can be written in various ways, but each of them apparently too complicated to permit the estimation of multiple products of $(A*G)^2$, leading to a closed form for the quantity $(A*G)\cot[(A*G)]$. To obtain a simple form for the latter, we introduce the model approximation

$$(\mathbf{A}\cdot\hat{G})^2_{ab} = P\delta_{ab} + Q\hat{G}_a\hat{G}_b, \tag{2.10}$$

which, effectively, pretends that the internal group structure here is that of SU(2). There are other terms which could possibly be added to the rhs of (2.10), such as $d_{asr}\hat{G}_r * d_{bst}\hat{G}_t$, and $d_{abr} * d_{rst}\hat{G}_s\hat{G}_t$, which can appear automatically for $N > 2$. A compact method of writing the exact rhs of (2.8) uses the "normal form" to express the combination in terms of Gell–Mann matrices and the corresponding d_{abc} of SU(L), with $L = N^2 - 1$, and is noted in the Appendix.

With the model approximation of (2.10), the only reflection of arbitrary N lies in the coefficients P and Q, which are determined by insisting that the combination continue to be orthogonal to either $\hat{G}_a$ or $\hat{G}_b$; and that the normalization be appropriate to $\mathrm{tr}[(\mathbf{A}*\hat{G})^2]$. In this way, one finds that $-Q = P = C_A(N)/(L-1) = N/(N^2-2)$, where C_A is the eigenvalue of the first Casimir operator in the adjoint representation. Repeated products of (2.10) can then be summed to yield a differential equation for $\hat{G}$,

$$\frac{d\hat{G}_a}{d\tau} = -f_{abc}\hat{E}_b\hat{G}_c + \sqrt{P}\cot(\sqrt{P}G)(\hat{E}_a - \hat{G}_a(\hat{E}\cdot\hat{G})). \tag{2.11}$$

With the model (2.11), it will be possible to formulate an "averaged" approximation to the coefficient functions of $U(t;E)$; the technique to be used resembles that of the SU(2) calculation, but is different in detail because of the need to work with a relatively large number of $\hat{G}_a$ components, growing as N^2 for large N. If an equation, similar to (2.11), could be obtained without recourse to the model approximation of (2.10), one would be in somewhat better shape as far as the FS is concerned; for the leading dependence of $\hat{G}$ in the large-ρ limit, however, it will be argued below that the model estimates are sufficient.

We now form the following three quantities, $I = \Sigma_{abc}\hat{G}_a\rho_b\hat{E}_c$, $J = \Sigma_a\hat{E}_a\hat{G}_a$, and $K = \Sigma_a\rho_a\hat{G}_a$, and ask for the equations satisfied by these objects in an averaged sense, retaining only bilinear products of rapidly oscillating dependence, such as $\hat{E}_a\hat{E}_b$ and J itself. As in the SU(2) case, a cursory examination of any numerical output suggests that J is approximately a constant, of order $1/\rho$, which means that $\hat{G}$ will have small but rapidly oscillating components in phase with those of $\hat{E}$. If it turns out that K has a slowly varying part, then that will also be true of $\hat{G}$.

With the aid of (2.11), one builds the relations

$$\frac{dI}{d\tau}=\frac{N}{L}(K-\rho^2 J)-JI\sqrt{P}\cot(\sqrt{P}G)\,, \tag{2.12}$$

$$\frac{dJ}{d\tau}=I+\sqrt{P}\cos(\sqrt{P}G)(1-J^2)\,, \tag{2.13}$$

$$\frac{dK}{d\tau}=-I-\sqrt{P}JK\cot(\sqrt{P}G)\,, \tag{2.14}$$

where the last two of these equations follow from (2.11) without approximation, while the first has used the averaging replacements

$$\langle\hat{E}_a\hat{E}_b\rangle=\delta_{ab}/L\,, \tag{2.15}$$

and

$$\langle\hat{E}_a\hat{G}_b\rangle=\delta_{ab}\langle J\rangle/L+f_{abc}\rho_c\langle I\rangle/N\rho^2\,. \tag{2.16}$$

In writing (2.15) and (2.16), the tacit assumption has been made that there are in fact $L=N^2-1$ rapidly oscillating components $\hat{E}_a$; and that their time average over rapid fluctuations is of the same form as is their stochastic average. For other situations, such as the SU(3) example of the next section, the normalization factor (L/N) of (2.12) will be changed; but this change will effect only the FS dependence. Equation (2.16) represents a statement of internal consistency, which is compatible with the definitions of both I and J.

As in the SU(2) case, we use the "experimental" fact that J can be considered to have a small constant, or averaged value, on which is superimposed rapid fluctuations. Neglecting, or averaging over such fluctuations, one then sets $dJ/dt=0$ to obtain from (2.13)

$$I\sim-\sqrt{P}(1-J^2)\cot(\sqrt{P}G)\,, \tag{2.17}$$

whose derivative then yields

$$\frac{dI}{d\tau}\sim-\sqrt{P}(1-J^2)(-J\sqrt{P})\cdot\frac{1}{\sin^2(\sqrt{P}G)}\,. \tag{2.18}$$

Comparing (2.18) with (2.12), one finds that K must, on the average, be considered as a constant, of value

$$K=J\rho^2+(L/N)PJ(1-J)\,. \tag{2.19}$$

Since the lhs of (2.14) is to vanish, one then concludes that

$$I=-JK\sqrt{P}\cot(\sqrt{P}G)\,. \tag{2.20}$$

Finally, a comparison of (2.20) with (2.17) yields

$$JK=(1-J^2)\,, \tag{2.21}$$

which, together with (2.19), serves to determine J as a function of ρ,

$$(L/N)PJ(1-J^2)+J\rho^2=(1-J^2)/J\,,$$

or

$$J^2=\frac{(1+(L/N)P+\rho^2)}{2P\cdot(L/N)}\times\left[1-\left(1-\left[\frac{2\sqrt{PL/N}}{(1+(L/N)P+\rho^2)}\right]^2\right)^{1/2}\right]. \tag{2.22}$$

The negative sign of the square root has been chosen so that J tends to zero as ρ increases. For large ρ, one has

$$J^2\sim(1+(L/N)P+\rho^2)^{-1}+\cdots\,, \tag{2.23}$$

that is, $J\sim1/\rho$. Factors depending on N and L appear only in higher-order corrections to the leading RPI behavior, and even though this is just a model computation, we argue below that the same feature will be true in an exact rendering of the leading, large-ρ behavior.

From (2.21) and (2.23) we now infer that K should be assigned an average vaue, $K=(1-J^2)/J$, and from the definition of K one concludes that, to leading order, there must be assigned an averaged, nonzero value: $\hat{G}_a=\hat{\rho}_a$. Our model solution, denoted by the quantity U, is then

$$U=e^{i\lambda_a\hat{\rho}_aG}\,,\quad G\sim\int_0^t dt'\frac{E}{\rho}\,. \tag{2.24}$$

This result is equivalent to that of the previous (and somewhat different) SU(2) calculation in the limit of very large ρ. Since the ρ_a are specified input constants, one can proceed from (2.24) to the coefficient functions F_0, F_a by the use of the normal form, (1.6). If the differences between products of the averages of $\hat{G}$ components and the averages of the same products is a FS effect, as is true in all the SU(2) work and in the SU(3) example of the next section, then the leading, large-ρ behavior of the coefficient functions is obtained from (2.24).

C. Guessing the solution

How could the use of the exact (2.7) change the model result (2.24) that followed from (2.11)? The only (reasonable) difference would be that $\hat{G}_a$ would have a constant (averaged over the rapid fluctuations) part that now depends on unit vectors $\hat{W}_a^{(i)}(\rho)$ more general and more complicated than the $\hat{W}_a^{(1)}=\hat{\rho}_a$ alone. Such unit vectors, orthogonal to the ρ_a, can be constructed out of the gradients of the $N-1$ independent invariants $\mathrm{tr}([\lambda\cdot\hat{\rho}]^l)$; e.g., if $d_{abc}\hat{\rho}_b\hat{\rho}_c=\frac{1}{3}(d/d\hat{\rho}_a)(d_{rst}\hat{\rho}_r\hat{\rho}_s\hat{\rho}_t)$, then one choice for $\hat{w}^{(2)}{}_a$ is

$$(d_{abc}\hat{\rho}_b\hat{\rho}_c-\hat{\rho}_a d_{\alpha\beta\gamma}\hat{\rho}_\alpha\hat{\rho}_\beta\hat{\rho}_\gamma)\cdot\left(\sum d^2_{abc}\hat{\rho}_b\hat{\rho}_c-[d_{rst}\hat{\rho}_r\hat{\rho}_s\hat{\rho}_t]^2\right)^{-1}.$$

One would then expect to find a representation for the averaged $\hat{G}$ in the form of a sum of such terms,

$$\hat{G}_a=\sum_i c_i\hat{W}_a^{(i)}(\hat{\rho})\,, \tag{2.25}$$

where the coefficients c_i remain to be determined, but where the overall normalization is chosen to satisfy $\hat{G}^2=1$. Under a set of reasonable assumptions, we now obtain the leading behavior of these coefficients.

To see how this goes, return to the exact equation (2.7) for $\hat{G}$, and calculate the time variation of the same I,J,K quantities as before. One can no longer carry through the process of writing the quantity

$$Q_{ab}=[\mathbf{A}\cdot\hat{\mathbf{G}}\cot(\mathbf{A}\cdot\hat{\mathbf{G}}G)]_{ab}$$

in closed form; but since Q_{ab} will only be used to estimate leading orders of magnitude when averaged, or contracted with an $\hat{E}_c$, this will not be too important. Based on the SU(2) work, on the model of the previous section, and on the SU(3) example of the next section, we now introduce two assumptions which form the basis of the argument to follow.

(1) Contraction with any $\hat{E}_c(t)$ produces for an arbitrary $\mathscr{F}(\hat{G})$ a result of one nominal order lower in ρ,

$$\langle \hat{E}_a \mathcal{F}(\hat{G}) \rangle = \langle \hat{E}_a \hat{G}_c \rangle \frac{\delta}{\delta \hat{G}_c} \mathcal{F}(\hat{G})|_{\hat{G} \to \hat{G}^{(0)}}, \quad (2.26)$$

where $\hat{G}^{(0)}$ denotes the constant dependence of $\hat{G}$ of $O(1)$, after averaging over the rapid fluctuations. Again, we adopt the internally consistent relation (2.16), and assume an averaging over all $L = N^2 - 1$ components $\hat{E}_a$. By this assumption, $\langle I \rangle \sim O(1)$, $\langle J \rangle \sim O(1/\rho)$, and $\langle K \rangle \sim O(\rho)$.

(2) The time average $\langle q \rangle$ over rapid fluctuations of any quantity $q(\tau)$ may still depend on a slowly varying τ dependence; and $\langle dq/d\tau \rangle = d\langle q \rangle/d\tau$. All slowly varying dependence—such as that of $G = \int d\tau' J$—will be treated as constants during the averaging over rapid fluctuations. [Were this not the case, then some of the $O(1/\rho)$ could conceivably be associated with slow variation of the G dependence.]

Since no approximation is now considered for the Q_{ab}, and since $\hat{G}_a Q_{ab} = Q_{ab} \hat{G} = 0$ by virtue of the definition of A_{ab}, we can drop the corresponding term on the rhs of the exact (2.7). The counterparts of Eqs. (2.12) through (2.14) then become

$$\frac{dI}{d\tau} = \frac{N}{L}(K - J\rho^2) + O\left(\frac{1}{\rho^2}\right), \quad (2.27)$$

$$\frac{dJ}{d\tau} = I - \frac{1}{L} tr\, Q, \quad (2.28)$$

$$\frac{dK}{d\tau} = -I + \rho_a \langle \hat{E}_b \hat{G}_c \rangle \frac{\partial}{\partial \hat{G}_c} Q_{ab}|_{\hat{G} = \hat{G}^{(0)}}, \quad (2.29)$$

where the averaging symbols $\langle\ \rangle$ have been suppressed, as in Sec. II B, but where an averaging has been performed to obtain each of these three equations. If, again, we call upon the "experimental" knowledge that J is a constant, in the large-ρ limit, we may infer from (2.28) that

$$I \approx (1/L) \mathrm{tr}\, Q, \quad (2.30)$$

and then, from (2.27), that

$$K \approx J\rho^2 + \frac{1}{N} \frac{d}{d\tau} \mathrm{tr}[\ ]. \quad (2.31)$$

Substituting into (2.29), and comparing with (2.30), one can see that $\mathrm{tr}\, Q(\tau)$ satisfies a forced harmonic oscillator equation, with frequency $(1/2\pi)^*(N/L)^{1/2}$ and driving term given by $\rho_a \langle \hat{E}_b \hat{G}_c \rangle (\partial/\partial \hat{G}_c) Q_{ab}|_{\hat{G}^{(0)}}$. More detailed information is difficult to extract; but for our purposes essentially all that is needed is (2.31), which says that to leading order, $K = J^*\rho^2$. In terms of $\hat{G}$, this means that $\Sigma_a \hat{\rho}_a \hat{G}_a = \rho J$, and one can conclude that $\hat{G}$ is given by a series of terms of the form of (2.25), of which the coefficient multiplying $\hat{\rho}$, say c_1, is given by $\rho^* J$. In the previous model calculation, we knew that $c_1 = 1$, but that information is missing here. In general, $\hat{G}$ may very well have projections in (group) directions orthogonal to $\hat{\rho}$, as explained just before (2.25).

Suppose that $\hat{G}$ does have a projection in the direction of a unit vector $\hat{v}(\hat{\rho})$ perpendicular to $\hat{\rho}$. Analogous to the quantities K and I, one may define the sums $M = \Sigma_a v_a \hat{G}_a$ and $P = \Sigma_a f_{abc} \hat{G}_a v_b \hat{E}_c$, where M measures the projection of $\hat{G}$ in the $\hat{v}$ direction, and the averaging of P must generate a term of $O(1/\rho)$. Following the same averaging techniques, one then finds that

$$\frac{dM}{d\tau} = -P + v_a \langle \hat{E}_b \hat{G}_c \rangle \frac{\partial}{\partial \hat{G}_c} Q_{ab}|_{\hat{G}^{(0)}} \quad (2.32)$$

and

$$\frac{dP}{d\tau} = -\frac{N}{L} J \mathbf{v} \cdot \boldsymbol{\rho} + \frac{N}{L} M, \quad (2.33)$$

where the first term on the rhs of (2.33) has been included, for clarity, even if the value of this term is zero; another pair of terms, not shown, have vanished by symmetry. That first term is zero because $\mathbf{v}$ has been chosen orthogonal to $\boldsymbol{\rho}$; but if it did not vanish it would contribute a term of order unity to (2.33). Because all rapid fluctuations of P have been averaged away, so that $dP/d\tau$ is at least of $O(1/\rho)$, the inescapable conclusion is that M must also be of $O(1/\rho)$. If this arbitrarily chosen projection of $\hat{G}$ is $O(1/\rho)$, then to $O(1)$, the properly normalized $\hat{G}$ can only be given by $\hat{G} = \hat{\rho}$; then, $c_1 = 1$, and $J \sim 1/\rho$.

By this construction, the model result (2.24) is correct for the non-FS terms of $O(1)$ in the coefficient functions F_0, F_a built from (2.24) with the aid of the normal form representation. This statement will next be verified with the aid of a simple but nontrivial SU(3) calculation. It must be remembered, however, that in the passage from the $\hat{G}$ description to that of the coefficient functions, we are going to assume that the difference between the average of a product of $\hat{G}$ components and the product of the corresponding averages is of higher order in $1/\rho$. In fact, in the discussion of Sec. II B, we have cavalierly ignored the difference between the average of products of the I, J, K and the corresponding products of the averages; this is obviously wrong, since it is easy to see that

$$\langle KJ \rangle \sim (1 + 1/L)\langle K \rangle \langle J \rangle,$$

$$\langle IJ \rangle \sim (1 + 2/L)\langle I \rangle \langle J \rangle, \quad \langle J^2 \rangle \sim \langle J \rangle^2 + 1/L.$$

In spite of these omissions, the $O(1)$ results of the model are unchanged. For the $\hat{G}_a$, the neglect of such correlations does seem to be a viable assumption, as is born out by the calculations (and figures) of the next section.

III. AN SU(3) EXAMPLE

The simplest, nontrivial example of an SU(3), OE has but a single component of ρ, driving two components of $\hat{E}_a$ each in a different "sector" of SU(3). That is, we suppose that the only nonzero components are $\hat{E}_2 = \cos(\omega t)$ and $\hat{E}_5 = \sin(\omega t)$, so that $\rho_a = \delta_{a7} \cdot \rho$ with $\rho = 2^*(\omega/E)$. Hence

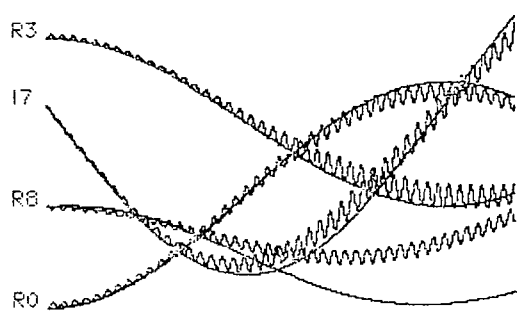

FIG. 1. A superposition of $O(1)$ averaged approximations and four numerically integrated functions, for $E = 10$ and $\omega = 60$.

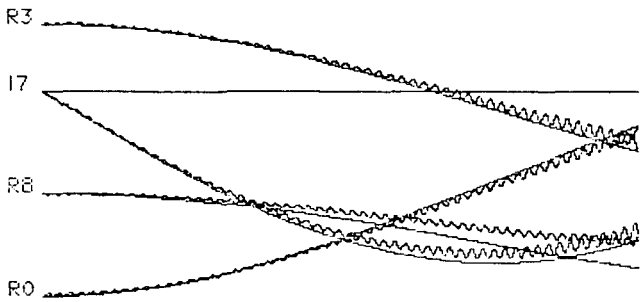

FIG. 2. The same as in Fig. 1, except that $\omega = 90$.

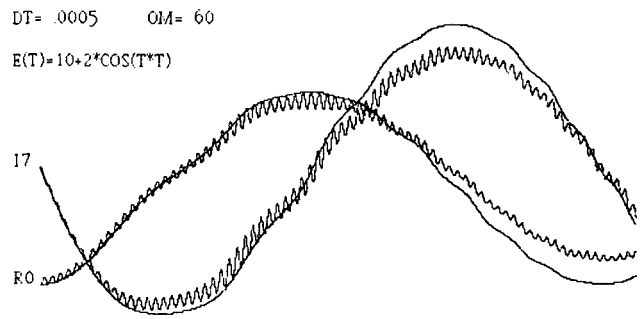

FIG. 4. The same as in Fig. 3, except that $\Delta t = 0.0005$.

the leading $\hat{G}_a^{(0)} = \delta_{a7}$, and the nonzero coefficient functions F_0, F_a are easily worked out to be

$$F_0 = \tfrac{1}{3}[1 + 2\cos G\,], \tag{3.1}$$

and

$$F_7 = i \sin G, \quad F_3 = \tfrac{1}{2}[1 - \cos G\,], \quad F_8 = (1/2\sqrt{3})[1 - \cos G\,]. \tag{3.2}$$

A superposition of these four predictions with those of the numerically integrated solutions is given in Figs. 1 and 2, in the simplest possible situation of constant ω and E, for two different values of ρ. (Here and subsequently we use the notation R_a and I_a to specify, for this special input, the nonzero real and imaginary parts, respectively, of the F_a.) Only for R_8 is the agreement less than satisfactory; and this will be discussed separately, below.

At larger t values, however, there is another source of error in the graphical presentation of these results, as the differences between the numerical integrations and the true solutions become evident. This is to be expected since (1.8) and (1.1) become the same only in the limit $\Delta t \to 0$; for finite Δt the numerically integrated functions start to separate from the exact solutions, and this separation becomes greater the larger the value of t. In Figs. 3 and 4, produced for the same t-dependent $\rho(t)$ but for different values of Δt, this "Δt" effect can clearly be seen. One infers that, in the limit of very small Δt, the smoothly varying part of the exact functions are well described by (3.1) and (3.2), even when ρ depends upon t; the only requirement is that $\rho \gg 1$.

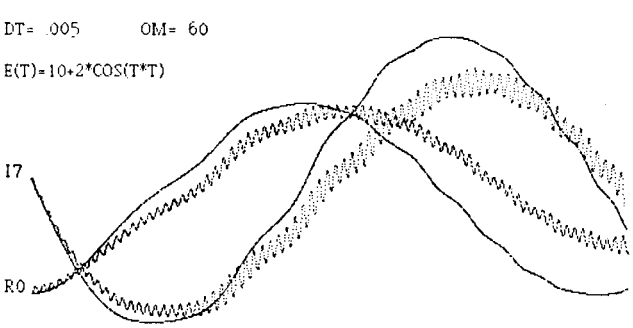

FIG. 3. A superposition of approximate and numerically integrated functions for R_0 and I_7, with $\omega = 60$ and $E(t) = 10 + 2^* \cos(t^2)$, for $\Delta t = 0.005$.

Finally, one comes to the remaining five functions whose $O(1)$ dependence is predicted to be zero by this analysis, I_2, I_5, R_1, R_4, and R_6. Figure 5 clearly shows that these functions decrease at least as fast as $1/\rho$; and it also shows the existence of higher harmonics present in those curves, which becomes especially clear for larger values of ρ.

Although no statement of FS has been attempted in this paper, it is perhaps worth remarking that an attempt to write down appropriate corrections to all of these functions can be organized by using the model forms of Sec. II B, for this simplest case of SU(3). This is not completely correct for reasons mentioned above: neglect of the $O(1/\rho)$ terms due to the difference of averages of products and products of averages, and due to the uncertainties of the model itself. But it is at least interesting to see how closely such FS can approximate the structure of the numerically integrated functions. For a set of $O(1/\rho)$ corrections to the $\hat{G}_a$, computed in a straightforward way and converted to the coefficient functions, this is shown in Fig. 6. Clearly, one is on the right track, even if other, neglected effects become important at later times.

IV. SUMMARY AND CRITIQUE

A partial, first step towards the SC, RFI approximation of SU(N) OE's has been suggested in this paper, in which the leading terms of that OE are specified. A comparison with the numerically integrated functions of a simple but nontrivial SU(3) example illustrates the sort of agreement that may be expected.

While interesting, and of probable future use in estimating functional integrals over distributions close to white-noise Gaussian, the analysis is incomplete in that it is unable to extract the $O(1/\rho)$ corrections. The method of analysis

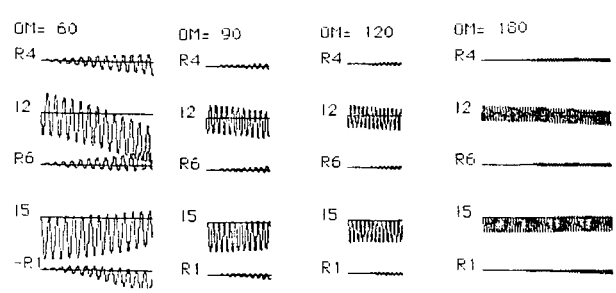

FIG. 5. Numerically integrated functions which are (at least) of $O(1/\rho)$ illustrated for $E = 10$ and four choices of ρ.

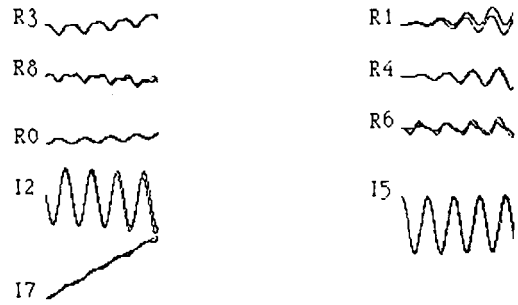

FIG. 6. An example of FS for the set of nine functions, in the SU(3) example described in the text.

also leaves unproved an assertion on the sources of $O(1/\rho)$ dependence. It will therefore be worthwhile to list those inadequacies of this treatment; if such difficulties could be overcome, one would have a much more satisfying approach to this entire problem.

(1) The difference between an average of products of the components $\hat{G}_a$ and the product of corresponding averages has been treated as an effect of relative $O(1/\rho)$. This could be proved if one were sure that slowly varying $G(\tau)$ dependence does not contribute $O(1/\rho)$ terms which have here been associated with the $\hat{G}_a$ dependence. In this paper, averages over rapid fluctuations have been taken under the assumption that the $G(\tau)$ factors may be treated as constants, for which case the $O(1/\rho)$ behavior is as stated. With one possible exception, the SU(3) example seems to bear this out; but no proof has been given.

(2) The averaging process used has been the simplest possible, with every contraction of an $\hat{E}$ component with a $\hat{G}$ component assumed to be of $O(1/\rho)$, and all noncontracted $\hat{G}$ components replaced by their $O(1)$ average value. In effect, no higher harmonics of the RFI were considered, even though one can see them appearing in Fig. 5. Further, the averaging process depends on the number of nonzero $\hat{E}$ components, with different normalization factors appearing in different situations.

With one exception—the discrepancy associated with R_8 in Figs. 1 and 2—these assumptions appear to work quite well. Within the context of this approach, one can try to understand possible reasons for such poor accuracy. Since R_8 really depends on $\hat{G}_7^2$, one might expect that this is an indication that the difference between $\langle \hat{G}_7^2 \rangle$ and $\langle \hat{G}_7 \rangle^2$ can be of $O(1)$. However, R_3 also depends upon $\hat{G}_7^2$, and there is no corresponding effect there. [R_8 has been numerically calculated for a variety of large ρ values, and one finds no appreciable difference between the vertical separation of the approximate and numerically integrated curves, measured at the same geometric points (e.g., at that value of t corresponding to the intersection of R_3 and R_0).] One might next ask if this is a "Δt" effect, due to the inaccuracies of the numerical integration; but the answer is, again, no. The only possibility left is an error in the numerical integration, although that has also been rechecked. As it stands, the author has no explanation for the discrepancy of the R_8 curves.

Many of these difficulties would disappear if a useful representation of $(A*G)^2$ could be found, similar to that of the model of Sec. II B. This is important because it surely holds the key to a development of the FS corrections. In the previous SU(2) analysis, it was possible to calculate the $O(1/\rho)$ corrections, which can always be relevant to some physical situation. For example, in the problem of neutrino–antineutrino oscillations in the presence of a constant plus-time-periodic magnetic field,[7] the probability $|\bar{N}|^2$ of finding an antineutrino at time t, when there was none at time $t=0$, is just one of finding the FS to the OE representing the solution to a pair of coupled, first-order differential equations for the neutrino and antineutrino amplitudes. In the notation of Ref. 7, the answer to that problem is

$$|\bar{N}|^2 = \left[\left(\frac{\omega_m}{\omega_B}\right)\frac{\sin((2\omega_B r/\omega)\sin^2(\omega t/2) - \omega_B t)}{[1 - r\sin(\omega t)]}\right]^2$$

for fixed ωt, $r \neq \pm 1$, and $\omega_m \ll \omega_B$. It corresponds to the square of certain $O(1/\rho)$ terms ($F_1^2 + F_2^2$, in the notation of Ref. 2), where $\rho = 2(\omega_B/\omega_m)|1 - r\sin(\omega t)|$.

In spite of the limitations described above, it is hoped that this paper will point the way towards a systematic method for the approximation of OE's in the RFI limit.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the kind hospitality of colleagues at Physique Théorique, Université de Nice, and at the Observatoire de Nice, during the academic year 1985–1986, where most of this work was performed.

An NSF–CNRS travel grant was useful in the preliminary stages of this project. This work was supported in part by the U. S. Department of Energy Contract No. DE-AC0276A03130.A009-Task A.

APPENDIX: DERIVATION OF THE "NORMAL FORM"

The derivation of (1.6) may be sketched as follows. One writes an arbitrary $\mathcal{F}(\boldsymbol{\lambda}\cdot\mathbf{G})$ in terms of its Fourier transform,

$$\mathcal{F}(\boldsymbol{\lambda}\cdot\mathbf{G}) = \int_{-\infty}^{+\infty} d\omega\, \tilde{\mathcal{F}}(\omega) e^{i\omega\boldsymbol{\lambda}\cdot\mathbf{G}},$$

and considers the unitary exponential written in terms of its coefficient functions,

$$e^{i\omega\boldsymbol{\lambda}\cdot\mathbf{G}} = F_0 + i\boldsymbol{\lambda}\cdot\mathbf{F}.$$

Let the N eigenvalues of $\boldsymbol{\lambda}\cdot\mathbf{G}$ be denoted by $\xi_l(\mathbf{G})$. Then

$$F_0 = \frac{1}{N}\,\mathrm{tr}\,[e^{i\omega\boldsymbol{\lambda}\cdot\mathbf{G}}] = \frac{1}{N}\sum_{l=1}^{N} e^{i\omega\xi_l},$$

and

$$F_i = -\frac{i}{2}\,\mathrm{tr}[\lambda_i e^{i\omega\boldsymbol{\lambda}\cdot\mathbf{G}}] = -\frac{N}{2\omega}\frac{\partial}{\partial G_i}F_0.$$

Hence

$$\mathcal{F}(\boldsymbol{\lambda}\cdot\mathbf{G}) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dz\, \mathcal{F}(z)\int_{-\infty}^{+\infty} d\omega\, e^{-i\omega z} \times \sum_{l=1}^{N} e^{i\omega\xi_l}\left\{\frac{1}{N} + \frac{\lambda_i}{2}\frac{\partial\xi_l}{\partial G_i}\right\},$$

or

$$\mathscr{F}(\boldsymbol{\lambda}\cdot\mathbf{G}) = \sum_{l=1}^{N} \mathscr{F}(\xi_l)\left\{\frac{1}{N} + \frac{1}{2}\lambda_i \frac{\partial \xi_l}{\partial G_i}\right\},$$

which is the "normal form" quoted in the text. A similar representation may be given for any $\mathscr{F}(\mathbf{A}^*\mathbf{G})$ in terms of the SU(L) defining representation matrices Λ_α, with $L = N^2 - 1$ and $\mathbf{A}\cdot\mathbf{G} = \psi_\alpha \cdot \Lambda_\alpha$,

$$\mathscr{F}(\mathbf{A}\cdot\mathbf{G}) = \sum_{l=1}^{L} \mathscr{F}(\xi_l[\psi])\left\{\frac{1}{L} + \frac{1}{2}\Lambda_\alpha \frac{\partial \xi_l[\psi]}{\partial \psi_\beta}\right\}.$$

[1] M. E. Brachet and H. M. Fried, Phys. Lett. A **103**, 309 (1984).
[2] M. E. Brachet and H. M. Fried, J. Math. Phys. **28**, 15 (1987).
[3] Another approach to Wilson loops, involving a strong coupling estimate by the extraction of relevant infrared effects, directly in the continuum, may be found in H-T. Cho, Ph. D. thesis, Brown University, 1987.
[4] H. M. Fried and J. Tessendorf, J. Math. Phys. **25**, 1144 (1984). The SU(3) OE discussed and approximated in the adiabatic limit there is not unitary, in contrast to that of the present paper.
[5] The author is indebted to P. Sorba for pointing out the origin of this "normal form."
[6] This is a variant of the Runge–Kutta method. The use of this sort of algorithm was kindly suggested to the author by D. Weingarten.
[7] W. P. Trower and N. Zovko, Phys. Rev. D **25**, 3088 (1982). The author thanks N. Zovko for an interesting discussion.

Existence and approximation of the solutions of some nonlinear problems

S. R. Vatsya
Atomic Energy of Canada Limited, Whiteshell Nuclear Research Establishment, Pinawa, Manitoba, Canada ROE 1L0

(Received 11 September 1986; accepted for publication 7 January 1987)

Nonlinear equations defined by a positive definite, elliptic operator and nonlinear functions are considered. It is proved that a unique solution exists for a wider class of problems than previously determined, which may be approximated by an iterative method. These results are shown to hold for an equation arising in some reaction-diffusion phenomena.

I. INTRODUCTION

Keller[1] and Pennline[2] have considered two-point boundary value problems of the form

$$-\frac{d^2u}{dx^2}=f(x;u), \quad 0<x<1, \quad u(0)=u(1)=0, \tag{1}$$

where f is a function of u. Equation (1) is equivalent to the integral equation

$$u(x)=\int_0^1 d\xi\, g_k(x,\xi)[ku(\xi)+f(\xi;u(\xi))],$$

where $g_k(x,\xi)$ is the Green's function defining the inverse of the operator $(-(d^2/dx^2)+k)$ with the stated boundary conditions, whenever it exists.

Keller[1] has shown that if, for all u and all $x\in[0,1]$, $\partial f/\partial u=f'$ is continuous and $0\geqslant f'\geqslant -N$ with some $N\geqslant 0$, then Eq. (1) has a unique solution u given by

$$u=\lim_{n\to\infty} u_n,$$

where $u_0(x)=0$,

$$u_{n+1}(x)=\int_0^1 d\xi\, g_k(x,\xi)[ku_n(\xi)+f(\xi;u_n(\xi))], \quad n=0,1,2,...,$$

and $k\geqslant N$. The convergence is uniform with respect to $x\in[0,1]$. Pennline[2] points out that it is possible to take $k\geqslant N/2$, which improves the rate of convergence. Also an attempt was made at relaxing the conditions further.

The work of Ref. 2 was motivated by the equation

$$\frac{d^2v}{dx^2}=\phi^2v^\nu, \quad \nu\geqslant 1, \quad \phi^2\geqslant 0, \quad 0<x<1,$$

$$v'(0)=0, \quad v(1)=1, \tag{2}$$

which arises in the problem of steady-state, isothermal, reaction-diffusion of a substance involving nth-order kinetics.[3] By setting $u=1-v$, Eq. (2) reduces to an equivalent equation,

$$-\frac{d^2u}{dx^2}=\phi^2(1-u)^\nu, \quad 0<x<1,$$

$$u'(0)=0, \quad u(1)=0. \tag{3}$$

While the results of Refs. 1 and 2 are not applicable in this case, a similar but independent treatment was used in Ref. 2 to conclude that, with a restriction on the values of ϕ, Eq. (2) has a unique solution which may be approximated by the iterative method.

Let D be an open, bounded subset of $R^l, l\geqslant 1$, and let ∂D, the boundary of D, be piecewise continuous. Also let H be the Hilbert space of square integrable functions of x on $\bar{D}=D\cup\partial D$ with the norm denoted by $\|\cdot\|$. In the present note, we consider the equation

$$Lu=f(u), \tag{4}$$

where L is a positive definite operator from H to H. Let λ be the greatest lower bound of L. It will be assumed that $(L+k)^{-1}$ for each $k>-\lambda$ is an integral operator with its kernel, still denoted by $g_k(x,\xi)$, being a non-negative, bounded function on $\bar{D}\times\bar{D}$. As a result of the boundedness of $g_k(x,\xi)$, $(L+k)^{-1}$ is a Hilbert–Schmidt operator and hence λ is an eigenvalue. These conditions are satisfied, in particular, for the case when L is the elliptic operator defined by

$$(Lu)(x)=-\sum_{i,j=}\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u(x)}{\partial x_j}\right)+a_0(x)u(x), \quad x\in D;$$

$$\alpha(x)u(x)+\beta(x)\frac{\partial u(x)}{\partial\eta}=0, \quad x\in\partial D;$$

$$\alpha(x) \text{ and } \beta(x)\geqslant 0, \quad \alpha(x) \text{ or } \beta(x)>0, \quad \alpha(x)\not\equiv 0;$$

where $\partial/\partial\eta$ is the conormal derivative, and the coefficients, $a_{ij}(x)$, are the elements of a continuously differentiable matrix-valued function on D that is bounded below by a positive constant and $a_0(x)\geqslant 0$ is continuous.[4] It is clear that the operators appearing in Eqs. (1) and (3) are special cases of the elliptic operator considered here.

In addition to the generalization described above, we extend the results of Refs. 1 and 2 to include f with f' bounded above by a positive constant instead of zero. We show further that this weaker condition on f' need be satisfied on a conveniently characterized narrower set of functions only. As an application, it is deduced that Eq. (3) has a unique solution for all values of ϕ which may be approximated by the iterative method. Present results complement, also, those of Ref. 5 where Eq. (4) with f belonging to a different class was studied.

 0022-2488/87/061283-04$02.50

II. EXISTENCE AND APPROXIMATION OF u

Let S be a given set of functions and let $|||\cdot|||$ denote the supremum norm, i.e., $|||v||| = \mathrm{Sup}_{x\in\bar{D}}|v(x)|$ whenever it exists. Condition (C1) will be said to be satisfied on S if

(C1) $S\subseteq H$, S is convex and it is closed with respect to $|||\cdot|||$.

A set S is convex iff $v,w\in S$ implies that $[tv+(1-t)w]\in S$ for each $t\in[0,1]$, and S is closed in the stated sense iff, with a given sequence $\{v_m\}\subset S$ of bounded functions $|||v_n - v_m|||\to_{n,m\to\infty} 0$ implies that $v = \lim_{n\to\infty} v_n \in S$. The set S may or may not be closed with respect to $\|\cdot\|$.

With (C1) satisfied, let f be such that

(C2) there exist constants γ and N such that for all $v\in S$ and all $x\in\bar{D}$, $\lambda > \gamma \geqslant f'(v) \geqslant -N$, $N\geqslant 0$;

(C3) there exists a $v\in S$ such that $f(v)\in H$.

If (C1) and (C2) are satisfied, then (C3) implies that $f(v)\in H$ for each $v\in S$ for the following. With v, $w\in S$,

$$f(w) = f(v) + \int_0^1 dt\, f'[tv+(1-t)w](v-w).$$

From the convexity of S and (C2), it follows that

$$\gamma \geqslant \int_0^1 dt\, f'[tv+(1-t)w] \geqslant -N, \tag{5}$$

implying that $|f(w)| \leqslant |f(v)| + \max(\gamma,N)|v-w|$. Since $|v-w|\in H$, $f(v)\in H$ implies that $f(w)\in H$.

In view of the above,

$$A_k v = (L+k)^{-1}(kv + f(v)),\quad v\in S,$$

defines a one-parameter family of operators A_k from S to H and $k\in(-\lambda,\infty)$. Furthermore, the fixed points of A_k and the solutions of Eq. (4) are in a one-to-one correspondence. The parameter k will be assumed to be restricted by

(C4) $k \geqslant \frac{1}{2}[\max(\gamma,N) - \gamma] = \bar{k}$.

We shall also assume

(C5) for some k satisfying (C4), $A_k S\subseteq S$.

With $u_0\in S$, let $u_{n+1} = A_k u_n$, $n = 0,1,2,\ldots$; $\{u_n\}$ will be called the iterative sequence generated by u_0. It follows, by induction, that $\{u_n\}\subset S$, for all k in conformity with (C4) and (C5).

Theorem 1: Let S, f, and k be such that the conditions (C1)–(C5) are satisfied, and let $\{u_n\}$ be the iterative sequence generated by a $u_0\in S$. Then A_k has a unique fixed point u in S and

$$|||u_n - u||| \underset{n\to\infty}{\to} 0.$$

Proof: We divide the proof in the following five steps.

Step 1: A_k is a contraction of S in H.

Proof: Since $A_k S\subseteq S$, A_k is a map from S to S. With $v,w\in S$, we have

$$\begin{aligned} A_k v - A_k w &= (L+k)^{-1}[k(v-w) + f(v) - f(w)] \\ &= (L+k)^{-1} B_k(v,w)(v-w), \end{aligned}$$

where $B_k(v,w)$ is the operation of multiplication defined by

$$B_k(v,w)h = \left(k + \int_0^1 dt\, f'[tv+(1-t)w]\right)h,\quad h\in H.$$

Since S is convex and (C2) holds, the inequality given by Eq. (5) is valid, which in view of (C4) yields

$$\begin{aligned} k+\gamma &\geqslant k + \int_0^1 dt\, f'[tv+(1-t)w] \\ &\geqslant -(k+\gamma),\quad k+\gamma\geqslant 0. \end{aligned}$$

Hence, $\|B_k(v,w)\| \leqslant (k+\gamma)$. Also, $\|(L+k)^{-1}\| \leqslant 1/(k+\lambda)$. It follows that

$$\|A_k v - A_k w\| \leqslant \mu_k \|v-w\|,$$

where

$$\mu_k = (k+\gamma)/(k+\lambda) < 1.$$

Step 2: For $v\in S$, $A_k v$ is bounded.

Proof: For each $x\in\bar{D}$,

$$\begin{aligned} |(A_k v)(x)| &= \left|\int_{\bar{D}} d\xi\, g_k(x,\xi)[kv(\xi) + f(\xi;v(\xi))]\right| \\ &\leqslant \|kv + f(v)\| \left[\int_{\bar{D}} d\xi\, g_k^2(x,\xi)\right]^{1/2} \end{aligned}$$

by the Schwarz inequality. Since $g_k(x,\xi)$ is bounded and the measure of $\bar{D}$ is finite, the bracketed term is bounded, say by M. Thus

$$|(A_k v)(x)| \leqslant M\|kv + f(v)\|.$$

Step 3: For $v,w\in S$, $|||A_k v - A_k w||| \leqslant M(k+\gamma) \times \|v-w\|$.

Proof: For each $x\in\bar{D}$,

$$\begin{aligned} |(A_k v - A_k w)(x)| &= \left|\int_{\bar{D}} d\xi\, g_k(x,\xi) \right. \\ &\qquad \left. \times [B_k(v,w)(v-w)](\xi)\right| \\ &\leqslant M\|B_k(v,w)\|\,\|v-w\|, \end{aligned}$$

as in Step 2. The result follows by observing that $\|B_k(v,w)\| \leqslant (k+\gamma)$, as in Step 1.

Step 4: $\{u_n\}$ is a Cauchy sequence with respect to $|||\cdot|||$.

Proof: A proof of the fact that $\{u_n\}\subset S$ is a Cauchy sequence in H, i.e.,

$$\|u_n - u_m\| \underset{n,m\to\infty}{\to} 0,$$

follows from Step 1. Since $\{u_n\}_0^\infty \subset S$, from Step 2, $\{u_n\}_1^\infty$ is bounded. From Step 3,

$$\begin{aligned} |||u_{n+1} - u_{m+1}||| &= |||A_k u_n - A_k u_m||| \\ &\leqslant M(k+\gamma)\|u_n - u_m\| \underset{n,m\to\infty}{\to} 0 \end{aligned}$$

for $\{u_n\}$ is a Cauchy sequence in H.

Step 5: Result of Theorem 1.

Proof: Since S is closed with respect to $|||\cdot|||$, it follows from Step 4 that there exists a $u\in S$ such that

$$\lim_{n\to\infty} |||u_n - u||| = \lim_{n\to\infty} |||A_k u_n - u||| = 0.$$

Since A_k is defined on S, $A_k u$ is well defined. From Step 3,

$$|||A_k u_n - A_k u|||\leqslant M(k+\gamma)\|u_n - u\|$$
$$\leqslant \bar{M}(k+\gamma)|||u_n - u|||$$
$$\underset{n\to\infty}{\to} 0,$$

where $\bar{M}$ is a constant. This implies that $u = A_k u$. To show the uniqueness, let $A_k \tilde{u} = \tilde{u} \in S$. Then from Step 1,

$$\|u - \tilde{u}\| = \|A_k u - A_k \tilde{u}\|$$
$$\leqslant \mu_k \|u - \tilde{u}\|$$
$$= 0$$

for $\mu_k < 1$. This implies that $u = \tilde{u}$ almost everywhere. However, $u = \tilde{u}$ everywhere, for, using Step 3,

$$|||u - \tilde{u}||| = |||A_k u - A_k \tilde{u}|||$$
$$\leqslant M(k+\gamma)\|u - \tilde{u}\|$$
$$= 0.$$

Since $\mu_k = (k+\gamma)/(k+\lambda)$ of Theorem 1, Step 1, is an increasing function of k on $(-\lambda, \infty)$, the rate of convergence is expected to improve with decreasing values of k within the restriction imposed by (C4). Thus, for computational purposes, it is desirable to take $k = \bar{k}$. Further, it is to one's advantage to obtain tight bounds on f', which can be checked by considering the behavior of $\bar{\mu} = (\bar{k}+\gamma)/(\bar{k}+\lambda)$ with respect to the variations of γ and N. This point was discussed also in Ref. 2.

For f in a smaller class of functions, a stronger restriction on k enables one to obtain a monotonically convergent $\{u_n\}$, which we show in Proposition 1. This result may prove useful when a slower rate of convergence may be tolerated in favor of the monotonicity.

Proposition 1: In addition to the conditions of Theorem 1, let $u_0 = 0 \in S$, $f(0) \geqslant 0$ for all $x \in \bar{D}$, and let $k \geqslant N$. Then, in addition to the result of Theorem 1, $\{u_n\}$ converges monotonically to u from below.

Proof: We need show only that $\{u_n\}$ is a nondecreasing sequence. Since $g_k(x,\xi) \geqslant 0$,

$$u_1(x) = \int_{\bar{D}} d\xi\, g_k(x,\xi) f(\xi;0) \geqslant 0 = u_0(x).$$

Let $u_n \geqslant u_{n-1}$; we have that

$$(u_{n+1} - u_n)(x) = \int_{\bar{D}} d\xi\, g_k(x,\xi)\,[B_k(u_n, u_{n-1})(u_n - u_{n-1})](\xi)$$
$$\geqslant 0$$

for

$$k + \int_0^1 dt\, f'[t u_n + (1-t)u_{n-1}] \geqslant 0$$

implying that

$$[B_k(u_n, u_{n-1})(u_n - u_{n-1})](\xi) \geqslant 0.$$

The result follows by the induction principle.

We have used the non-negativity of $g_k(x,\xi)$ in Proposition 1, which was not needed in Theorem 1.

If we take $S = H$ in Theorem 1, then (C1) is clearly satisfied. If (C2) and (C3) hold and k is chosen according to (C4), then (C5) is also satisfied. Therefore the result of Theorem 1 is valid with $S = H$. This generalizes the result of Ref. 1. However, the requirement that (C2) be valid for all $v \in H$ is too stringent to be satisfied by a large number of problems of interest. An attempt at relaxing this requirement was made in Ref. 2, but the new conditions are too limiting to be useful. In particular, it was assumed that, for all non-negative v bounded by a fixed constant, $0 \geqslant f(v) \geqslant -\frac{1}{2}(\delta + N)v$, with some δ, $N \geqslant 0$. By setting $v = 0$ one has that $f(0) = 0$, which is serious limitation.

In the following, we give some useful characterizations of sets that may replace S in Theorem 1 and Proposition 1.

Let $f(0) \in H$ and

$$\bar{u}_\xi = (L+\xi)^{-1}|f(0)|, \quad \xi > -\lambda.$$

Non-negativity of $g_\xi(x,\xi)$ implies that $\bar{u}_\xi \geqslant 0$. It is also bounded as in Step 2 of Theorem 1. Since

$$(L+k)\bar{u}_\xi = |f(0)| + (k-\zeta)\bar{u}_\xi$$

it follows that

$$\bar{u}_\xi = (L+k)^{-1}[|f(0)| + (k-\zeta)\bar{u}_\xi], \quad k > -\lambda.$$

Let

$$Q_\xi = \{v:\ |v| \leqslant \bar{u}_\xi\}$$

and let

$$P_\xi = \{v:\ 0 \leqslant v \leqslant \bar{u}_\xi\}.$$

It is straightforward to check that (C1) and (C3) hold on Q_ξ and P_ξ. Assuming that (C2) is satisfied, one may restrict k according to (C4). In Proposition 2, we determine conditions that imply (C5) with $S = Q_\xi$ and $S = P_\xi$.

Proposition 2: Let the symbols be as above. If there exist ξ and k such that, for all $x \in \bar{D}$,

(i) for each $v \in P_\xi$, $\quad 0 \leqslant kv + f(v) \leqslant f(0) + (k-\zeta)\bar{u}_\xi$,

(ii) for each $v \in Q_\xi$, $\quad 0 \leqslant |kv + f(v)| \leqslant |f(0)| + (k-\zeta)\bar{u}_\xi$,

then (i) $A_k P_\xi \subseteq P_\xi$ and (ii) $A_k Q_\xi \subseteq Q_\xi$.

Proof: We give a proof of (i); a proof of (ii) follows by similar arguments by estimating $|A_k v|$ instead of $A_k v$.

For each $v \in P_\xi$, we have

$$0 \leqslant (A_k v)(x) = \int_{\bar{D}} d\xi\, g_k(x,\xi)\,[kv(\xi) + f(\xi; v(\xi))]$$
$$\leqslant \int_{\bar{D}} d\xi\, g_k(x,\xi)\,[f(\xi;0) + (k-\zeta)\bar{u}_\xi(\xi)]$$
$$= \bar{u}_\xi(x).$$

It is clear that, if (C2) holds on P_ξ, Q_ξ and with k in conformity with (C4), the corresponding condition of Proposition 2 also holds; then (C1)–(C5) are satisfied with $S = P_\xi$, $S = Q_\xi$, implying the validity of the result of

Theorem 1. In case of P_ζ, if $u_0 = 0$ and $k \geqslant N$, then, from Proposition 1, a monotonic convergence results.

III. APPLICATION

It follows from the results of Sec. III of Ref. 2 that Eq. (3) has a unique solution u, satisfying

$$0 \leqslant u(x) \leqslant 1 - (\cosh \phi x)/(\cosh \phi) = \chi(x)$$

for each ϕ such that

$$(\cosh \phi - 1)/(\cosh \hat{k} - 1) \geqslant \phi^2/(\hat{k}^2 \cosh \phi),$$

where

$$\hat{k} = \tfrac{1}{2}\nu\phi^2[1 + (1/\cosh \phi)^{\nu-1}].$$

Also the iterative sequence generated by $\chi(x)$, with $k = \hat{k}$, was shown to converge to $u(x)$. The author considered Eq. (2); the results given here for Eq. (3) follow by setting $u = 1 - v$. In the following we use the results of Sec. II to show that the existence and convergence results hold for Eq. (3) for all real values of ϕ, with any $k \geqslant \hat{k}$ and any $u_0(x)$ such that $0 \leqslant u_0(x) \leqslant \chi(x)$.

First we note that, in this case, $l = 1$, $\overline{D} = [0,1]$, and $\lambda = \pi^2/4$. For $\nu = 1$, Eq. (3) is solved easily, yielding $u = \chi$.

Let $\nu \geqslant 2$. With $\zeta = \phi^2$, $\bar{u}_\zeta = \chi \leqslant 1$, and $0 \leqslant f(0) = \phi^2 \in H$. Let P_ζ be defined as in Sec. II, i.e., $v \in P_\zeta$ implies that $0 \leqslant v \leqslant \bar{u}_\zeta \leqslant 1$. Now for $v \in P_\zeta$,

$$-\delta = -\nu\phi^2(1/\cosh \phi)^{\nu-1} \geqslant f' \geqslant -\nu\phi^2.$$

Thus (C2) is satisfied with $\gamma = -\delta$ and $N = \nu\phi^2$. Let

$$k \geqslant \tfrac{1}{2}(\nu\phi^2 + \delta),$$

which satisfies (C4). For each $v \in P_\zeta$,

$$\phi^2 + (k - \phi^2)v = kv + \phi^2(1 - v) \geqslant kv + f(v) \geqslant 0.$$

For $\nu \geqslant 2$, $k \geqslant \phi^2$; consequently,

$$\phi^2 + (k - \phi^2)v \leqslant \phi^2 + (k - \phi^2)\bar{u}_\zeta = f(0) + (k - \zeta)\bar{u}_\zeta,$$

and the condition (i) of Proposition 2 is satisfied.

Since the conditions (C1)–(C5) of Theorem 1 are satisfied, Eq. (4) has a unique solution, $u \in P_\zeta$, for each value of ϕ and the iterative sequence generated by any $u_0 \in P_\zeta$ converges uniformly to u. If we take $u_0 = 0$ and $k \geqslant \nu\phi^2$, then the convergence is also monotonic from below.

It is not necessary to restrict $\nu \geqslant 2$. In fact, $\nu \geqslant 1$ may be allowed to be a continuous parameter. For the above arguments to hold, all one has to do is to take

$$k \geqslant \max(\phi^2, \tfrac{1}{2}(\nu\phi^2 + \delta)).$$

[1] H. B. Keller, *Numerical Methods for Two-Point Boundary Value Problems* (Blaisdell, Waltham, MA, 1968), pp. 108–109.
[2] J. A. Pennline, Math. Comp. **37**, 127 (1981).
[3] R. Aris, *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysis* (Clarendon, Oxford, 1975), Vol. 1, pp. 101–239.
[4] N. Aronszajn and K. Smith, Am. J. Math. **79**, 611 (1957); see also, H. B. Keller, in *Bifurcation Theory and Nonlinear Eigenvalue Problems*, edited by J. B. Keller and S. Antman (Benjamin, New York, 1969), pp. 217–255.
[5] S. R. Vatsya, J. Math. Phys. **22**, 2977 (1981).

On the existence of infinitely many resonances in the scattering problem based on moment conditions and entire functions

J. S. Hwang
Institute of Mathematics, Academia Sinica, Taipei, Taiwan, China 11529

(Received 7 August 1986; accepted for publication 30 January 1987)

It has been well proved that there are infinitely many resonances in the scattering problem provided the potential $V(t)$ satisfies the moment conditions $\int_0^\infty |V(t)|t^n\,dt = O(n^{(1-\epsilon)n})$, $\epsilon>0$, $n=0,1,\ldots$. In particular, if $V(t)$ has a compact support then the result of Rollnik [H. Rollnik, Z. Phys. **145**, 654 (1956)] and Regge [T. Regge, Nuovo Cimento, **8**, 671 (1958)] is obtained and if $|V(t)|\sim\exp(-\tau t^{1+\epsilon})$, $\tau,\epsilon>0$, we have the results of Sartori [L. Sartori, J. Math. Phys. **4**, 1408 (1963)].

I. INTRODUCTION

It is well known (see Agranovich and Marchenko,[1] p. 20) that the Schrödinger equation

$$y'' + z^2 y = V(x)y, \quad 0<x<\infty,$$

has a solution $E(x,z)$ given by

$$E(x,z) = e^{izx} + \int_x^\infty K(x,t)e^{izt}\,dt, \tag{1}$$

where the kernel $K(x,t)$ satisfies

$$|K(x,t)| \leqslant \tfrac{1}{2}e^{\sigma_1(x)}\sigma[(x+t)/2], \tag{2}$$

where

$$\sigma(x) = \int_x^\infty |V(t)|dt < \infty$$

and

$$\sigma_1(x) = \int_x^\infty |V(t)|t\,dt < \infty.$$

For $x=0$, the function $f(z)=E(0,z)$ is analytic in $\operatorname{Im} z>0$ (the Jost function, see Newton,[2] p. 340) and as usual, a zero of $f(z)$ is called a resonance in the scattering problem (excluding the bound state, see Ref. 2, p. 360). Naturally, we may ask under what condition on the potential $V(x)$ can there be infinitely many resonances? In this paper, we answer this question by the following.

Theorem 1: There are infinitely many resonances if the Jost function $f(z)$ is entire and the potential satisfies the moment conditions

$$\int_x^\infty |V(t)|t^n\,dt = O(n^{(1-\epsilon)n}), \quad x\geqslant 0, \quad \epsilon>0, \quad n=0,1,\ldots. \tag{3}$$

As a consequence of Theorem 1, we obtain the following known result due to Rollnik[3] and Regge.[4]

Corollary 2: There are infinitely many resonances if the potential has a compact support.

To prove Corollary 2, we need only observe that $|V(t)|=0$ for all $t\geqslant T$, where $T>0$ is fixed, so that

$$\int_x^\infty |V(t)|t^n\,dt \leqslant CT^{n+1}, \quad \text{for some } C>0.$$

This satisfies the moment conditions (3). Clearly, the Jost function is entire and hence the assertion follows from Theorem 1.

Although the regular solution is always an entire function, the Jost solution and the scattering operator are only entire functions for potentials which decrease faster than an exponential at infinity, see Nussenzveig (Ref. 5, p. 201).

II. ENTIRE FUNCTIONS

As we saw in the Introduction, the resonances in the scattering problem are the same as the zeros of a certain class of entire functions. To study such a class, we call an entire function $f\in B$ if $f(z)$ is bounded in a half-plane H and $f(z)$ has a radial limit l, as $z\to\infty$, along the boundary of H, where we require f be nonconstant.

We note that the function $f(z)=E(0,z)$ defined in (1) belongs to class B but not the function e^{izx}, $x>0$. Functions in class B can possibly have no zeros as will be seen from the following.

Theorem 3: If $f\in B$, then $e^f\in B$ and the function $f(z)$ tends to l uniformly in H, as $z\to\infty$, so that $f(z)$ has at most finitely many zeros in H provided $l\neq 0$. If $f(z)$ is of exponential type τ and belongs to L^2 on the real axis, then $f(z)$ can be represented by

$$f(z) = \int_{-\tau}^{\tau} e^{izt}V(t)dt, \quad V\in L^2(-\tau,\tau).$$

Proof: We first observe that both of the boundedness and the radial limit of $f\in B$ are preserved by the exponential function and hence $e^f\in B$.

Next, in view of Montel's theorem (see Boas,[6] p. 5), we see that the function $f(z)$ tends to l uniformly in H, as $z\to\infty$, so that it cannot have infinitely many zeros in H provided $l\neq 0$.

Finally, the representation of $f(z)$ follows from the Paley–Wiener theorem (see Ref. 6, p. 103) and the proof is complete.

Note that the hypothesis $l\neq 0$ in the above result is necessary. For instance, the function

$$f(z) = (e^{iz} - e^{-1})/(z-i)$$

belongs to B in the upper half-plane with $l=0$ and has infinitely many zeros there at $i+2n\pi$, $n=1,2,\ldots$.

 0022-2488/87/061287-03$02.50

Note that the function e^f is of infinite order and therefore the condition on the finite order is what we need to have infinitely many zeros as will be seen from the following.

Theorem 4: If $f \in B$ and if f is of finite order, then f has infinitely many zeros.

Proof: Suppose on the contrary that f has only finitely many zeros. Then f can be represented by

$$f(z) = e^{P_n(z)} Q_m(z),$$

where P_n and Q_m are polynomials of degree n and m, respectively. Without loss of generality, we may assume that the half-plane in question is the upper half-plane $U = \{z: \operatorname{Im} z > 0\}$. We set

$$P_n(z) = a_0 + a_1 z + \cdots + a_n z^n, \quad a_n = re^{i\alpha} \neq 0, \quad 0 \leqslant \alpha < 2\pi.$$

Then it is easy to see that $P_n(z) \to \infty$, as $z \to \infty$ along the ray $\Gamma = \{\mathrm{Re}^{i(2\pi - \alpha)/n}: \ 0 \leqslant R < \infty\}$. Clearly, if $n > 1$ then the ray Γ lies on U which contradicts the boundedness of f in U. Hence we must have $n = 1$ and

$$f(z) = e^{irz} Q_m(z).$$

This in turn implies that $f(z) \to \infty$, as $z \to \infty$ along the real axis, a contradiction again. We thus conclude that f has infinitely many zeros.

We note that any entire function which is not a polynomial takes every value in its range with one possible exception infinitely often due to Picard's theorem (see Titchmarsh,[7] p. 277). The function e^f, $f \in B$, has the exceptional value 0. However, any function considered in Theorem 4 has no exceptional values in its range as will be seen from the following extension.

Corollary 5: Under the hypothesis of Theorem 4, the function f assumes every value infinitely often.

Proof: For any value v, the function $f - v \in B$ and of finite order. Hence the assertion follows from Theorem 4.

Theorem 6: Let $f(z) = c + \int_x^\infty K(x,t) e^{izt}\, dt$, where c is a constant, $x \geqslant 0$ is fixed, and $K(x,t)$ satisfies

$$\int_x^\infty |K(x,t)| t^n\, dt = O(n^{(1-\epsilon)n}), \quad \epsilon > 0, \quad n = 0,1,\ldots . \tag{4}$$

If $f(z)$ is an entire function, then it is of order $\leqslant 1/\epsilon$ and has infinitely many zeros.

Proof: Since $f(z)$ is entire, it can be expanded as

$$f(z) = \sum_{n=0}^{\infty} a_n z^n,$$

where

$$a_n = \frac{(-i)^n}{n!} \int_x^\infty K(x,t) t^n\, dt.$$

Using (4) and Stirling's formula, we see that $f(z)$ is an entire function of order (see Ref. 6, p. 9)

$$\rho = \lim_{n \to \infty} \sup \frac{n \log n}{\log(1/|a_n|)} \leqslant \lim_{n \to \infty} \frac{n \log n}{n\epsilon(\log n - 1)} = \frac{1}{\epsilon}.$$

Since the function $f(z)$ is bounded in $\operatorname{Im} z \geqslant 0$ and $f(z) \to c$, as $z \to \pm \infty$ along the real axis, it follows from Theorem 4 that f has infinitely many zeros. This completes the proof.

Note that from Corollary 5 we see that the function

$$F(z) = \int_x^\infty K(x,t) e^{izt}\, dt$$

assumes any value infinitely often in the lower half-plane.

With the help of Theorem 6, we are now ready to prove Theorem 1.

III. PROOF OF THEOREM 1

According to (2), we have the inequality

$$|K(x,t)| \leqslant \tfrac{1}{2} e^{\sigma_1(x)} \sigma((x+t)/2).$$

It follows from (3) that

$$\begin{aligned}
\int_x^\infty |K(x,t)| t^n\, dt &\leqslant \frac{1}{2} e^{\sigma_1(x)} \int_x^\infty \left(\int_{(x+t)/2}^\infty |V(s)|\, ds \right) t^n\, dt \\
&= \frac{1}{2} e^{\sigma_1(x)} \int_x^\infty |V(s)|\, ds \int_x^{2s-x} t^n\, dt \\
&< \frac{1}{n+1} e^{\sigma_1(x)} \int_x^\infty |V(s)| (2s - x)^{n+1}\, ds \\
&= \frac{2^{n+1}}{n+1} e^{\sigma_1(x)} O((n+1)^{(1-\epsilon)(n+1)}) \\
&= O(n^{(1-\epsilon)n}).
\end{aligned}$$

Hence the kernel $K(x,t)$ satisfies condition (4) in Theorem 6 and the assertion follows from that theorem.

As a consequence of Theorem 1, we obtain Corollary 2 and the following result of Sartori.[8]

Corollary 7: There are infinitely many resonances if the potential satisfies

$$|V(t)| \sim \exp(-\tau t^{1+\epsilon}), \quad \tau, \epsilon > 0. \tag{5}$$

Proof: Clearly condition (5) gives

$$\int_0^\infty \exp(-\tau t^{1+\epsilon}) t^n\, dt = ((1+\epsilon)\tau^N)^{-1} \Gamma(N+1),$$

where $N = (1+n)/(1+\epsilon)$ and $\Gamma(N+1) = N! < N^N < n^{(1-\epsilon/2)n}$ for all $n > 2(\epsilon(1-\epsilon))^{-1}$, $0 < \epsilon < 1$ (it suffices to prove the case $\epsilon < 1$). It follows that

$$\int_0^\infty |V(t)| t^n\, dt = O(n^{(1-\delta)n}), \quad 0 < \delta < \frac{\epsilon}{2}.$$

This together with Theorem 1 yields the assertion.

IV. CONCLUSIONS AND OPEN PROBLEMS

Note that the method used does not generalize to potentials with exponential or slower decrease at infinity.

In view of the function defined in Theorem 3,

$$f(z) = (e^{iz} - e^{-1})/(z - i),$$

which has infinitely many zeros in U, we may ask whether there is a potential $V(t)$ such that the function

$$f_V(z) = \int_0^\infty e^{izt} V(t)\, dt$$

is entire and has infinitely many zeros in U. In general, we do not know the answer. However, if we replace U by the closure $\overline{U}$, then such a potential does exist. For instance, we may take $V(t) = 1$, $0 \leqslant t \leqslant 2\pi$, and $V(t) = 0$, $t > 2\pi$. Then

$$f_V(z) = (e^{i2\pi z} - 1)/(iz),$$

which has infinitely many zeros at $z = 1,2,\ldots$.

Naturally, we may ask whether there is a potential such that

$$(e^{iz} - e^{-1})/(z - i) = \int_0^\infty e^{izt} V(t)dt. \tag{6}$$

The answer should be negative. To see this, we expand both sides of (6) and we then have

$$\int_0^\infty t^n e^{-t} V(t)dt = e^{-1}, \quad n = 0,1,\ldots . \tag{7}$$

We conjecture that no summable function $f(t)$ can satisfy the following condition of constant moment:

$$\int_0^\infty t^n f(t)dt = c \neq 0, \quad n = 0,1,\ldots . \tag{8}$$

Clearly, the nonexistence of the potential $V(t)$ in (7) is a particular case in (8). However, if $c = 0$, then there does exist a summable function satisfying (8). For instance, the function (see Shohat and Tamarkin,[9] p. 22)

$$f(t) = (\sin t^{1/4})\exp(-t^{1/4}), \quad 0 \leqslant t < \infty,$$

is summable and satisfies (8) when $c = 0$.

Of course, if we consider the distribution $d\mu(t)$ in place of the density $f(t)dt$, then there does exist a distribution satisfying

$$\int_0^\infty t^n \, d\mu(t) = c, \quad n = 0,1,\ldots . \tag{9}$$

In fact, we can define the point mass: $\mu(1) = c$, $t \geqslant 1$, and $\mu(t) = 0$, $t < 1$, then (9) obviously holds.

We now explain the reason for our conjecture with regard to (8). More precisely, we shall prove that no potentials with compact supports can satisfy (8). Suppose on the contrary that there is such a potential $V(t) = 0$, $t \geqslant T$, satisfying (8). Then we have

$$\int_0^T t^n V(t)dt = c \neq 0, \quad n = 0,1,\ldots . \tag{10}$$

Subtracting two consecutive terms in (10) gives

$$\int_0^T t^n (1 - t) V(t)dt = 0, \quad n = 0,1,\ldots .$$

It follows from the Müntz–Szász theorem (see Rudin,[10] p. 304) that $(1 - t)V(t) = 0$ or $V(t) = 0$ holds almost everywhere on $(0, T)$. This in turn implies that $c = 0$, a contradiction. We thus conclude that no potentials with compact supports can satisfy (8).

We note that the above assertion can also be proved via a different approach (Ref. 9, p. 5). Also note that some extensions of the Müntz–Szász theorem with regard to the completeness have been done by the author.[11]

In closing this note, we finally pose the following moment problem: What is a necessary and sufficient condition on a sequence $\{\mu_n\}$ of real numbers such that there is a potential $V(t)$ satisfying

$$\int_0^\infty t^n V(t)dt = \mu_n, \quad n = 0,1,\ldots ?$$

ACKNOWLEDGMENT

I am indebted to the referee for his valuable comments.

[1] Z. S. Agranovich and V. A. Marchenko, *The Inverse Problem of Scattering Theory* (Gordon and Breach, New York, 1963).

[2] R. G. Newton, *Scattering Theory of Waves and Particles* (McGraw-Hill, New York, 1982), 2nd ed.

[3] H. Rollnik, "Zur theorie der zerfallenden zustände," Z. Phys. **145**, 654 (1956).

[4] T. Regge, "Analytic properties of the scattering matrix," Nuovo Cimento **8**, 671 (1958).

[5] H. M. Nussenzveig, *Causality and Dispersion Relations* (Academic, New York, 1972).

[6] R. P. Boas, *Entire Functions* (Academic, New York, 1954).

[7] E. C. Titchmarsh, *The Theory of Functions* (Oxford U. P., New York, 1939), 2nd ed.

[8] L. Sartori, "Asymptotic behavior of Schrödinger scattering amplitudes," J. Math. Phys. **4**, 1408 (1963).

[9] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments* (Am. Math. Soc., Providence, RI, 1943).

[10] W. Rudin, *Real and Complex Analysis* (McGraw-Hill, New York, 1966).

[11] J. S. Hwang, "A note on Bernstein and Müntz–Szász theorems with applications to the order statistics," Ann. Inst. Statist. Math. **30**, 167 (1978), Part A.

Do trilinear commutation relations in quantum mechanics admit coordinate space realization in three dimensions?

Ranjan Bhattacharya and Siddhartha Bhowmick
Department of Physics, Jadavpur University, Calcutta 700032, India

(Received 7 October 1985; accepted for publication 14 January 1987)

The trilinear commutation relations involving coordinates and momenta introduced by Wigner [E. P. Wigner, Phys. Rev. 77, 711 (1950)] are generalized to three dimensions. It is shown that the only realizable coordinate space representation of the momenta implies the usual bilinear commutation relations.

I. INTRODUCTION

It is well known that in quantum mechanics trilinear commutation relations involving coordinates and momenta, rather than the usual bilinear ones, can be introduced for a one-dimensional harmonic oscillator.[1] A nontrivial coordinate space representation of the momentum operator necessitates the use of wave functions that are not analytic in the usual sense. It seems worthwhile to investigate whether similar conclusions are imperative for the same problem in three dimensions.

The trilinear commutation relations for the one-dimensional oscillator follow by requiring that the equations of motion obtained from the Heisenberg equations are the same as the classical ones. The trilinear relations thus obtained are

$$[q,\{q,p\}] = 2iq, \quad [p,\{q,p\}] = -2ip\,, \tag{1}$$

where the brackets $\{\ \}$ and $[\ ]$ refer, respectively, to an anticommutator and a commutator. Yang[2] found the coordinate representation of the momentum operator p,

$$p = -i\frac{d}{dq} + i\frac{C}{q}R\,, \tag{2}$$

where C is a real constant, and R the inversion operator, i.e.,

$$RqR^{-1} = -q\,, \quad R\frac{d}{dq}R^{-1} = -\frac{d}{dq}\,. \tag{3}$$

In trying to solve the oscillator problem with the representation (2), Yang concluded that C must be zero if one requires the wave functions to be analytic. Ohnuki and Kamefuchi,[3,4] by introducing generalized functions, or "hyperfunctions" as wave functions, solved the oscillator problem with $C \neq 0$.

II. TRILINEAR COMMUTATION RELATIONS IN THREE DIMENSIONS

We now consider what generalizations can be made when three degrees of freedom, rather than a single one, are involved. We, therefore, consider a three-dimensional oscillator with the Hamiltonian

$$H = \frac{1}{2}\sum_k (p_k^2 + x_k^2)\,. \tag{4}$$

Here, and subsequently, all the indices run over the values 1, 2, and 3. The demand that the classical equations of motion follow from the Heisenberg equations leads to the trilinear relations

$$[x_k,\{p_l,x_k\}] = 2i\delta_{kl}x_k, \quad [p_k,\{x_l,p_k\}] = -2i\delta_{kl}p_k\,. \tag{5}$$

We assume relations more general than the ones above and write

$$\begin{aligned}
&\text{(i)}\ [x_k,\{p_l,x_m\}] = 2i\delta_{kl}x_m\,,\\
&\text{(ii)}\ [p_k,\{x_l,p_m\}] = -2i\delta_{kl}p_m\,,\\
&\text{(iii)}\ [p_k,\{x_l,x_m\}] = -2i\delta_{kl}x_m - 2i\delta_{km}x_l\,,\\
&\text{(iv)}\ [x_k,\{p_l,p_m\}] = 2i\delta_{kl}p_m, + 2i\delta_{km}p_l\,.
\end{aligned} \tag{6}$$

Not all of the above relations are independent. In fact, (iii) and (iv) follow from (i) and (ii), respectively, by using the (generalized) Jacobi identity. If we now introduce the operators a_k and $a_k^\dagger$ defined by

$$a_k \equiv (1/\sqrt{2})(x_k + ip_k), \quad a_k^\dagger \equiv (1/\sqrt{2})(x_k - ip_k)\,, \tag{7}$$

the relations (6) become in terms of these operators,

$$\begin{aligned}
&[a_k,\{a_l^\dagger,a_m\}] = 2\delta_{kl}a_m\,,\\
&[a_k,\{a_l^\dagger,a_m^\dagger\}] = 2\delta_{kl}a_m^\dagger + 2\delta_{km}a_l^\dagger\,,\\
&[a_k,\{a_l,a_m\}] = 0\,.
\end{aligned} \tag{8}$$

These are the well-known trilinear relations for the creation and annihilation operators for a system of para-Bose oscillators.[5]

III. COORDINATE SPACE REALIZATION

To find the coordinate representation of the momenta p_i satisfying the trilinear relations (6), consider the operator

$$S_{ij} \equiv [x_i,p_j] - i\delta_{ij}\,. \tag{9}$$

Therefore

$$S_{ij}^\dagger = -S_{ij}\,. \tag{10}$$

Then using the relations (6), we have

$$\{p_k,S_{ij}\} = \{x_k,S_{ij}\} = 0\,. \tag{11}$$

Taking the matrix element of the second relation of (11) between the states $|\vec{x}'\rangle$ and $|\vec{x}''\rangle$, we get

$$\langle\vec{x}'|\{x_k,S_{ij}\}|\vec{x}''\rangle = (x_k' + x_k'')\langle\vec{x}'|S_{ij}|\vec{x}''\rangle = 0\,. \tag{12}$$

Or,

 0022-2488/87/061290-03$02.50

$$\langle \vec{x}'|S_{ij}|\vec{x}''\rangle = 2iC_{ij}(\vec{x}')\delta(\vec{x}'+\vec{x}''), \tag{13}$$

where the $2i$ has been introduced for convenience. Thus

$$S_{ij} = 2iC_{ij}(\vec{x})R. \tag{14}$$

Here R is the inversion operator, i.e.,

$$R\Psi(\vec{x}) = \Psi(-\vec{x}).$$

We also note, using (10),

$$C^*_{ij}(\vec{x}) = C_{ij}(-\vec{x}). \tag{15}$$

The matrix element of S_{ij} is also given by

$$\langle \vec{x}'|S_{ij}|\vec{x}''\rangle = (x_i' - x_i'')\langle \vec{x}'|p_j|\vec{x}''\rangle - i\delta_{ij}\delta(\vec{x}'-\vec{x}'')$$
$$= 2iC_{ij}(\vec{x}')\delta(\vec{x}'+\vec{x}''),$$

using (9) and (13).

We therefore have

$$\langle \vec{x}'|p_j|\vec{x}''\rangle = -i\frac{\partial}{\partial x_j}\delta(\vec{x}'-\vec{x}'') + \frac{iC_{ij}(\vec{x}')\delta(\vec{x}'+\vec{x}'')}{x_i'}$$
$$+ B_j(\vec{x}')\delta(\vec{x}'-\vec{x}''), \tag{16}$$

where we have used

$$x_i\frac{\partial}{\partial x_j}\delta(\vec{x}) = -\delta_{ij}\delta(\vec{x}).$$

Here $B_j(\vec{x}')$ is an arbitrary real function as p_j is Hermitian. Since the right-hand side of (16) cannot depend on the index i, $C_{ij}(\vec{x}')$ must be proportional to x_i'. Now, from the requirement of the correct transformation properties of p_j under rotations, we must have

$$C_{ij}(\vec{x}') = x_i x_j f(|\vec{x}|), \quad B_j(\vec{x}) = x_j g(|\vec{x}|). \tag{17}$$

Hence we have from (16),

$$\langle \vec{x}|p_j|\Psi\rangle = -i\frac{\partial}{\partial x_j}\Psi(\vec{x}) + ix_j f(|\vec{x}|)\Psi(-\vec{x})$$
$$+ x_j g(|\vec{x}|)\Psi(\vec{x}), \tag{18}$$

whence,

$$p_j = -i\frac{\partial}{\partial x_j} + ix_j f(|\vec{x}|)R + x_j g(|\vec{x}|). \tag{19}$$

We know, however, that the arbitrary function $B_j(\vec{x})$ in (16) can be gauged away with a unitary transformation if the integrability condition $\partial B_j/\partial x_i = \partial B_i/\partial x_j$ holds. We find from (17) that this condition is obviously satisfied. It is easy to see that under this transformation $C_{ij}(\vec{x}) \to C'_{ij}(\vec{x})$ which also satisfies (15). The argument leading to the explicit form of C_{ij} in (17) also holds for C'_{ij}. Therefore we have

$$p_j = -i\frac{\partial}{\partial x_j} + ix_j f(|\vec{x}|)R. \tag{20}$$

The form of f can now be determined from the first of the relations (10). We have, using (14),

$$\{p_k, S_{ij}\} = \left\{-i\frac{\partial}{\partial x_k} + ix_k f(|\vec{x}|)R,\ 2iC_{ij}(\vec{x})R\right\} = 0. \tag{21}$$

This gives, finally,

$$\delta_{ik}x_j f(|\vec{x}|)R + \delta_{jk}x_i f(|\vec{x}|)R + x_i x_j\frac{\partial}{\partial x_k}f(|\vec{x}|)R = 0. \tag{22}$$

Or, since $R^{-1} = R$ exists,

$$\delta_{ik}x_j f(|\vec{x}|) + \delta_{jk}x_i f(|\vec{x}|) + \frac{x_i x_j x_k}{|\vec{x}|}\frac{d}{d|\vec{x}|}f(|\vec{x}|) = 0. \tag{23}$$

Equation (23) has the solution

$$f(|\vec{x}|) = c\delta(|\vec{x}|). \tag{24}$$

This does not contribute to p_j as $x_j\delta(|\vec{x}|) = 0$. Therefore we have for the coordinate representation of the momentum operator,

$$p_i = -i\frac{\partial}{\partial x_i}. \tag{25}$$

It is clear that this representation necessarily implies the canonical bilinear commutation relations between x and p. This, of course, does not rule out the possible existence of other inequivalent representations. In fact, the usual parastatistics does provide a representation of trilinear commutation relations in terms of the so called generalized Bose numbers (see the Appendix). We therefore conclude that the generalization of Wigner's trilinear commutation relations for coordinates and momenta does not admit a nontrivial coordinate representation in three dimensions.

IV. CONCLUSION

To summarize, in one dimension a nontrivial coordinate representation of the momentum operator satisfying trilinear commutation relations is possible at the expense of introducing generalized functions as wave functions. A generalization to three dimensions shows that the only realizable coordinate space representation of the momentum operator implies the usual bilinear commutation relations. In Schrödinger quantum mechanics, therefore, there is no place for trilinear commutation relations in three dimensions.

ACKNOWLEDGMENTS

We thank the referee for pointing out the general solution (24) of Eq. (23). We are grateful to Dr. A. Chatterjee and Professor B. Dutta Roy for giving more help than they admit.

APPENDIX: GENERALIZED BOSE NUMBER REPRESENTATION

Representations of trilinear commutation relations exist in terms of generalized Bose numbers[4] which are the analogs of Green components.[5] Let

$$x_k = \sum_{\alpha=1}^{p} x_k^\alpha \quad \text{and} \quad p_k = -i\sum_{\alpha=1}^{p}\frac{\partial}{\partial x_k^\alpha},$$

where

$$\{x_k^\alpha, x_l^\beta\} = 0, \quad \alpha \neq \beta, \quad [x_k^\alpha, x_l^\alpha] = 0,$$

and

$$\left\{\frac{\partial}{\partial x_k^\alpha}, \frac{\partial}{\partial x_l^\beta}\right\} = 0, \quad \alpha \neq \beta, \quad \left[\frac{\partial}{\partial x_k^\alpha}, \frac{\partial}{\partial x_l^\alpha}\right] = 0,$$

with

$$\left\{\frac{\partial}{\partial x_k^\alpha}, x_l^\beta\right\} = 0\,, \quad \alpha \neq \beta, \quad \left[\frac{\partial}{\partial x_k^\alpha}, x_l^\alpha\right] = \delta_{kl}\,.$$

Then the above x and p satisfy the trilinear commutation relations (6).

[1]E. P. Wigner, Phys. Rev. **77**, 711 (1950).
[2]L. M. Yang, Phys. Rev. **84**, 788 (1951).
[3]Y. Ohnuki and S. Kamefuchi, J. Math. Phys. **19**, 67 (1978).
[4]Y. Ohnuki and S. Kamefuchi, *Quantum Field Theory and Parastatistics* (Springer, Berlin, 1982).
[5]H. S. Green, Phys. Rev. **90**, 270 (1953); O. W. Greenberg and A. Messiah, *ibid.* **138**, B1155 (1965).

Electromagnetic scattering of an arbitrary plane wave from a spherical shell with a circular aperture

Richard W. Ziolkowski
Electronics Engineering Department, Lawrence Livermore National Laboratory, P.O. Box 5504, L-156, Livermore, California 94550

William A. Johnson[a)]
Division 7553, Sandia National Laboratories, Albuquerque, New Mexico 87185

(Received 23 July 1985; accepted for publication 30 January 1987)

The problem of the scattering of an electromagnetic plane wave with arbitrary polarization and angle of incidence from a perfectly conducting spherical shell with a circular aperture is solved with a generalized dual series approach. This canonical problem encompasses coupling to an open spherical cavity and scattering from a spherical reflector. In contrast to the closed sphere problem, the electromagnetic boundary conditions couple the TE and TM modes. A pseudodecoupling of the resultant dual series equations system into dual series problems for the TE and TM modal coefficients is accomplished by introducing terms that are proportional to the associated Legendre functions P_0^{-m}. The solutions of the TE and TM dual series problems require the further introduction of terms proportional to P_n^{-m}, where $0 \leqslant n < m$. These functions effectively complete the standard spherical harmonic basis set when an aperture is present and guarantee the satisfaction of Meixner's edge conditions. Having generated the modal coefficients, all desired electromagnetic quantities follow immediately. Numerical results for the currents induced on the open spherical shell and for the energy density of the field at its center are presented for the case of normal incidence.

I. INTRODUCTION

The number of electromagnetic boundary value problems that can be solved exactly is rather small, especially in three dimensions. The desire and the need for these canonical problems, however, is very strong. They reveal the basic physics underlying the phenomena and help establish insights that can usually be extrapolated to more general situations. Moreover, they act as valuable test cases for general numerical approaches to related problems.

The scattering of an electromagnetic plane wave from a perfectly conducting closed sphere is probably the best known three-dimensional canonical scattering problem. Its generalization, the scattering of a plane wave from a perfectly conducting spherical shell with a circular aperture, is important from both theoretical and practical points of view. In particular, when the circular hole has a relatively small angular extent, this problem allows one to study the coupling of a wave from an external source through an aperture into an enclosed region. On the other hand, when the shell has a relatively small angular extent, the problem describes the scattering of a plane wave from a spherical reflector. A complete solution to this canonical mixed boundary value problem is given in this paper.

A Debye potential formulation is employed, but in contrast to standard treatments in spherical geometries, the associated Legendre polynomials of negative order (P_n^{-m}, $n \geqslant m$) are utilized for the modal expansions. Enforcement of the electromagnetic boundary conditions leads to a coupled set of dual series equations for the TE and TM modal coefficients of each azimuthal mode. A pseudodecoupling ansatz is developed to allow separate treatment of the TE and TM dual series systems. It requires the introduction of terms proportional to the associated Legendre polynomials P_0^{-m} (m being the azimuthal mode number) which are homogeneous solutions of the boundary condition equations. Solutions of the resulting "uncoupled" TE and TM dual series systems are given. They require the further introduction of terms proportional to the associated Legendre polynomials whose degree is less than its order: P_n^{-m}, where $0 \leqslant n < m$. These terms guarantee satisfaction of Meixner's edge conditions and effectively complete the spherical harmonic basis set in the presence of the aperture. Infinite systems of Fredholm equations of the second kind for the modal coefficients are obtained. A rigorous truncation procedure is given that leads to a straightforward numerical evaluation of those coefficients. Results for the currents induced on the open spherical shell and for an energy density ratio as a function of the fundamental parameter ka ($2\pi \times$ radius/wavelength) are presented for the case of normal incidence. It is shown analytically that the behavior of the currents near the edge of the aperture are in agreement with Meixner's edge conditions; the graphical results further confirm this. The energy density scans highlight the resonance features of the coupling physics.

This paper is organized as follows. In Sec. II the coupled dual series systems are derived for the scattering of a general plane wave from an open spherical shell. The decoupling ansatz is presented is Sec. III, and the resulting TE and TM dual series systems are solved in Sec. IV. The results are then restricted to the normal incidence case in Sec. V. In Sec. VI the currents induced on the open spherical shell are given for various values of ka, aperture size, and the two allowed angles of incidence. Their modal structure is exhibited with a set of three-dimensional color figures. The energy density scans are discussed in Sec. VII.

[a)] Present address: Division 1265, Sandia National Laboratories, Albuquerque, New Mexico 87185.

 0022-2488/87/061293-22$02.50

There have been several reports of solutions to the normally incident case of the open spherical shell problem from both analytical[1-10] and numerical[11-15] points of view. In the numerical papers, various convergence problems and erroneous results are encountered. Of the analytical papers only Ref. 9 seems to lead to correct results for the scattering problem. Unfortunately, direct comparisons for that case are difficult because the dual series systems and their solutions (which were checked with our validation scheme) differ from those obtained here and no current or field values were calculated there. Moreover, the basic tenets of Ref. 9 appear to be restricted to the normal incidence case. The errors in Refs. 1–8 and 10 are either that the dual series systems were solved incorrectly or, more fundamentally, that the wrong dual series systems were solved. The latter stems from the erroneous assumption that the TE and TM dual series are completely decoupled. This error is identical to the one made by Meixner in his original Debye potential solution to the scattering of a plane wave from a circular hole in a perfectly conducting ground plane.[16] In analogy with our approach, Meixner corrected that error in Ref. 17 by introducing additional potentials that were homogeneous solutions of the equations resulting from enforcement of the electromagnetic boundary conditions. The coefficients of these potentials were chosen to guarantee that the fields satisfy the correct edge behavior, hence accounting for the presence of the aperture. The pseudodecoupling ansatz can be shown to be equivalent to a gauge transformation, which in analogy with Dirac string analyses, involves discontinuous potentials, the gauge conditions being identical to the pseudodecoupling constraint conditions.[18]

The results for normal incidence were closely compared with those generated with a general, numerical surface patch scattering code; and these comparisons were reported in Ref. 19. The agreement is excellent; and since that code has been validated with a variety of different scattering problems, this lends further credence to the validity of the solution presented below. The present work represents a generalization of related aperture coupling work[20-24] to three dimensions. A more detailed presentation is available.[25] It includes many complementary results that were omitted here simply because of length considerations.

II. REDUCTION TO COUPLED DUAL SERIES PROBLEM

Consider the problem configuration shown in Fig. 1. A perfectly conducting open thin spherical shell is represented by the surface $r = a$, $0 \leqslant \theta < \theta_0$ in the spherical coordinate system (r,θ,ϕ) erected at the shell's center. The negative z axis of that system passes through the center of the aperture, the latter being defined as $\{(r,\theta,\phi)\,|\,r = a \text{ and } \theta_0 < \theta \leqslant \pi\}$. The opening angle of the aperture, θ_{ap}, is defined simply as $\theta_{ap} = \pi - \theta_0$. The medium inside and outside the shell is free space. The unit vectors $(\hat{r},\hat{\theta},\hat{\phi})$ are defined in the standard manner in the directions of positively increasing coordinate values.

Mathematically, we are seeking, for an arbitrary incident plane wave, the field scattered by the open spherical shell. This scattered field must satisfy the Sommerfeld radiation condition as $r \to \infty$. The total field (incident + scattered) must satisfy (1) the electromagnetic conditions, $E_{\tan} = 0$ on the metallic shell and $H_{\tan}$ continuous over the aperture; and (2) Meixner's edge conditions,[17,26] i.e., the total energy of the field must be finite near the aperture rim.

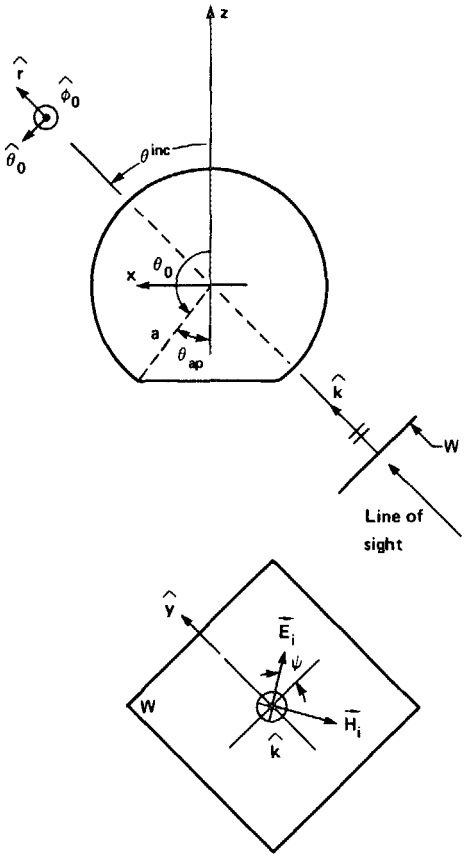

FIG. 1. Configuration of the scattering of an arbitrary plane wave from a spherical shell having a circular aperture.

A. Debye expansions

A plane wave with electric field strength E_0 is incident on the open spherical shell. It is characterized by a wave vector $\mathbf{k}$, which for convenience is assumed to lie in the xz plane; an incident angle θ^{inc} with respect to the z axis so that $\mathbf{k}\cdot\hat{z} = \cos\theta^{inc}$; and a polarization angle ψ between $\mathbf{E}$ and the projection of the positive z axis on the incident wave front. The incident field has the form

$$\begin{bmatrix} \mathbf{E}^{inc} \\ Z_0\mathbf{H}^{inc} \end{bmatrix} = -E_0 e^{i\mathbf{k}\cdot\mathbf{r}} \begin{bmatrix} (\cos\psi)\hat{\theta}_0 - (\sin\psi)\hat{\phi}_0 \\ (\sin\psi)\hat{\theta}_0 + (\cos\psi)\hat{\phi}_0 \end{bmatrix}, \quad (2.1)$$

where $\hat{\theta}_0$ and $\hat{\phi}_0$ are the incident polarization vectors and where, as throughout this paper, an $e^{-i\omega t}$ time dependence has been assumed and suppressed. The free-space impedance Z_0 is related to the free-space admittance $Y_0 = (\epsilon/\mu)^{1/2}$ as $Z_0 = Y_0^{-1}$ and the wave number $k = \omega(\epsilon\mu)^{1/2}$, where ϵ and μ denote, respectively, the permittivity and permeability of free space. The incident field parameters are indicated in Fig. 1. The incident electric field is polarized perpendicular to the edge of the aperture when $\psi = 0$ and is polarized parallel to it when $\psi = \pi/2$. Since any incident plane wave can be reduced to a linear superposition of these waves, only they

TABLE I. The electric and magnetic field components in spherical coordinates in terms of Debye potentials.

$$E_r = -\frac{1}{i\omega\epsilon}\left\{\frac{\partial^2}{\partial r^2}+k^2\right\}(r\Psi)$$

$$E_\theta = -\frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}(r\Phi) - \frac{1}{(i\omega\epsilon)r}\frac{\partial^2}{\partial r\,\partial\theta}(r\Psi)$$

$$E_\phi = \frac{1}{r}\frac{\partial}{\partial\theta}(r\Phi) - \frac{1}{(i\omega\epsilon)r\sin\theta}\frac{\partial^2}{\partial r\,\partial\phi}(r\Psi)$$

$$H_r = -\frac{1}{i\omega\mu}\left\{\frac{\partial^2}{\partial r^2}+k^2\right\}(r\Phi)$$

$$H_\theta = \frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}(r\Psi) - \frac{1}{(i\omega\mu)r}\frac{\partial^2}{\partial r\,\partial\theta}(r\Phi)$$

$$H_\phi = -\frac{1}{r}\frac{\partial}{\partial\theta}(r\Psi) - \frac{1}{(i\omega\mu)r\sin\theta}\frac{\partial^2}{\partial r\,\partial\phi}(r\Phi)$$

need to be addressed explicitly. Moreover, with the intrinsic symmetry of the field components in Maxwell's equations, the $\psi=\pi/2$ case is readily obtained from the $\psi=0$ case. Consequently, we restrict our considerations to the $\psi=0$ case with no loss in generality.

Following standard analyses of problems in a spherically symmetric geometry, we employ a Debye potential formalism.[27–29] In particular, the electric and magnetic fields are expressed in terms of the two vector potentials $\Phi\mathbf{r}$ and $\Psi\mathbf{r}$ as

$$\mathbf{E} = -\operatorname{curl}(\Phi\mathbf{r}) - (i\omega\epsilon)^{-1}\operatorname{curl}\operatorname{curl}(\Psi\mathbf{r}), \tag{2.2}$$

$$\mathbf{H} = +\operatorname{curl}(\Psi\mathbf{r}) - (i\omega\mu)^{-1}\operatorname{curl}\operatorname{curl}(\Phi\mathbf{r}), \tag{2.3}$$

where the radial vector $\mathbf{r}=r\hat{r}$. Their components are given explicitly in Table I. The scalar functions Φ and Ψ may represent any combination of the incident and scattered fields. The function Φ defines the field TE with respect to r, Ψ the field TM with respect to r. The descriptor "with respect to r" is assumed and suppressed throughout the rest of this paper.

The spherical wave expansion of the incident field (2.1) with $\psi=0$ given, for instance, in Ref. 30 or Ref. 31, can be generated with the Debye scalar potentials, Φ^{inc} and Ψ^{inc}, defined below. Since the scattered potentials, Φ^s and Ψ^s, assume an analogous form, we have

$$\begin{pmatrix}\Phi^{\text{inc}}\\ \Phi^s\end{pmatrix} = -E_0\sum_{m=0}^{\infty}\begin{pmatrix}\Phi_m^{\text{inc}}\\ \Phi_m^s\end{pmatrix}\sin m\phi, \tag{2.4}$$

$$\begin{pmatrix}\Psi^{\text{inc}}\\ \Psi^s\end{pmatrix} = Y_0E_0\sum_{m=0}^{\infty}\begin{pmatrix}\Psi_m^{\text{inc}}\\ \Psi_m^s\end{pmatrix}\cos m\phi, \tag{2.5}$$

where the azimuthal modal coefficients

$$\Phi_m^{\text{inc}} = \sum_{n=m}^{\infty}\gamma_{mn}\left[\frac{mP_n^m(\cos\theta^{\text{inc}})}{\sin\theta^{\text{inc}}}\right]j_n(kr)P_n^{-m}(\cos\theta), \tag{2.6}$$

$$\Psi_m^{\text{inc}} = \sum_{n=m}^{\infty}\gamma_{mn}\left[-\left(\frac{\partial}{\partial\theta}P_n^m\right)(\cos\theta^{\text{inc}})\right] \times j_n(kr)P_n^{-m}(\cos\theta), \tag{2.7}$$

$$\Phi_m^s = \sum_{n=m}^{\infty}A_{mn}P_n^{-m}(\cos\theta)\begin{cases}j_n(kr)h_n(ka) & (r<a),\\ j_n(ka)h_n(kr) & (r>a),\end{cases} \tag{2.8}$$

$$\Psi_m^s = \sum_{n=m}^{\infty}B_{mn}P_n^{-m}(\cos\theta) \times\begin{cases}j_n(kr)[kah_n(ka)]' & (r<a),\\ [kaj_n(ka)]'h_n(kr) & (r>a),\end{cases} \tag{2.9}$$

where

$$\gamma_{mn} = (-1)^{m+1}i^n[(2n+1)/n(n+1)]\epsilon_m(1-\delta_{0n}). \tag{2.10}$$

The terms j_n and h_n are, respectively, the spherical Bessel and Hankel (of the first kind) functions of order n. The associated Legendre polynomials of degree n, order $\pm m$, are denoted by $P_n^{\pm m}$. The prime in an expression $[xf_n(x)]'$ denotes the derivative with respect to x. The term $\epsilon_m=2$ for $m\neq0$ and $\epsilon_m=1$ for $m=0$. Kronecker's delta $\delta_{ij}=0$ for $i\neq j$ and $\delta_{jj}=1$ for any j. Because the term corresponding to both $m=0$ and $n=0$ is identically zero in the incident field, we set the corresponding scattered potential coefficients identically to zero: $A_{00}\equiv B_{00}\equiv0$.

These representations of the interior and exterior scattered potentials have been chosen so that Φ^s and $\partial_r(r\Psi^s)$ are continuous at $r=a$, thereby ensuring the continuity of the tangential scattered electric field components E_θ and E_ϕ across that surface. The resultant fields satisfy Sommerfeld's radiation condition; the dependence of the scattered fields on $h_n(kr)$ for $r>a$ ensures their decay to zero as $r\to\infty$. The modal coefficients A_{mn} and B_{mn} are the quantities that must be determined.

B. Electromagnetic boundary conditions

The electromagnetic boundary conditions: $E_{\tan}=0$ on the metal and $H_{\tan}$ continuous in the aperture, are now enforced. Referring to Table I, $E_\theta=0$ on the metal if

$$\sum_{m=0}^{\infty}\cos m\phi\left\{\frac{m}{\sin\theta}(\Phi_m^{\text{inc}}+\Phi_m^s) - \frac{Y_0}{i\omega\epsilon}\frac{1}{r}\frac{\partial^2}{\partial r\,\partial\theta}[r(\Psi_m^{\text{inc}}+\Psi_m^s)]\right\}_{r=a}=0,$$

$E_\phi=0$ on the metal if

$$\sum_{m=0}^{\infty}\sin m\phi\left\{-\frac{\partial}{\partial\theta}(\Phi_m^{\text{inc}}+\Phi_m^s) + \frac{Y_0}{i\omega\epsilon}\frac{m}{r\sin\theta}\frac{\partial}{\partial r}[r(\Psi_m^{\text{inc}}+\Psi_m^s)]\right\}_{r=a}=0,$$

H_θ is continuous across the aperture if for $\epsilon\to0$

$$\sum_{m=0}^{\infty}\sin m\phi\left\{\frac{-mY_0}{\sin\theta}(\Psi_m^{\text{inc}}+\Psi_m^s) + \frac{1}{i\omega\mu}\frac{1}{r}\frac{\partial^2}{\partial r\,\partial\theta}[r(\Phi_m^{\text{inc}}+\Phi_m^s)]\right\}_{r=a-\epsilon}^{r=a+\epsilon}=0,$$

and H_ϕ is continuous across the aperture if for $\epsilon\to0$

$$\sum_{m=0}^{\infty}\cos m\phi\left\{-Y_0\frac{\partial}{\partial\theta}(\Psi_m^{\text{inc}}+\Psi_m^s) + \frac{1}{i\omega\mu}\frac{m}{r\sin\theta}\frac{\partial}{\partial r}[r(\Phi_m^{\text{inc}}+\Phi_m^s)]\right\}_{r=a-\epsilon}^{r=a+\epsilon}=0.$$

Because the azimuthal eigenfunctions $\sin m\phi$ and $\cos m\phi$ form an orthogonal set over $[0,2\pi]$, these conditions must be satisfied on a mode by mode basis. They require satisfac-

tion of the following coupled set of dual series equations for the modal coefficients A_{mn} and B_{mn}:

$$ika \sum_{n=m}^{\infty} \{A_{mn} j_n(ka) h_n(ka) - f_{mn}\} m P_n^{-m}(\cos\theta)$$
$$= \sin\theta\, \partial_\theta \sum_{n=m}^{\infty} \{B_{mn} [kaj_n(ka)]' [kah_n(ka)]' - g_{mn}\} P_n^{-m}(\cos\theta) \quad (0 \leqslant \theta < \theta_0), \tag{2.11a}$$

$$\sum_{n=m}^{\infty} A_{mn} m P_n^{-m}(\cos\theta) = -ika \sin\theta\, \partial_\theta \sum_{n=m}^{\infty} B_{mn} P_n^{-m}(\cos\theta) \quad (\theta_0 < \theta \leqslant \pi), \tag{2.11b}$$

$$ika \sin\theta\, \partial_\theta \sum_{n=m}^{\infty} \{A_{mn} j_n(ka) h_n(ka) - f_{mn}\} P_n^{-m}(\cos\theta)$$
$$= \sum_{n=m}^{\infty} \{B_{mn} [kaj_n(ka)]' [kah_n(ka)]' - g_{mn}\} m P_n^{-m}(\cos\theta) \quad (0 \leqslant \theta < \theta_0), \tag{2.12a}$$

$$\sin\theta\, \partial_\theta \sum_{n=m}^{\infty} A_{mn} P_n^{-m}(\cos\theta) = -ika \sum_{n=m}^{\infty} B_{mn} m P_n^{-m}(\cos\theta) \quad (\theta_0 < \theta \leqslant \pi), \tag{2.12b}$$

where

$$f_{mn} = \gamma_{mn} [m P_n^m(\cos\theta^{\text{inc}})/\sin\theta^{\text{inc}}]\, j_n(ka), \tag{2.13}$$

$$g_{mn} = \gamma_{mn} \left[\left(\frac{\partial}{\partial\theta} P_n^m\right)(\cos\theta^{\text{inc}})\right] [kaj_n(ka)]'. \tag{2.14}$$

Equations (2.11) result from the E_θ and H_ϕ boundary conditions and are naturally paired because their θ and ϕ dependencies are the same; Eqs. (2.12) result from the E_ϕ and H_θ boundary conditions. The absence of any spherical Bessel or Hankel function in (2.11b) and (2.12b) results from application of the modified Wronskian relation

$$j_n(x)[xh_n(x)]' - h_n(x)[xj_n(x)]' = i/x. \tag{2.15}$$

Note that $\partial_\theta \equiv \partial/\partial\theta$.

III. PSEUDODECOUPLING ANSATZ

In the problem of plane wave scattering from a solid sphere it is known[27] that the TE and TM portions of the problem may be decoupled. Satisfaction of independent boundary conditions applied directly to the TE and TM Debye potentials leads to series defined over the entire θ interval, $[0,\pi]$; and orthogonality arguments then produce a complete decoupling. Introducing the hole results in mixed boundary conditions over partial θ intervals and a coupling of the TE and TM modes. Nonetheless, one might anticipate some form of TE/TM decoupling even in this case if the proper set of basis functions were employed.

Consider the associated Legendre functions of negative order $P_n^{-m}(\cos\theta)$. For all n and m they are known independent solutions to Legendre's equation[32-35]:

$$\mathcal{L}_\theta P_n^{-m}(\cos\theta) = -n(n+1)(\sin^2\theta) P_n^{-m}(\cos\theta), \tag{3.1}$$

where the operator

$$\mathcal{L}_\theta \equiv (\sin\theta\, \partial_\theta)(\sin\theta\, \partial_\theta) - m^2. \tag{3.2}$$

They have the integral representations[32]

$$P_n^{-m}(\cos\theta) = (-1)^m \left(\frac{2}{\pi}\right)^{1/2} \frac{\csc^m\theta}{\Gamma(m+\frac{1}{2})} \times \int_0^\theta \frac{\cos[(n+\frac{1}{2})t]\,dt}{[\cos t - \cos\theta]^{1/2-m}}. \tag{3.3}$$

Note that our definition of P_n^{-m} differs from that in Ref. 32 by the factor $(-1)^m$. The related functions

$$\overline{P}_n^{-m}(\cos\theta) \equiv (-1)^{n+m} P_n^{-m}(\cos(\pi-\theta)) = (-1)^m \left(\frac{2}{\pi}\right)^{1/2} \frac{\csc^m\theta}{\Gamma(m+\frac{1}{2})} \times \int_\theta^\pi \frac{\sin[(n+\frac{1}{2})t]\,dt}{[\cos\theta - \cos t]^{1/2-m}} \tag{3.4}$$

also satisfy (3.1). For $n \geqslant m$, these functions are identical by the standard symmetry relation

$$\overline{P}_n^{-m}(\cos\theta) \equiv (-1)^{n+m} P_n^{-m}(\cos(\pi-\theta)) \equiv P_n^{-m}(\cos\theta). \tag{3.5}$$

However, for $0 \leqslant n < m$, this relation no longer holds true. In particular, $P_n^{-m}(\cos\theta)$ is finite at $\theta = 0$ but infinite at $\theta = \pi$. On the other hand, $\overline{P}_n^{-m}(\cos\theta)$ is finite at $\theta = \pi$ but infinite at $\theta = 0$. This behavior is immediately apparent for $n = 0$ where

$$P_0^{-m}(\cos\theta) = [(-1)^m/m!]\tan^m(\theta/2), \tag{3.6a}$$

$$\overline{P}_0^{-m}(\cos\theta) = (1/m!)\cot^m(\theta/2). \tag{3.6b}$$

Return now to the dual series systems (2.11) and (2.12). They are self-consistent if

$$\mathcal{L}_\theta \sum_{n=m}^{\infty} \{A_{mn} j_n(ka) h_n(ka) - f_{mn}\} P_n^{-m}(\cos\theta) = 0$$
$$\mathcal{L}_\theta \sum_{n=m}^{\infty} \{B_{mn} [kaj_n(ka)]' [kah_n(ka)]' - g_{mn}\} P_n^{-m}(\cos\theta) = 0 \quad (0 \leqslant \theta < \theta_0);$$

$$\mathscr{L}_\theta \sum_{n=m}^{\infty} A_{mn} P_n^{-m}(\cos\theta) = 0$$
$$\mathscr{L}_\theta \sum_{n=m}^{\infty} B_{mn} P_n^{-m}(\cos\theta) = 0 \qquad (\theta_0 < \theta \leqslant \pi).$$

Bounded homogeneous solutions of these equations are admissible and are proportional to $P_0^{-m}(\cos\theta)$ and $\overline{P}_0^{-m}(\cos\theta)$ over their respective intervals. Solutions to the TE dual series system

$$\sum_{n=m}^{\infty} \{A_{mn} j_n(ka) h_n(ka) - f_{mn}\} P_n^{-m}(\cos\theta) = \alpha_m P_0^{-m}(\cos\theta) \quad (0 \leqslant \theta < \theta_0), \tag{3.7a}$$

$$\sum_{n=m}^{\infty} A_{mn} P_n^{-m}(\cos\theta) = \overline{\alpha}_m \overline{P}_0^{-m}(\cos\theta) \quad (\theta_0 < \theta \leqslant \pi), \tag{3.7b}$$

for $m \geqslant 1$ and to the TM dual series system

$$\sum_{n=m}^{\infty} \{B_{mn} [kaj_n(ka)]' [kah_n(ka)]' - g_{mn}\} P_n^{-m}(\cos\theta) = \beta_m P_0^{-m}(\cos\theta) \quad (0 \leqslant \theta < \theta_0), \tag{3.8a}$$

$$\sum_{n=m}^{\infty} B_{mn} P_n^{-m}(\cos\theta) = \overline{\beta}_m \overline{P}_0^{-m}(\cos\theta) \quad (\theta_0 < \theta \leqslant \pi), \tag{3.8b}$$

for $m \geqslant 0$ are therefore solutions to (2.11) and (2.12) provided that the "decoupling" constants α_m, β_m, $\overline{\alpha}_m$, and $\overline{\beta}_m$ are constrained by those coupled dual series equations. Since

$$\sin\theta\, \partial_\theta \begin{bmatrix} P_0^{-m}(\cos\theta) \\ \overline{P}_0^{-m}(\cos\theta) \end{bmatrix} = m \begin{bmatrix} +P_0^{-m}(\cos\theta) \\ -\overline{P}_0^{-m}(\cos\theta) \end{bmatrix}, \tag{3.9}$$

the required constraint relations for $m \geqslant 1$ are simply

$$\beta_m = ika\alpha_m, \tag{3.10}$$
$$\overline{\alpha}_m = ika\overline{\beta}_m. \tag{3.11}$$

There is no $m = 0$ constraint relation because there is no $m = 0$ TE dual series equation.

Consequently, although the TE and TM portions of the dual series systems (2.11) and (2.12) have been decoupled, the TE and TM modal coefficients are still coupled through these "decoupling" constant constraint relations. This explains the connotation "pseudodecoupling ansatz." The solutions to the TE and TM dual series systems (3.7) and (3.8) subject to the constraints (3.10) and (3.11) comprise the desired result.

IV. TE AND TM DUAL SERIES SOLUTIONS

The TE and TM dual series systems can be reduced to more manageable and physically revealing forms with several manipulations. First, by introducing for $n \geqslant 1$ the functions χ_n^ϕ and χ_n^ψ so that

$$j_n(ka) h_n(ka) = (1 + \chi_n^\phi)/ika(2n+1), \tag{4.1}$$
$$[kaj_n(ka)]' [kah_n(ka)]' = -[n(n+1)/ika(2n+1)](1 + \chi_n^\psi), \tag{4.2}$$

the dual series systems (3.7) and (3.8) can be rewritten as

$$\sum_{n=m}^{\infty} \frac{A_{mn}}{n+\frac{1}{2}} (1 + \chi_n^\phi) P_n^{-m} = 2ika\alpha_m P_0^{-m} + 2ika \sum_{n=m}^{\infty} f_{mn} P_n^{-m} \quad (0 \leqslant \theta < \theta_0),$$

$$\sum_{n=m}^{\infty} A_{mn} P_n^{-m} = \overline{\alpha}_m \overline{P}_0^{-m} \quad (\theta_0 < \theta \leqslant \pi);$$

$$\sum_{n=m}^{\infty} B_{mn} \frac{n(n+1)}{n+\frac{1}{2}} (1 + \chi_n^\psi) P_n^{-m} = -2ika\beta_m P_0^{-m} - 2ika \times \sum_{n=m}^{\infty} g_{mn} P_n^{-m} \quad (0 \leqslant \theta < \theta_0),$$

$$\sum_{n=m}^{\infty} B_{mn} P_n^{-m} = \overline{\beta}_m \overline{P}_0^{-m} \quad (\theta_0 < \theta \leqslant \pi).$$

As shown in the Appendix, in the quasistatic limit

$$\lim_{ka \to 0} \chi_n^\phi = 0 + \mathscr{O}((ka)^2), \tag{4.3a}$$

$$\lim_{ka \to 0} \chi_n^\psi = 0 + \mathscr{O}((ka)^2). \tag{4.3b}$$

Thus, in analogy with the dual series treatments of the two-dimensional slit cylinder coupling problems given in Refs. 20–23, the static terms have been extracted. These TE and TM dual series must now be solved subject to Meixner's edge conditions; i.e., one must account for the singular behavior near the rim of the aperture required by the finite energy condition. The large n behavior of the solution coefficients is responsible for this edge behavior. Since, as shown in the Appendix, for large values of this index

$$\lim_{n \to \infty} \chi_n^\phi \sim \mathscr{O}(n^{-2}), \tag{4.4a}$$

$$\lim_{n \to \infty} \chi_n^\psi \sim \mathscr{O}(n^{-2}), \tag{4.4b}$$

the terms proportional to χ_n^ϕ and χ_n^ψ are of order n^{-2} smaller than the static pieces. To enhance the isolation of the large index behavior in the TM systems, we introduce the additional functions

$$\tilde{\chi}_n^\psi = n(n+1)(1 + \chi_n^\psi)/(n+\tfrac{1}{2})^2 - 1 = -\{1 + [4ika/(2n+1)][kaj_n(ka)]' \times [kah_n(ka)]'\}, \tag{4.5}$$

which exhibit the limiting behaviors

$$\lim_{n \to \infty} \tilde{\chi}_n^\psi = \mathscr{O}(n^{-2}) \text{ and } \lim_{ka \to 0} \tilde{\chi}_n^\psi = -(2n+1)^{-2}, \tag{4.6}$$

and rewrite the TM dual series systems as

$$\sum_{n=m}^{\infty} B_{mn} \left(n + \frac{1}{2}\right) (1 + \tilde{\chi}_n^\psi) P_n^{-m} = -2ika\beta_m P_0^{-m} - 2ika \times \sum_{n=m}^{\infty} g_{mn} P_n^{-m} \quad (0 \leqslant \theta < \theta_0), \tag{4.7}$$

$$\sum_{n=m}^{\infty} B_{mn} P_n^{-m} = \overline{\beta}_m \overline{P}_0^{-m} \quad (\theta_0 < \theta \leqslant \pi).$$

We then treat the terms proportional to χ_n^ϕ in (4.2) and $\tilde{\chi}_n^\psi$ in (4.7) as forcing terms by moving them to the right-hand sides. This isolates the pieces responsible for the singularities near the rim of the aperture on the left-hand sides. Defining the forcing terms

$$F_{mn} = 2ikaf_{mn} - [A_{mn}/(n+\tfrac{1}{2})]\chi_n^\phi, \tag{4.8a}$$

$$G_{mn} = -2ikag_{mn} - \tilde{\chi}_n^\psi (n+\tfrac{1}{2})B_{mn}, \tag{4.8b}$$

the TE and TM dual series systems for $m \geqslant 1$ become

$$\sum_{n=m}^{\infty} \frac{A_{mn}}{n+\frac{1}{2}} P_n^{-m} = 2ika\alpha_m P_0^{-m} + \sum_{n=m}^{\infty} F_{mn} P_n^{-m} \quad (0 \leqslant \theta < \theta_0), \tag{4.9a}$$

$$\sum_{n=m}^{\infty} A_{mn} \overline{P}_n^{-m} = \bar{\alpha}_m \overline{P}_0^{-m} \quad (\theta_0 < \theta \leqslant \pi); \tag{4.9b}$$

$$\sum_{n=m}^{\infty} B_{mn}\left(n+\frac{1}{2}\right) P_n^{-m} = -2ika\beta_m P_0^{-m} + \sum_{n=m}^{\infty} G_{mn} P_n^{-m} \quad (0 \leqslant \theta < \theta_0), \tag{4.10a}$$

$$\sum_{n=m}^{\infty} B_{mn} \overline{P}_n^{-m} = \bar{\beta}_m \overline{P}_0^{-m} \quad (\theta_0 < \theta \leqslant \pi). \tag{4.10b}$$

Equation (3.5) has been invoked to convert the P_n^{-m} to their duals $\overline{P}_n^{-m}$ over the aperture interval. This form of the dual series systems strongly suggests the solution process we introduce below.

For $m = 0$, the TM dual series becomes

$$\sum_{n=1}^{\infty} B_{0n}\left(n+\frac{1}{2}\right) P_n = -2ika\beta_0 + \sum_{n=1}^{\infty} G_{0n} P_n \quad (0 \leqslant \theta < \theta_0), \tag{4.10a'}$$

$$\sum_{n=1}^{\infty} B_{0n} P_n = \bar{\beta}_0 \quad (\theta_0 < \theta \leqslant \pi), \tag{4.10b'}$$

since $P_n^0 \equiv P_n$, Legendre's polynomial, and $B_{00} \equiv g_{00} \equiv 0$.

The singular behavior of the fields near the aperture rim ($\theta = \theta_0$) is reflected in the corresponding behavior of the current components

$$J_\theta(\theta,\phi) = H_\phi^<(a,\theta,\phi) - H_\phi^>(a,\theta,\phi) = \left[\frac{-Y_0 E_0}{(ka)^2}\right] \sum_{m=0}^{\infty} \cos m\phi \sum_{n=m}^{\infty} \left\{ A_{mn}\left[\frac{mP_n^{-m}(\cos\theta)}{\sin\theta}\right] + ikaB_{mn}\left[\frac{\partial}{\partial\theta} P_n^{-m}(\cos\theta)\right]\right\}, \tag{4.11a}$$

$$J_\phi(\theta,\phi) = H_\theta^>(a,\theta,\phi) - H_\theta^<(a,\theta,\phi) = \left[\frac{+Y_0 E_0}{(ka)^2}\right] \sum_{m=0}^{\infty} \sin m\phi \sum_{n=m}^{\infty} \left\{ A_{mn}\left[\frac{\partial}{\partial\theta} P_n^{-m}(\cos\theta)\right] + ikaB_{mn}\left[\frac{mP_n^{-m}(\cos\theta)}{\sin\theta}\right]\right\}, \tag{4.11b}$$

where, for instance, $H_\phi^<(a,\theta,\phi)$ $[H_\phi^>(a,\theta,\phi)]$ is the ϕ component of the magnetic field for $r \leqslant a$ $(r \geqslant a)$ evaluated at $r = a$. Because the angular dependence of the terms depending on A_{mn} and B_{mn} in these expressions is distinct, we use it to guide our constructions of the physically correct TE and TM solutions. These solutions require several summation formulas:

$$\sum_{n=0}^{\infty} P_n^{-m}(\cos\theta)\partial_{\theta_0}^j \cos\left(n+\frac{1}{2}\right)\theta_0 = 0 \quad (0 \leqslant \theta < \theta_0), \tag{4.12a}$$

$$\sum_{n=0}^{\infty} \overline{P}_n^{-m}(\cos\theta)\partial_{\theta_0}^j \sin\left(n+\frac{1}{2}\right)\theta_0 = 0 \quad (\theta_0 < \theta \leqslant \pi), \tag{4.12b}$$

$$\sum_{n=0}^{\infty} P_n^{-m}(\cos\theta) \frac{\sin(n+\frac{1}{2})\theta_0}{n+\frac{1}{2}} = \sum_{n=0}^{\infty} \frac{(-1)^n P_n^{-m}(\cos\theta)}{n+\frac{1}{2}} \quad (0 \leqslant \theta < \theta_0), \tag{4.12c}$$

$$\sum_{n=0}^{\infty} \overline{P}_n^{-m}(\cos\theta) \frac{\cos(n+\frac{1}{2})\theta_0}{n+\frac{1}{2}} = \sum_{n=0}^{\infty} \frac{\overline{P}_n^{-m}(\cos\theta)}{n+\frac{1}{2}} \quad (\theta_0 < \theta \leqslant \pi), \tag{4.12d}$$

$$\sum_{n=0}^{\infty} P_n^{-m}(\cos\theta) \frac{\cos(n+\frac{1}{2})\theta_0}{(n+\frac{1}{2})^2} = (\pi - \theta_0) \sum_{n=0}^{\infty} (-1)^n \frac{P_n^{-m}(\cos\theta)}{n+\frac{1}{2}} \quad (0 \leqslant \theta < \theta_0), \tag{4.12e}$$

which are derived from the basic expressions [see Ref. 36, Eq. (3.71)]

$$\sum_{n=0}^{\infty} P_n^{-m}(\cos\theta)\cos\left(n+\frac{1}{2}\right)\psi = \begin{cases} \dfrac{(-1)^m(\pi/2)^{1/2}}{\Gamma(m+\frac{1}{2})} \dfrac{[\cos\psi - \cos\theta]^{m-1/2}}{\sin^m\theta} & (0 < \psi < \theta), \\ 0 & (\theta < \psi < \pi) \end{cases} \tag{4.13a}$$

(modified to our sign convention) and its dual

$$\sum_{n=0}^{\infty} \overline{P}_n^{-m}(\cos\theta)\sin\left(n+\frac{1}{2}\right)\psi = \begin{cases} 0 & (0 < \psi < \theta), \\ \dfrac{(\pi/2)^{1/2}}{\Gamma(m+1/2)} \dfrac{[\cos\theta - \cos\psi]^{m-1/2}}{\sin^m\theta} & (\theta < \psi < \pi), \end{cases} \tag{4.13b}$$

and the identities (see Ref. 36, 1.19 and 1.16)

$$\sum_{n=0}^{\infty}(-1)^n\frac{\cos(n+\frac{1}{2})t}{n+\frac{1}{2}}=\frac{\pi}{2}\quad(0\leqslant t<\pi),\tag{4.14a}$$

$$\sum_{n=0}^{\infty}\frac{\sin(n+\frac{1}{2})t}{n+\frac{1}{2}}=\frac{\pi}{2}\quad(0\leqslant t<\pi).\tag{4.14b}$$

A. TE dual series solution

We would like to reduce the associated Legendre function dual series system (4.9) to one in sines and cosines. This conversion would appear to be straightforward with the representations (3.3) and (3.4) and with an interchange of the summations and integrations. However, consider Meixner's edge conditions (see Ref. 26, Sec. 9.2), which, when applied to the field generated by the TE Debye potential, imply that as the edge is approached along the surface $r=a$,

$$H_\phi^s(a,\theta,\phi)\sim\partial_r(r\Phi_m^s)|_{r=a}\sim(\theta_0-\theta)^{+1/2},$$
$$H_\theta^s(a,\theta,\phi)\sim\partial_r\partial_\theta(r\Phi_m^s)|_{r=a}\sim(\theta_0-\theta)^{-1/2}.$$

The corresponding portions of J_θ and J_ϕ near $\theta=\theta_0$ must behave, respectively, as

$$\sum_{n=m}^{\infty}A_{mn}P_n^{-m}(\cos\theta)\sim(\theta_0-\theta)^{+1/2},\tag{4.15a}$$

$$\sum_{n=m}^{\infty}A_{mn}\,\partial_\theta P_n^{-m}(\cos\theta)\sim(\theta_0-\theta)^{-1/2},\tag{4.15b}$$

where $m\geqslant 1$. Analogously, the θ dependency of Eq. (4.9a) near $\theta=\theta_0$ differs from that of (4.9b) by $(\theta_0-\theta)^{+1}$. The factor $(n+\frac{1}{2})^{-1}$ in (4.9a) is responsible for this difference. Thus, with [see Ref. 37, (8.10.7) and (6.1.37)]

$$\lim_{n\to\infty}P_n^{-m}(\cos\theta)\sim n^{-m+1/2}\times\frac{\cos[(n+\frac{1}{2})\theta-m\pi/2-\pi/4]}{(\pi\sin\theta/2)^{1/2}}$$

and (4.13), Meixner's conditions are satisfied if

$$\lim_{n\to\infty}A_{mn}\sim n^{m-1}.\tag{4.16}$$

The simple summation–integration interchange is then not directly permitted because it will introduce terms that are proportional to delta functions and their derivatives; i.e., with (1.135) from Ref. 36 one finds, for instance, that near $\theta=\theta_0$

$$\sum_{n=0}^{\infty}n^j\cos\left(n+\frac{1}{2}\right)\theta\cos\left(n+\frac{1}{2}\right)\theta_0\sim\delta^{(j)}(\theta_0-\theta),$$

the jth derivative of the Dirac distribution. Interchange can be accomplished by preconditioning the dual series as follows.

We need to introduce terms into (4.9) that will cancel the potential delta function contributions. This is accomplished with Eqs. (4.12)–(4.14). In particular, we define the modified solution coefficients

$$\tilde{A}_{mn}=\sum_{j=0}^{m-1}a_{mj}\,\partial_{\theta_0}^j\sin\left(n+\frac{1}{2}\right)\theta_0+\begin{cases}A_{mn} & (n\geqslant m\geqslant 1),\\ 0 & (0\leqslant n<m),\end{cases}\tag{4.17}$$

and the modified forcing term coefficients

$$\tilde{F}_{mn}=F_{mn}\quad(n\geqslant m=1),\tag{4.18a}$$

$$\tilde{F}_{mn}=\frac{-(ka)^2}{2}\sum_{j=0}^{m-2}a_{m(j+1)}\,\partial_{\theta_0}^j\frac{\cos(n+\frac{1}{2})\theta_0}{(n+\frac{1}{2})^2}+\begin{cases}F_{mn} & (n\geqslant m\geqslant 2),\\ 0 & (0\leqslant n<m),\end{cases}\tag{4.18b}$$

so that the TE dual series systems for $m\geqslant 1$ become

$$\sum_{n=0}^{\infty}\frac{\tilde{A}_{mn}}{n+\frac{1}{2}}P_n^{-m}=2ika\alpha_mP_0^{-m}+(a_{m0}-\kappa_m^E)\sum_{n=0}^{\infty}\frac{(-1)^nP_n^{-m}}{n+\frac{1}{2}}+\sum_{n=0}^{\infty}\tilde{F}_{mn}P_n^{-m}\quad(0\leqslant\theta<\theta_0),\tag{4.19a}$$

$$\sum_{n=0}^{\infty}\tilde{A}_{mn}\overline{P}_n^{-m}=\bar{\alpha}_m\overline{P}_0^{-m}\quad(\theta_0<\theta\leqslant\pi).\tag{4.19b}$$

The constants

$$\kappa_m^E=\frac{-(ka)^2}{2}\begin{cases}0, & \text{for } m=1,\\ (\pi-\theta_0)a_{m1}, & \text{for } m=2,\\ (\pi-\theta_0)a_{m1}-a_{m2}, & \text{for } m\geqslant 3.\end{cases}\tag{4.20}$$

The additional unknown coefficients a_{mj} $(j=0,1,\ldots,m-1)$ provide the extra degrees of freedom needed to remove the unphysical singularities and permit the desired summation–integration interchange. In particular, their values will be fixed by our solution process so that for any $m\geqslant 1$

$$\lim_{n\to\infty}\tilde{A}_{mn}\sim\mathcal{O}(n^{-1}),\tag{4.17'}$$

$$\lim_{n\to\infty}\tilde{F}_{mn}\sim\mathcal{O}(n^{-3}).\tag{4.18'}$$

Note that it can be inferred from the form of (4.19) that we have completed the basis function set for this open geometry by including the associated Legendre polynomials P_n^{-m} and $\overline{P}_n^{-m}$ for $0\leqslant n<m$.

Inserting (3.3) and (3.4) into (4.19) and interchanging the summations and integrations, the desired TE dual series are generated:

$$\sum_{n=0}^{\infty}\frac{\tilde{A}_{mn}}{n+\frac{1}{2}}\cos\left(n+\frac{1}{2}\right)t=\frac{\pi}{2}(a_{m0}-\kappa_m^E)+2ika\alpha_m\cos\frac{t}{2}+\sum_{n=0}^{\infty}\tilde{F}_{mn}\cos\left(n+\frac{1}{2}\right)t\quad(0\leqslant t<\theta_0),\tag{4.21a}$$

$$\sum_{n=0}^{\infty}\tilde{A}_{mn}\sin\left(n+\frac{1}{2}\right)t=\bar{\alpha}_m\sin\frac{t}{2}\quad(\theta_0<t\leqslant\pi).\tag{4.21b}$$

A solution of (4.21) is constructed as in Refs. 20–24 by first making the metal and aperture equations display the same t dependence. Two possibilities exist: integrating (4.21b) or differentiating (4.21a). Only the former guarantees satisfaction of (4.17'). Applying $\int_t^\pi dt$ to (4.21b) leads to the dual series system

$$\sum_{n=0}^{\infty} \frac{\tilde{A}_{mn}}{n+\frac{1}{2}} \cos\left(n+\frac{1}{2}\right) t = \begin{cases} \frac{\pi}{2}(a_{m0}-\kappa_m^E) + 2ika\alpha_m \cos\frac{t}{2} + \sum_{n=0}^{\infty} \tilde{F}_{mn} \cos\left(n+\frac{1}{2}\right) t & (0 \leqslant t < \theta_0), \\ 2\bar{\alpha}_m \cos\frac{t}{2} & (\theta_0 < t \leqslant \pi). \end{cases} \tag{4.22}$$

Since the left-hand side of (4.22) is now defined over the entire $[0,\pi]$ interval, Fourier inversion then yields the coefficients

$$\frac{\tilde{A}_{ml}}{l+\frac{1}{2}} = (a_{m0}-\kappa_m^E)\frac{\sin(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}} + (2ika\alpha_m - 2\bar{\alpha}_m)\Lambda_{0l}^E + 2\bar{\alpha}_m\delta_{0l} + \sum_{n=0}^{\infty} \tilde{F}_{mn}\Lambda_{nl}^E \quad (l=0,1,\ldots), \tag{4.23}$$

where the inversion terms

$$\Lambda_{nl}^E = \frac{2}{\pi}\int_0^{\theta_0} \cos\left[\left(n+\frac{1}{2}\right)\psi\right]\cos\left[\left(l+\frac{1}{2}\right)\psi\right] d\psi = \begin{cases} \frac{1}{\pi}\left[\frac{\sin(n-l)\theta_0}{n-l} + \frac{\sin(n+l+1)\theta_0}{n+l+1}\right] & (n \neq l), \\ \frac{1}{\pi}\left[\theta_0 + \frac{\sin(2l+1)\theta_0}{2l+1}\right] & (n=l). \end{cases} \tag{4.24}$$

Explicitly, (4.23) means

$$\sum_{j=1}^{m-1} a_{mj}\,\partial_{\theta_0}^j \frac{\sin(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}} = (2ika\alpha_m - 2\bar{\alpha}_m)\Lambda_{0l}^E + 2\bar{\alpha}_m\delta_{0l} - \kappa_m^E \frac{\sin(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}} + \sum_{n=0}^{\infty} \tilde{F}_{mn}\Lambda_{nl}^E \quad (l=0,1,\ldots,m-1), \tag{4.23'}$$

$$\frac{A_{ml}}{l+\frac{1}{2}} = -\sum_{j=1}^{m-1} a_{mj}\,\partial_{\theta_0}^j \frac{\sin(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}} + (2ika\alpha_m - 2\bar{\alpha}_m)\Lambda_{0l}^E + 2\bar{\alpha}_m\delta_{0l} - \kappa_m^E \frac{\sin(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}} + \sum_{n=0}^{\infty} \tilde{F}_{mn}\Lambda_{nl}^E \quad (l=m,m+1,\ldots). \tag{4.23''}$$

Furthermore, inversion requires continuity of the right-hand side of (4.22) across $t=\theta_0$. This yields

$$a_{m0} = \kappa_m^E - \frac{2}{\pi}\left[(2ika\alpha_m - 2\bar{\alpha}_m)\cos\frac{\theta_0}{2} + \sum_{n=0}^{\infty} \tilde{F}_{mn}\cos\left(n+\frac{1}{2}\right)\theta_0\right]. \tag{4.25}$$

The system (4.23) and (4.25) is an infinite system of linear equations for the TE solution coefficients. The first $(m+1)$ of these, (4.23') and Eq. (4.25), would not have appeared without the introduction of the terms a_{mj}, α_n, and $\bar{\alpha}_n$; hence, the terms $\cos(n+\frac{1}{2})\theta$ and $\sin(n+\frac{1}{2})\theta$ for $n=0,1,\ldots,m-1$ into (4.21). The Λ_{0l}^E $(l=0,1,\ldots)$ terms originate in this completion of the expansion. Moreover, since there are no solution coefficients A_{mn} $(0 \leqslant n < m)$ in those first $(m+1)$ equations, we may view them as orthogonality relations. They determine the interchange coefficients a_{mj} $(j=1,\ldots,m-1)$ and a relation between the decoupling coefficients α_m and $\bar{\alpha}_m$. In a similar fashion, the TM case generates a relation between β_m and $\bar{\beta}_m$. The remaining two degrees of freedom are determined by the constraint relations (3.10) and (3.11). The coefficient a_{m0} is defined by (4.25) and is coupled to all of the other a_{mj} $(j=1,\ldots,m-1)$ through the relation for $(2ika\alpha_m - 2\bar{\alpha}_m)$. However, it does not contribute directly to the solution coefficients A_{ml} $(l=m,m+1,\ldots)$. It only provides that degree of freedom needed to insure continuity across the boundary between the metal and aperture intervals.

Equations (4.23) are solutions of the original dual series (4.9) subject to Meixner's edge conditions (4.15a). They are general solutions if these results are independent of the decoupling and interchange constants. This has been confirmed numerically for $m=1,2,3$. A rich set of new associated Legendre polynomial identities is obtained from this validation process.[25] We consider explicitly only the $m=1$ relations since they are employed for the normal incidence case discussed below.

For $m=1$, the solution system

$$\frac{A_{1l}}{l+\frac{1}{2}} = (2ika\alpha_1 - 2\bar{\alpha}_1)\Lambda_{0l}^E + \sum_{n=1}^{\infty} F_{1n}\Lambda_{nl}^E \quad (l=1,2,\ldots), \tag{4.26a}$$

$$0 = (2ika\alpha_1 - 2\bar{\alpha}_1)\Lambda_{00}^E + 2\bar{\alpha}_1 + \sum_{n=1}^{\infty} F_{1n}\Lambda_{n0}^E, \tag{4.26b}$$

gives the coefficients

$$\frac{A_{1l}}{l+\frac{1}{2}} = \sum_{n=1}^{\infty} F_{1n}\Gamma_{1,nl}^E + 2\bar{\alpha}_1 L_{1l}^E \quad (l=1,2,\ldots), \tag{4.27a}$$

where

$$\Gamma_{1,nl}^E = \Lambda_{nl}^E - \Lambda_{n0}^E\Lambda_{0l}^E/\Lambda_{00}^E, \tag{4.27b}$$

$$L_{1l}^E = -\Lambda_{0l}^E/\Lambda_{00}^E. \tag{4.27c}$$

Substituting these expressions into the $m=1$ versions of (4.9a) and (4.9b), the original dual series system is satisfied since *on the metal* $(0 \leqslant \theta < \theta_0)$

$$\sum_{l=1}^{\infty} \Gamma_{1,nl}^E P_l^{-1} - P_n^{-1} - L_{1n}^E P_0^{-1} = 0 \quad (n=1,2,\ldots), \tag{4.28a}$$

$$\sum_{l=1}^{\infty} L_{1l}^E P_l^{-1} - \frac{\Lambda_{00}^E - 1}{\Lambda_{00}^E} P_0^{-1} = 0, \tag{4.28b}$$

and *in the aperture* $(\theta_0 < \theta \leqslant \pi)$

$$\sum_{l=1}^{\infty} (2l+1)\Gamma_{nl}^E P_l^{-1} = 0 \quad (n=1,2,\ldots), \tag{4.29a}$$

$$\sum_{l=1}^{\infty} (2l+1) L_{1l}^E - \overline{P}_0^{-1} = 0. \tag{4.29b}$$

When evaluated over the metal interval $(0 \leqslant \theta < \theta_0)$, the left-hand sides of Eqs. (4.29) yield Eqs. (4.15).

B. TM dual series solution

We proceed as in the TE case. Consider the dual series systems (4.10) and (4.10'). Meixner's edge conditions applied to the fields generated by the TM Debye potential imply that near $\theta = \theta_0$

$$H_\theta^s(a,\theta,\phi) \sim (r\Psi_m^s)|_{r=a} \sim (\theta_0 - \theta)^{+3/2},$$
$$H_\phi^s(a,\theta,\phi) \sim \partial_\theta (r\Psi_m^s)|_{r=a} \sim (\theta_0 - \theta)^{+1/2}.$$

Thus, from (4.11a) and (4.11b), the portions of J_θ and J_ϕ generated by Ψ_m^s near the aperture edge behave, respectively, as

$$\sum_{n=m}^{\infty} B_{mn} P_n^{-m}(\cos\theta) \sim (\theta_0 - \theta)^{+3/2}, \tag{4.30a}$$

$$\sum_{n=m}^{\infty} B_{mn}\, \partial_\theta P_n^{-m}(\cos\theta) \sim (\theta_0 - \theta)^{+1/2}, \tag{4.30b}$$

where $m \geqslant 0$. Analogously, the θ dependency of the metal equations near $\theta = \theta_0$ differs from that of the aperture equations by $(\theta_0 - \theta)^{-1}$. The factor $(n + \frac{1}{2})$ in the metal equations is responsible for this difference.

The requisite edge behavior (4.30a) is obtained if

$$\lim_{n\to\infty} B_{mn} \sim n^{m-2}. \tag{4.31}$$

Consider first the cases with $m > 0$. Anticipating the effects of the operator interchange, we introduce the modified coefficients

$$\tilde{B}_{mn} = \sum_{j=0}^{m-1} b_{mj}\, \partial_{\theta_0}^j \frac{\cos(n+\frac{1}{2})\theta_0}{n+\frac{1}{2}} + \begin{cases} B_{mn} & (n \geqslant m \geqslant 1), \\ 0 & (0 \leqslant n < m), \end{cases} \tag{4.32}$$

the modified forcing term coefficients

$$\tilde{G}_{mn} = G_{mn} \quad (n \geqslant m = 1), \tag{4.33a}$$

$$\tilde{G}_{mn} = \frac{(ka)^2}{2} \sum_{j=0}^{m-2} b_{m(j+1)}\, \partial_{\theta_0}^j \frac{\sin(n+\frac{1}{2})\theta_0}{n+\frac{1}{2}} + \begin{cases} G_{mn} & (n \geqslant m \geqslant 2), \\ 0 & (0 \leqslant n < m), \end{cases} \tag{4.33b}$$

and the constants

$$\kappa_m^H = \begin{cases} 0, & \text{for } m = 1, \\ [(ka)^2/2] b_{m1}, & \text{for } m \geqslant 2. \end{cases} \tag{4.33c}$$

The interchange constants b_{mj} $(j = 0,1,\ldots,m-1)$ will be adjusted so that for all $m \geqslant 1$

$$\lim_{n\to\infty} \tilde{B}_{mn} \sim \mathcal{O}(n^{-2}), \tag{4.32'}$$

$$\lim_{n\to\infty} \tilde{G}_{mn} \sim \mathcal{O}(n^{-3}). \tag{4.33'}$$

The dual series systems (4.10) become

$$\sum_{n=0}^{\infty} \left(n+\frac{1}{2}\right) \tilde{B}_{mn} P_n^{-m} = -2ika\beta_m P_0^{-m} + \sum_{n=0}^{\infty} \tilde{G}_{mn} P_n^{-m} - \kappa_m^H \times \sum_{n=0}^{\infty} \frac{(-1)^n P_n^{-m}}{n+\frac{1}{2}} \quad (0 \leqslant \theta < \theta_0), \tag{4.34a}$$

$$\sum_{n=0}^{\infty} \tilde{B}_{mn} \bar{P}_n^{-m} = \bar{\beta}_m \bar{P}_0^{-m} + b_{m0} \sum_{n=0}^{\infty} \frac{\bar{P}_n^{-m}}{n+\frac{1}{2}} \quad (\theta_0 < \theta \leqslant \pi). \tag{4.34b}$$

Introducing (3.3) and (3.4), interchanging summations and integrations, and applying the operator $\int_0^t dt$ to the resulting metal equation to attain similar t dependencies for both equations, the TM dual series prior to inversion are

$$\sum_{n=0}^{\infty} \tilde{B}_{mn} \sin\left(n+\frac{1}{2}\right) t = \begin{cases} -4ika\beta_m \sin\frac{t}{2} + \sum_{n=0}^{\infty} \frac{\tilde{G}_{mn}}{n+\frac{1}{2}} \sin\left(n+\frac{1}{2}\right) t - \kappa_m^H \frac{\pi}{2} t & (0 \leqslant t < \theta_0), \\ \bar{\beta}_m \sin\frac{t}{2} + b_{m0}\frac{\pi}{2} & (\theta_0 < t \leqslant \pi). \end{cases} \tag{4.35}$$

Introducing the terms

$$\mu_l(\theta_0) = \int_0^{\theta_0} t \sin\left(l+\frac{1}{2}\right) t = \frac{\sin(l+\frac{1}{2})\theta_0}{(l+\frac{1}{2})^2} - \theta_0 \frac{\cos(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}}, \tag{4.36}$$

$$\Lambda_{nl}^H = \frac{2}{\pi} \int_0^{\theta_0} \sin\left[\left(n+\frac{1}{2}\right)\psi\right] \sin\left[\left(l+\frac{1}{2}\right)\psi\right] d\psi = \begin{cases} \frac{1}{\pi}\left[\frac{\sin(n-l)\theta_0}{n-l} - \frac{\sin(n+l+1)\theta_0}{n+l+1}\right] & (n \neq l), \\ \frac{1}{\pi}\left[\theta_0 - \frac{\sin(2l+1)\theta_0}{2l+1}\right] & (n = l), \end{cases} \tag{4.37}$$

Fourier inversion leads to the coefficient expressions for $m \geqslant 1$

$$\tilde{B}_{ml} = -(4ika\beta_m + \bar{\beta}_m)\Lambda_{0l}^H + \bar{\beta}_m \delta_{0l} + b_{m0} \frac{\cos(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}} - \kappa_m^H \mu_l(\theta_0) + \sum_{n=0}^{\infty} \frac{\tilde{G}_{mn}}{n+\frac{1}{2}} \Lambda_{nl}^H \quad (l = 0,1,\ldots). \tag{4.38a}$$

These contain explicitly m orthogonality relations

$$\sum_{j=0}^{m-1} b_{mj}\, \partial_{\theta_0}^j \frac{\cos(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}} = -(4ika\beta_m + \bar{\beta}_m)\Lambda_{0l}^H + \bar{\beta}_m \delta_{0l} + b_{m0} \frac{\cos(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}} - \kappa_m^H \mu_l(\theta_0) + \sum_{n=0}^{\infty} \frac{\tilde{G}_{mn}}{n+\frac{1}{2}} \Lambda_{nl}^H \quad (l = 0,\ldots,m-1), \tag{4.38b}$$

and the coefficients

$$B_{ml} = -(4ika\beta_m + \bar{\beta}_m)\Lambda^H_{0l} + \bar{\beta}_m\delta_{0l} - \kappa^H_m\mu_l(\theta_0) - \sum_{j=1}^{m-1} b_{mj}\,\partial^j_{\theta_0}\frac{\cos(l+\frac{1}{2})\theta_0}{l+\frac{1}{2}} + \sum_{n=0}^{\infty}\frac{\tilde{G}_{mn}}{n+\frac{1}{2}}\Lambda^H_{nl}$$
$$(l = m, m+1, \ldots). \tag{4.38c}$$

Continuity across $t = \theta_0$ of the right-hand side of (4.35) gives

$$b_{m0} = -\kappa^H_m\theta_0 - \frac{2}{\pi}\left[(4ika\beta_m + \bar{\beta}_m)\sin\frac{\theta_0}{2} - \sum_{n=0}^{\infty}\frac{\tilde{G}_{mn}}{n+\frac{1}{2}}\sin\left(n+\frac{1}{2}\right)\theta_0\right]. \tag{4.39}$$

Note that b_{m0} does not contribute directly to the solution coefficients B_{ml} $(l = m, m+1, \ldots)$.

Equations (4.38) are general solutions of (4.10) subject to Meixner's edge conditions (4.30b). As with the TE case, this has been confirmed numerically for $m = 1,2,3$. Explicitly for $m = 1$ the resulting solution system

$$B_{1l} = -(4ika\beta_1 + \bar{\beta}_1)\Lambda^H_{0l} + \sum_{n=1}^{\infty}\frac{G_{1n}}{n+\frac{1}{2}}\Lambda^H_{nl} \quad (l = 1,2,\ldots), \tag{4.40a}$$

$$0 = \bar{\beta}_1 - (4ika\beta_1 + \bar{\beta}_1)\Lambda^H_{00} + \sum_{n=1}^{\infty}\frac{G_{1n}}{n+\frac{1}{2}}\Lambda^H_{n0} \tag{4.40b}$$

yields the coefficients

$$B_{1l} = \sum_{n=1}^{\infty}\frac{G_{1n}}{n+\frac{1}{2}}\Gamma^H_{1,nl} + \bar{\beta}_1 L^H_{1l} \quad (l = 1,2,\ldots), \tag{4.41a}$$

where

$$\Gamma^H_{1,nl} = \Lambda^H_{nl} - \Lambda^H_{n0}\Lambda^H_{0l}/\Lambda^H_{00}, \tag{4.41b}$$

$$L^H_{1l} = -\Lambda^H_{0l}/\Lambda^H_{00}. \tag{4.41c}$$

Satisfaction of the $m = 1$ versions of the original dual series system (4.10) is guaranteed since *on the metal* $(0 \leqslant \theta < \theta_0)$,

$$\sum_{l=1}^{\infty}(2l+1)\Gamma^H_{1,nl}P_l^{-1} - (2n+1)P_n^{-1} - L^H_{1n}P_0^{-1} = 0 \quad (n = 1,2,\ldots), \tag{4.42a}$$

$$\sum_{l=1}^{\infty}(2l+1)L^H_{1l}P_l^{-1} - \frac{\Lambda^H_{00}-1}{\Lambda^H_{00}}P_0^{-1} = 0, \tag{4.42b}$$

and since *in the aperture* $(\theta_0 < \theta \leqslant \pi)$,

$$\sum_{l=1}^{\infty}\Gamma^H_{1,nl}P_l^{-1} = 0, \tag{4.43a}$$

$$\sum_{l=1}^{\infty}L^H_{1l}P_l^{-1} - \bar{P}_0^{-1} = 0. \tag{4.43b}$$

Finally, consider the $m = 0$ case. Introducing the modified coefficients

$$\tilde{B}_{0n} = \begin{cases} B_{00} & (n = 0), \\ B_{0n} & (n \geqslant 1), \end{cases} \tag{4.44}$$

and the modified constants

$$\beta'_0 = \beta_0 - B_{00}/4ika, \tag{4.45a}$$

$$\bar{\beta}'_0 = \bar{\beta}_0 + B_{00}, \tag{4.45b}$$

the dual series system (4.8′) can be rewritten as

$$\sum_{n=0}^{\infty}\left(n+\frac{1}{2}\right)\tilde{B}_{0n}P_n = -2ika\beta'_0P_0 + \sum_{n=1}^{\infty}G_{0n}P_n \quad (0 \leqslant \theta < \theta_0), \tag{4.46a}$$

$$\sum_{n=0}^{\infty}\tilde{B}_{0n}P_n = \tilde{\beta}'_0P_0 \quad (\theta_0 < \theta \leqslant \pi). \tag{4.46b}$$

Our TM solution process leads to the solution coefficients

$$B_{0l} = \sum_{n=1}^{\infty}\frac{G_{0n}}{n+\frac{1}{2}}\Gamma^H_{0,nl} \quad (l = 1,2,\ldots), \tag{4.47a}$$

where

$$\Gamma^H_{0,nl} = \Lambda^H_{nl} - \frac{\sin(n+\frac{1}{2})\theta_0}{\sin\theta_0/2}\Lambda^H_{0l} \tag{4.47b}$$

and the constant

$$\bar{\beta}_0 = -\sum_{n=1}^{\infty}\frac{G_{0n}}{n+\frac{1}{2}}\Gamma^H_{0,n0}. \tag{4.47c}$$

The remaining constant β_0 follows immediately from the continuity condition:

$$(4ika\beta'_0 + \bar{\beta}'_0) \equiv 4ika\beta_0 + \bar{\beta}_0 = \sum_{n=1}^{\infty}\frac{G_{0n}}{n+\frac{1}{2}}\frac{\sin(n+\frac{1}{2})\theta_0}{\sin\theta_0/2}. \tag{4.48}$$

C. Coupled TE and TM solution systems

The modal coefficients of the original electromagnetics problem can now be constructed from the TE and TM dual series results. The TE solution systems for $m \geqslant 1$, (4.23), are still coupled to the corresponding TM solution systems (4.38) through the constraint relations (3.10) and (3.11). For $m = 0$ only TM coefficients exist, and they are generated from (4.47). For each m an infinite linear system of the form (an invertible Fredholm system of the second kind)

$$V_{ml} + \sum_{n=0}^{\infty}\mathscr{M}_{m,nl}V_{mn} = \sum_{n=0}^{\infty}\mathscr{N}_{m,nl}W_{mn} \quad (l = 0,1,2,\ldots) \tag{4.49}$$

is obtained and must be solved. A solution process analogous to the one developed in Ref. 20 can then be applied.

The infinite linear system (4.49) is reduced to a finite one by recognizing that as $n \to \infty$ several terms rapidly go to zero. In particular, the TE and TM solutions have been constructed so that for all m

$$\lim_{n\to\infty}\mathscr{M}_{m,nl} \sim \mathscr{O}(n^{-3}), \tag{4.50a}$$

$$\lim_{n\to\infty}W_{mn} \sim \mathscr{O}(n^{m-3/2-n}). \tag{4.50b}$$

Let us assume that N unknown coefficients are desired: $V_{m1},\ldots,V_{mN}$. Truncation then occurs in (4.49) after the N th term and the following square system results:

$$V_{ml} + \sum_{n=0}^{N}\mathscr{M}_{m,nl}V_{mn} = \sum_{n=0}^{N}\mathscr{N}_{m,nl}W_{mn} \quad (l = 0,1,\ldots,N). \tag{4.51}$$

This system can be solved numerically, for instance, by Gauss elimination. Any additional coefficients can then be

generated recursively from (4.51) by setting $l = N+1$, $N+2,\ldots,L$.

To illustrate this procedure, consider the $m=1$ case. Introducing the terms $\xi = -\bar{\beta}_1$, $\eta = \alpha_1$, $\bar{f}_{1n} = 2ikaf_{1n}$, $\bar{g}_{1n} = -2ikag_{1n}/(n+\frac{1}{2})$, and $\bar{A}_{1n} = A_{1n}/(n+\frac{1}{2})$ and combining the constraint conditions (3.10) and (3.11), the orthogonality relations (4.26b) and (4.40b), and the coefficient expressions (4.27) and (4.41), one obtains the solution system ($l = 1,2,\ldots$)

$$(2ikaL^{E}_{1l})\xi + \bar{A}_{1l} + \sum_{n=1}^{\infty}(\chi^{\phi}_{n}\Gamma^{E}_{1,nl})\bar{A}_{1n} = \sum_{n=1}^{\infty}\Gamma^{E}_{1,nl}\bar{f}_{1n}, \tag{4.52a}$$

$$(L^{H}_{1l})\xi + B_{1l} + \sum_{n=1}^{\infty}(\tilde{\chi}^{\psi}_{n}\Gamma^{H}_{1,nl})B_{1n} = \sum_{n=1}^{\infty}\Gamma^{H}_{1,nl}\bar{g}_{1n}, \tag{4.52b}$$

$$[2ika(1-\Lambda^{E}_{00})]\xi + (-2ika\Lambda^{E}_{00})\eta + \sum_{n=1}^{\infty}(\chi^{\phi}_{n}\Lambda^{E}_{n0})\bar{A}_{1n} = \sum_{n=1}^{\infty}\Lambda^{E}_{n0}\bar{f}_{1n}, \tag{4.52c}$$

$$(1-\Lambda^{H}_{00})\xi + [-4(ka)^2\Lambda^{H}_{00}]\eta + \sum_{n=1}^{\infty}(\tilde{\chi}^{\psi}_{n}\Lambda^{H}_{n0})B_{1n} = \sum_{n=1}^{\infty}\Lambda^{H}_{n0}\bar{g}_{1n}. \tag{4.52d}$$

These equations clearly are coupled and take the form of (4.49). The infinite system (4.52) is reduced to a finite one by noticing that

$$\lim_{n\to\infty}\chi^{\phi}_{n}\Gamma^{E}_{1,nl} \sim \lim_{n\to\infty}\chi^{\psi}_{n}\Gamma^{H}_{1,nl} \sim \mathcal{O}(n^{-3}), \tag{4.53a}$$

$$\lim_{n\to\infty}\bar{f}_{1n} \sim \lim_{n\to\infty}\bar{g}_{1n} \sim \mathcal{O}(n^{-(n+1/2)}). \tag{4.53b}$$

Assuming that the coefficients $\bar{A}_{1n}$ and B_{1n} are desired for $n = 1,\ldots,N$, the truncated solution system is

$$(2ikaL^{E}_{1l})\xi + \bar{A}_{1l} + \sum_{n=1}^{N}(\chi^{\phi}_{n}\Gamma^{E}_{1,nl})\bar{A}_{1n} = \sum_{n=1}^{N}\Gamma^{E}_{1,nl}\bar{f}_{1n} \quad (l = 1,2,\ldots,N), \tag{4.54a}$$

$$(L^{E}_{1l})\xi + B_{1l} + \sum_{n=1}^{N}(\tilde{\chi}^{\psi}_{n}\Gamma^{H}_{1,nl})B_{1n} = \sum_{n=1}^{N}\Gamma^{E}_{1,nl}\bar{g}_{1n} \quad (l = 1,2,\ldots,N), \tag{4.54b}$$

$$[2ika(1-\Lambda^{E}_{00})]\xi + (-2ika\Lambda^{E}_{00})\eta + \sum_{n=1}^{N}(\chi^{\phi}_{n}\Lambda^{E}_{n0})\bar{A}_{1n} = \sum_{n=1}^{N}\Lambda^{E}_{n0}\bar{f}_{1n}, \tag{4.54c}$$

$$(1-\Lambda^{H}_{00})\xi + [-4(ka)^2\Lambda^{H}_{00}]\eta + \sum_{n=1}^{N}(\tilde{\chi}^{\psi}_{n}\Lambda^{H}_{n0})B_{1n} = \sum_{n=1}^{\infty}\Lambda^{H}_{n0}\bar{g}_{1n}. \tag{4.54d}$$

Numerical results generated from this system will be presented below.

V. NORMAL INCIDENCE CASE

The plane wave is normally incident when $\theta^{\text{inc}} = 0$ or $\theta^{\text{inc}} = \pi$. The $\theta^{\text{inc}} = 0$ geometry is illustrated in Fig. 2. The

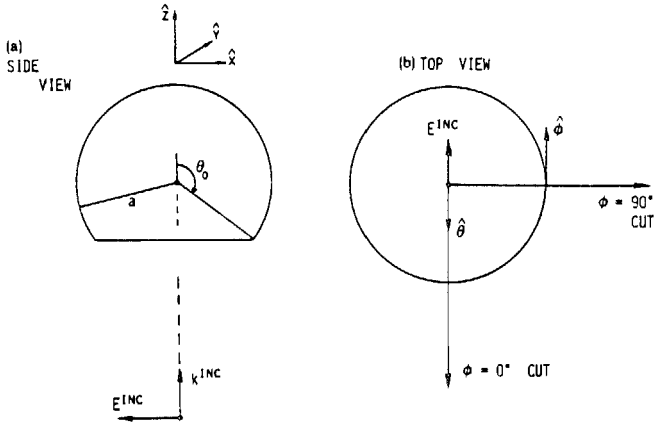

FIG. 2. Configuration of the scattering of a normally incident plane wave from a spherical shell having a circular aperture: (a) side view, (b) top view.

restriction to normal incidence provides a great simplification because

$$\left[\frac{mP^{m}_{n}(\cos\theta)}{\sin\theta}\right](\theta = \theta^{\text{inc}} = 0) = \left[+\frac{\partial}{\partial\theta}P^{m}_{n}(\cos\theta)\right](\theta = \theta^{\text{inc}} = 0) = \frac{n(n+1)}{2}\delta_{m1}, \tag{5.1}$$

$$\left[\frac{mP^{m}_{n}(\cos\theta)}{\sin\theta}\right](\theta = \theta^{\text{inc}} = \pi) = \left[-\frac{\partial}{\partial\theta}P^{m}_{n}(\cos\theta)\right](\theta = \theta^{\text{inc}} = \pi) = (-1)^{n+1}\frac{n(n+1)}{2}\delta_{m1}. \tag{5.2}$$

As a result, the potentials reduce to single sums involving only the $m = 1$ azimuthal mode:

$$\begin{pmatrix}\Phi^{\text{inc}}\\ \Phi^{s}\end{pmatrix} = -E_0\begin{pmatrix}\Phi^{\text{inc}}_{1}\\ \Phi^{s}_{1}\end{pmatrix}\sin\phi, \tag{5.3a}$$

$$\begin{pmatrix}\Psi^{\text{inc}}\\ \Psi^{s}\end{pmatrix} = Y_0E_0\begin{pmatrix}\Psi^{\text{inc}}_{1}\\ \Psi^{s}_{1}\end{pmatrix}\cos\phi. \tag{5.3b}$$

Consequently, the coupled dual series systems for normal incidence coincide with the $m = 1$ case treated in Sec. IV; the modal coefficients A_{1n} and B_{1n} are numerically generated from the solution system (4.54) with $\theta^{\text{inc}} = 0$ or π. The field components for normal incidence in terms of these coefficients are listed in Table II for convenient reference. These expressions isolate the coefficients, the r, the θ, and the ϕ dependencies; hence they are very useful for current, energy density, and cross section calculations.

An important analytical property of the dual series solution is simply revealed by the normal incidence case results. This is its trivial recovery of the scattering coefficients for the closed sphere case when $\theta_0 = \pi$. Let $\theta^{\text{inc}} = 0$. The terms

$$\Lambda^{E,H}_{nl}(\theta_0 = \pi) = \delta_{nl}, \tag{5.4}$$

TABLE II. Electric and magnetic field components for normal incidence.

Field components

$$E_r = E_0 \sum_{n=1}^{\infty} in(n+1)\tau_{1n} \frac{Z_n(kr)}{kr} P_n^{-1}(\cos\theta)\cos\phi$$

$$E_\theta = E_0 \sum_{n=1}^{\infty} \left\{\sigma_{1n} Z_n(kr)\bar{\nu}_{1n}(\theta) - i\tau_{1n} \frac{[krZ_n(kr)]'}{kr} \bar{w}_{1n}(\theta)\right\}\cos\phi$$

$$E_\phi = E_0 \sum_{n=1}^{\infty} \left\{\sigma_{1n} Z_n(kr)\bar{w}_{1n}(\theta) - i\tau_{1n} \frac{[krZ_n(kr)]'}{kr} \bar{\nu}_{1n}(\theta)\right\}\sin\phi$$

$$H_r = -Y_0 E_0 \sum_{n=1}^{\infty} in(n+1)\sigma_{1n} \frac{Z_n(kr)}{kr} P_n^{-1}(\cos\theta)\sin\phi$$

$$H_\theta = -Y_0 E_0 \sum_{n=1}^{\infty} \left\{\tau_{1n} Z_n(kr)\bar{\nu}_{1n}(\theta) - i\sigma_{1n} \frac{[krZ_n(kr)]'}{kr} \bar{w}_{1n}(\theta)\right\}\sin\phi$$

$$H_\phi = +Y_0 E_0 \sum_{n=1}^{\infty} \left\{\tau_{1n} Z_n(kr)\bar{w}_{1n}(\theta) - i\sigma_{1n} \frac{[krZ_n(kr)]'}{kr} \bar{\nu}_{1n}(\theta)\right\}\cos\phi$$

Incident field

$$Z_n(kr) = j_n(kr)$$

$$\sigma_{1n}^{\text{inc}} = i^n(2n+1)\begin{cases} 1, & \theta^{\text{inc}} = 0 \\ (-1)^{n+1}, & \theta^{\text{inc}} = \pi \end{cases}$$

$$\tau_{1n}^{\text{inc}} = i^n(2n+1)\begin{cases} -1, & \theta^{\text{inc}} = 0 \\ (-1)^{n+1}, & \theta^{\text{inc}} = \pi \end{cases}$$

Scattered field for $r < a$

$$Z_n(kr) = j_n(kr)$$

$$\sigma_{1n}^{<} = A_{1n} h_n(ka)$$

$$\tau_{1n}^{<} = B_{1n}[kah_n(ka)]'$$

Scattered field for $r > a$

$$Z_n(kr) = h_n(kr)$$

$$\sigma_{1n}^{>} = A_{1n} j_n(ka)$$

$$\tau_{1n}^{>} = B_{1n}[kaj_n(ka)]'$$

Terms

$$\bar{\nu}_{1n}(\theta) = \frac{P_n^{-1}(\cos\theta)}{\sin\theta} = \frac{-1}{n(n+1)} \frac{P_n^1(\cos\theta)}{\sin\theta}$$

$$\bar{w}_{1n}(\theta) = -\partial_\theta P_n^{-1}(\cos\theta) = \frac{1}{n(n+1)} \partial_\theta P_n^1(\cos\theta)$$

so that the modal coefficients

$$A_{1l} = \frac{ika(2l+1)f_{1l}}{1+\chi_l^\phi} = \frac{-i^l(2l+1)}{h_l(ka)}, \tag{5.5a}$$

$$B_{1l} = \frac{-4ikag_{1l}}{(2l+1)(1+\bar{\chi}_l^\psi)} = \frac{i^l(2l+1)}{[kah_l(ka)]'}. \tag{5.5b}$$

Referring to Eqs. (2.4)–(2.14) and to Table II, this means that (1) for $r \leqslant a$, $\Phi_{1<}^s = -\Phi_{1<}^{\text{inc}}$ and $\Psi_{1<}^s = -\Psi_{1<}^{\text{inc}}$ so that the total potentials, hence the fields, are identically zero there and the boundary conditions $E_{\tan}^s(r=a) = -E_{\tan}^{\text{inc}}(r=a)$ are satisfied and (2) for $r > a$, the standard results for the scattered potentials—fields given, for instance, in Ref. 38, Sec. 6.9, and Ref. 39, Sec. 16.9, are recovered.

VI. CURRENTS ON THE SPHERICAL SHELL

The most stringent test of the dual series solution is the calculation of the currents J_θ and J_ϕ on the open spherical shell. Verification of the required current behavior near the aperture edge is immediately apparent from graphical results. Moreover, the vanishing of the current in the aperture is an excellent test of the results and reflects the satisfaction of the corresponding TE and TM dual series equations in that region. For normal incidence the current expressions (4.11) simply become

$$J_\theta(\theta,\phi) = \left[\frac{-Y_0 E_0}{(ka)^2}\right]\cos\phi \sum_{n=1}^{\infty} \left[A_{1n} \frac{P_n^{-1}(\cos\theta)}{\sin\theta} + ikaB_{1n}\, \partial_\theta P_n^{-1}(\cos\theta)\right], \tag{6.1}$$

$$J_\phi(\theta,\phi) = \left[\frac{+Y_0 E_0}{(ka)^2}\right]\sin\phi \sum_{n=1}^{\infty} \left[A_{1n}\, \partial_\theta P_n^{-1}(\cos\theta) + ikaB_{1n} \frac{P_n^{-1}(\cos\theta)}{\sin\theta}\right]. \tag{6.2}$$

A. Analytical preconditioning

Consider first the quasistatic case where $ka = 0.01$, $\theta_0 = 120°$. In all of the examples $a = 1.0$. Simply performing the sums in (6.2) with the solution coefficients generated from Eqs. (4.54), we find that the number of terms required to track the square root singularity in J_ϕ is large. The polynomial sum $\Sigma_{n=1}^{L} A_{1n}\, \partial_\theta P_n^{-1}$ is the cause of this difficulty. However, the truncation number N (Sec. IV C) need not be large; and the remaining coefficients $n = N+1,\ldots,L$ are recursively defined from (4.54a) and (4.54b). This is demonstrated in Fig. 3 where the real part of $J_\phi(\theta,\pi/2)[(ka)^2/Y_0E_0]$ is given for various truncation numbers. In Figs. 3(a), 3(b), and 3(c) the truncation numbers $N = 5$ and $L = 50$, 500, and 5000, respectively. However, the results may be improved by treating the singularity analytically as follows.

Inserting the coefficient expressions (4.27a) and (4.41) into (6.1) and (6.2) and referring to the definitions given in Table III, the current components

$$J_\theta = \frac{-Y_0 E_0}{2(ka)^2}\cos\phi \left\{\sum_{n=1}^{N}\left[F_{1n}\frac{K_n^E(\theta)}{\sin\theta} + 4ika\frac{G_{1n}}{2n+1}\partial_\theta K_n^H(\theta)\right] - 2ka\xi\left[\frac{K_0^E(\theta)}{\sin\theta} + \partial_\theta K_0^H(\theta)\right]\right\}, \tag{6.3}$$

$$J_\phi = \frac{+Y_0 E_0}{2(ka)^2}\sin\phi \left\{\sum_{n=1}^{N}\left[F_{1n}\,\partial_\theta K_n^E(\theta) + 4ika\frac{G_{1n}}{2n+1}\frac{K_n^H(\theta)}{\sin\theta}\right] - 2ka\xi\left[\partial_\theta K_0^E(\theta) + \frac{K_0^H(\theta)}{\sin\theta}\right]\right\} \tag{6.4}$$

result. Two advantages of these expressions are immediate. First, the coefficients obtained from the matrix inversion can be used directly without calculating any additional coefficients by recursion. Second, the currents vanish analytically in the aperture. The terms proportional to F_{1n} and G_{1n} give

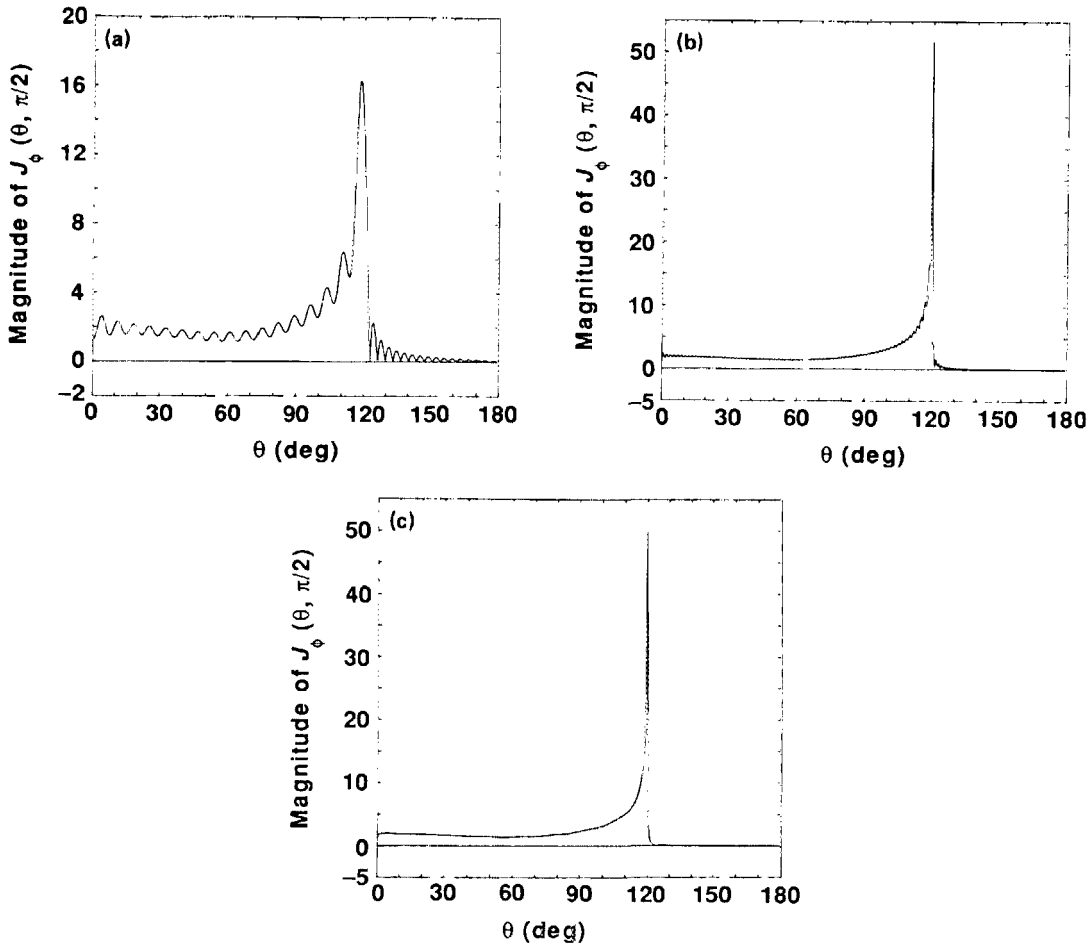

FIG. 3. Brute force summation of the $J_\phi(\theta,\pi/2)$ current expression requires a large number of terms to eliminate the numerical Gibbs' phenomena: (a) $N=5, L=50$; (b) $N=5, L=500$; (c) $N=5, L=5000$.

TABLE III. Special functions and relations for the current expressions.

$$K_0^H(\theta)=\sum_{l=1}^{\infty} L_{1l}^H(\theta_0)P_l^{-1}(\cos\theta)=\begin{cases}+\bar{P}_0^{-1}(\theta) & (\theta_0<\theta\leqslant\pi)\\ +\bar{P}_0^{-1}(\theta)+s_H(\theta) & (0\leqslant\theta<\theta_0)\end{cases}$$

$$K_0^E(\theta)=\sum_{l=1}^{\infty}(2l+1)L_{1l}^E(\theta_0)P_l^{-1}(\cos\theta)=\begin{cases}+\bar{P}_0^{-1}(\theta) & (\theta_0<\theta\leqslant\pi)\\ +\bar{P}_0^{-1}(\theta)+s_E(\theta) & (0\leqslant\theta<\theta_0)\end{cases}$$

$$K_n^H(\theta)=\sum_{l=1}^{\infty}\Gamma_{1,nl}^H(\theta_0)P_l^{-1}(\cos\theta)=\begin{cases}0 & (\theta_0<\theta\leqslant\pi)\\ S_n^H(\theta)+\Lambda_{n0}^H s_H(\theta) & (0\leqslant\theta<\theta_0)\end{cases}$$

$$K_n^E(\theta)=\sum_{l=1}^{\infty}(2l+1)\Gamma_{1,nl}^E(\theta_0)P_l^{-1}(\cos\theta)=\begin{cases}0 & (\theta_0<\theta\leqslant\pi)\\ S_n^E(\theta)+\Lambda_{n0}^E s_E(\theta) & (0\leqslant\theta<\theta_0)\end{cases}$$

where

$$s_H(\theta)=\frac{+4}{\pi\Lambda_{00}^H\sin\theta}\left\{\cos\frac{\theta_0}{2}\left(\cos^2\frac{\theta}{2}-\cos^2\frac{\theta_0}{2}\right)^{1/2}-\cos^2\frac{\theta}{2}\left[\arccos\left(\frac{\cos(\theta_0/2)}{\cos(\theta/2)}\right)\right]\right\}$$

$$s_E(\theta)=\frac{-4}{\pi\Lambda_{00}^E\sin\theta}\left\{\cos\frac{\theta_0}{2}\left(\cos^2\frac{\theta}{2}-\cos^2\frac{\theta_0}{2}\right)^{1/2}+\cos^2\frac{\theta}{2}\left[\arccos\left(\frac{\cos(\theta_0/2)}{\cos(\theta/2)}\right)\right]\right\}$$

$$u_n(\theta)=\frac{4}{\pi}\cos\left(n+\frac{1}{2}\right)\theta_0\frac{[2(\cos\theta-\cos\theta_0)]^{1/2}}{\sin\theta}$$

$$S_n^H(\theta)=\sum_{l=1}^{\infty}\Lambda_{nl}^H P_l^{-1}(\cos\theta)+\Lambda_{n0}^H\bar{P}_0^{-1}(\theta)$$

$$S_n^E(\theta)=u_n(\theta)+(2n+1)S_n^H(\theta)$$

TABLE IV. Derivative relations employed in the current calculations.

$$\partial_\theta s_H(\theta) = \frac{-4}{\pi \Lambda_{00}^H}\left\{\frac{-\cos\theta}{\sin^2\theta}\left[\cos\frac{\theta_0}{2}\left(\cos^2\frac{\theta}{2}-\cos^2\frac{\theta_0}{2}\right)^{1/2}-\cos^2\frac{\theta}{2}\arccos\left(\frac{\cos(\theta_0/2)}{\cos(\theta/2)}\right)\right]+\frac{1}{2}\arccos\left(\frac{\cos(\theta_0/2)}{\cos(\theta/2)}\right)\right\}$$

$$\partial_\theta s_E(\theta) = \frac{+4}{\pi \Lambda_{00}^E}\left\{\frac{-\cos\theta}{\sin^2\theta}\left[\cos\frac{\theta_0}{2}\left(\cos^2\frac{\theta}{2}-\cos^2\frac{\theta_0}{2}\right)^{1/2}\right.\right.$$

$$\left.\left.+\cos^2\frac{\theta}{2}\arccos\left(\frac{\cos(\theta_0/2)}{\cos(\theta/2)}\right)\right]-\frac{1}{2}\arccos\left(\frac{\cos(\theta_0/2)}{\cos(\theta/2)}\right)-\frac{\cos(\theta_0/2)}{2(\cos^2(\theta/2)-\cos^2(\theta_0/2))^{1/2}}\right\}$$

$$\partial_\theta u_n(\theta) = \frac{-8\cos(n+\frac{1}{2})\theta_0}{\pi\sin^2\theta(\cos^2(\theta/2)-\cos^2(\theta_0/2))^{1/2}}\left[\cos^2\frac{\theta}{2}\left(\cos^2\frac{\theta}{2}-\cos^2\frac{\theta_0}{2}\right)+\sin^2\frac{\theta}{2}\cos^2\frac{\theta_0}{2}\right]$$

$$\partial_\theta S_n^H(\theta) = \sum_{l=1}^{\infty}\Lambda_{nl}^H\partial_\theta P_l^{-1}-\Lambda_{n0}^H\frac{\bar{P}_0^{-1}}{\sin\theta}$$

$$\partial_\theta S_n^E(\theta) = \partial_\theta u_n(\theta) + (2n+1)\partial_\theta S_n^H(\theta)$$

no contributions in the aperture because K_n^E and K_n^H are zero there. The terms proportional to ξ also reduce to zero there by (3.9).

Restricting now our attention to the behavior of the current on the metal, (6.3) and (6.4) yield

$$J_\theta = -\frac{Y_0E_0}{2(ka)^2}\cos\phi\left\{\sum_{n=1}^{N}\left[F_{1n}\frac{K_n^E(\theta)}{\sin\theta}+4ika\frac{G_{1n}}{2n+1}\partial_\theta K_n^H(\theta)\right]-2ka\xi\left[\frac{s_E(\theta)}{\sin\theta}+\partial_\theta s_H(\theta)\right]\right\}\quad(0\leqslant\theta\leqslant\theta_0),\tag{6.3'}$$

$$J_\phi = +\frac{Y_0E_0}{2(ka)^2}\sin\phi\left\{\sum_{n=1}^{N}\left[F_{1n}\,\partial_\theta K_n^E(\theta)+4ika\frac{G_{1n}}{2n+1}\frac{K_n^H(\theta)}{\sin\theta}\right]-2ka\xi\left[\partial_\theta s_E(\theta)+\frac{s_H(\theta)}{\sin\theta}\right]\right\}\quad(0\leqslant\theta\leqslant\theta_0).\tag{6.4'}$$

Near the aperture edge $\theta=\theta_0$ the terms $K_n^E(\theta)$ and $K_n^H(\theta)$ behave, respectively, as $(\theta_0-\theta)^{1/2}$ and $(\theta_0-\theta)^{3/2}$. Consequently, the square root singularity in J_ϕ is generated by the term $\partial_\theta K_n^E(\theta)$ and, referring to Table IV, by the term $\partial_\theta s_E(\theta)$. If the former is generated numerically, a large number of coefficients are required. However, referring to

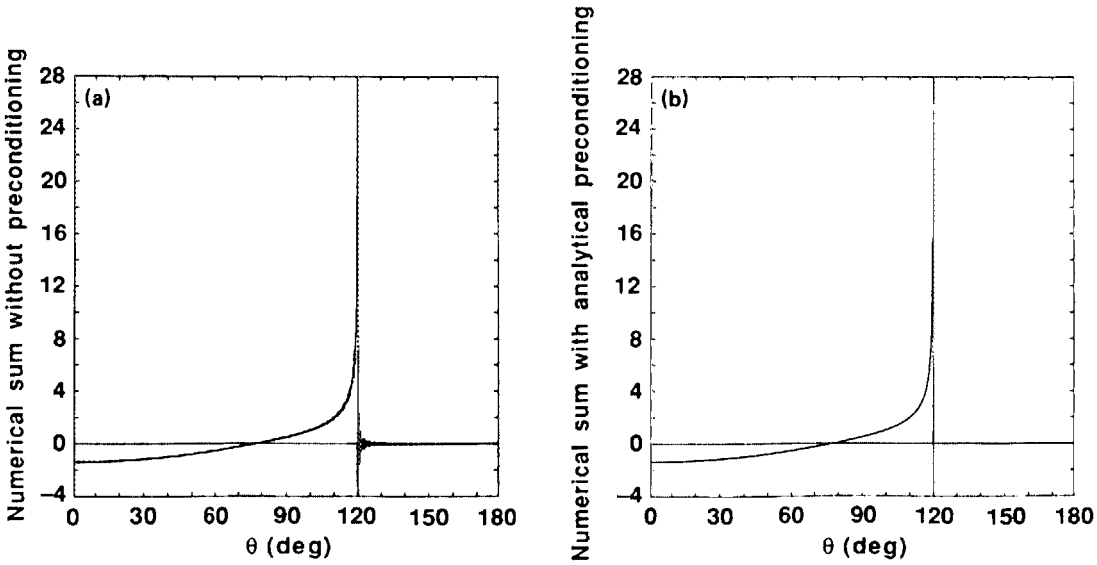

FIG. 4. The Gibbs' phenomena is removed by handling the edge singularity analytically. The dominant sum in the $J_\phi(\theta,\pi/2)$ expression (a) without analytical preconditioning, and (b) with analytical preconditioning.

Tables III and IV for $0 \leqslant \theta < \theta_0$, the relations

$$K_n^E(\theta) = S_n^E(\theta) + \Lambda_{n0}^E s_E(\theta)$$
$$= (2n+1)S_n^H(\theta) + u_n(\theta) + \Lambda_{n0}^E s_E(\theta), \quad (6.5)$$
$$\partial_\theta K_n^E(\theta) = (2n+1)\partial_\theta S_n^H(\theta) + \partial_\theta u_n(\theta) + \Lambda_{n0}^E \partial_\theta s_E(\theta) \quad (6.6)$$

indicate that one only needs to evaluate $S_n^H(\theta)$ and $\partial_\theta S_n^H(\theta)$ numerically. Since $S_n^H(\theta) = K_n^H(\theta) - \Lambda_{n0}^H s_H(\theta)$ over the metal and near the aperture edge $\partial_\theta K_n^H(\theta) \sim (\theta_0 - \theta)^{1/2}$ and $\partial_\theta s_H \sim (\theta_0 - \theta)^{1/2}$, the term $\partial_\theta S_n^H(\theta) \sim (\theta_0 - \theta)^{1/2}$ near $\theta = \theta_0$, which circumvents the numerical difficulties. The square root singularity is handled analytically through the terms $\partial_\theta u_n$ and $\partial_\theta s_E$. A comparison of $\partial_\theta K_1^E(\theta)$ evaluated directly and with (6.6) is given in Fig. 4. Each sum included 800 terms. As desired, the (numerical) oscillations were removed by the analytical preconditioning.

B. Numerical results

Because the current components

$$J_\theta(\theta,\phi) = J_\theta(\theta,0)\cos\phi, \quad (6.7)$$
$$J_\phi(\theta,\phi) = J_\phi(\theta,\pi/2)\sin\phi, \quad (6.8)$$

their important features are illustrated succinctly by considering $J_\theta(\theta,0)$ and $J_\phi(\theta,\pi/2)$. Examples of the scaled current terms $\mathcal{J}_\theta = J_\theta(\theta,0)[-2(ka)^2/Y_0E_0]$ and $\mathcal{J}_\phi = J_\phi(\theta,\pi/2)[2(ka)^2/Y_0E_0]$ are given in Figs. 5–10 for various ka, aperture sizes, and angles of incidence.

Values of $|\mathcal{J}_\theta|$ and $|\mathcal{J}_\phi|$ are given in Figs. 5 and 6 for the quasistatic limit ($ka = 0.01$), the angle of incidence $\theta^{\text{inc}} = 0.0$, and, respectively, the aperture angles $\theta_0 = 120°$ and $\theta_0 = 170°$. For both cases the truncation number $N = 10$. Essentially the same results were generated with $N = 3$. This low truncation number is typical for quasistatic cases because the $n = 1$ term dominates the behavior. The term $|\mathcal{J}_\theta|$ is given in Fig. 7 for $\theta^{\text{inc}} = 0.0$, $\theta_0 = 120°$, and the ka values 1.0, 3.0, 5.0, and 10.0. The corresponding graphs of $|\mathcal{J}_\phi|$ are given in Fig. 8. For all of these cases the truncation number was taken to be $N = 10(ka)$. This choice yields convergent results. The plots in Fig. 7 clearly demonstrate that our solution reproduces the required $(\theta_0 - \theta)^{1/2}$ behavior of J_θ near $\theta = \theta_0$; Fig. 8 demonstrates that the required square root singularity of J_ϕ near $\theta = \theta_0$ is present. In Fig. 9 the terms $\mathrm{Re}(\mathcal{J}_\theta)$ and $|\mathcal{J}_\theta|$ are plotted for $\theta_0 = 120°$, $ka = 1.0$, $\theta^{\text{inc}} = 0°$, and $\theta^{\text{inc}} = 180°$. Very different behaviors are obtained. When the wave is incident on the shell

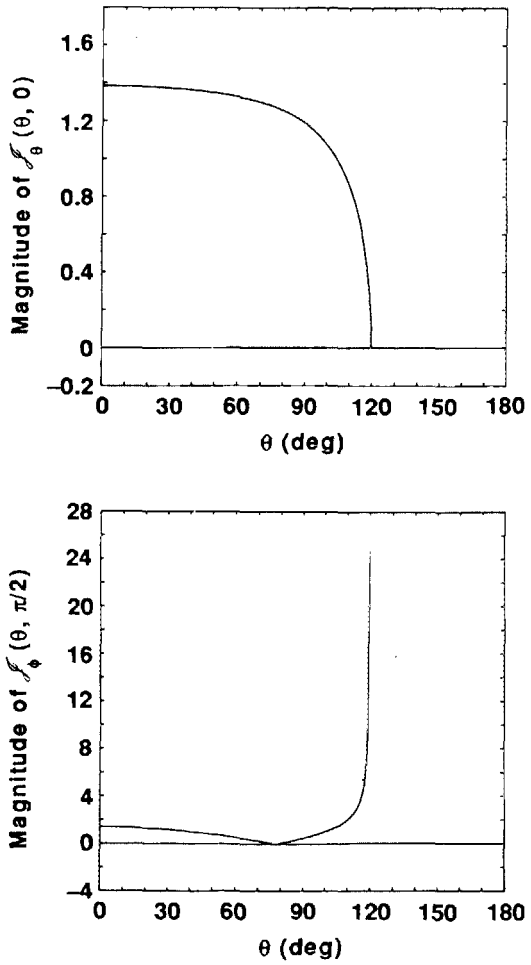

FIG. 5. The magnitudes of the current terms $\mathcal{J}_\theta(\theta,0)$ and $\mathcal{J}_\phi(\theta,\pi/2)$ induced on an open spherical shell with $\theta_0 = 120°$ when $ka = 0.01$ and $\theta^{\text{inc}} = 0.0°$.

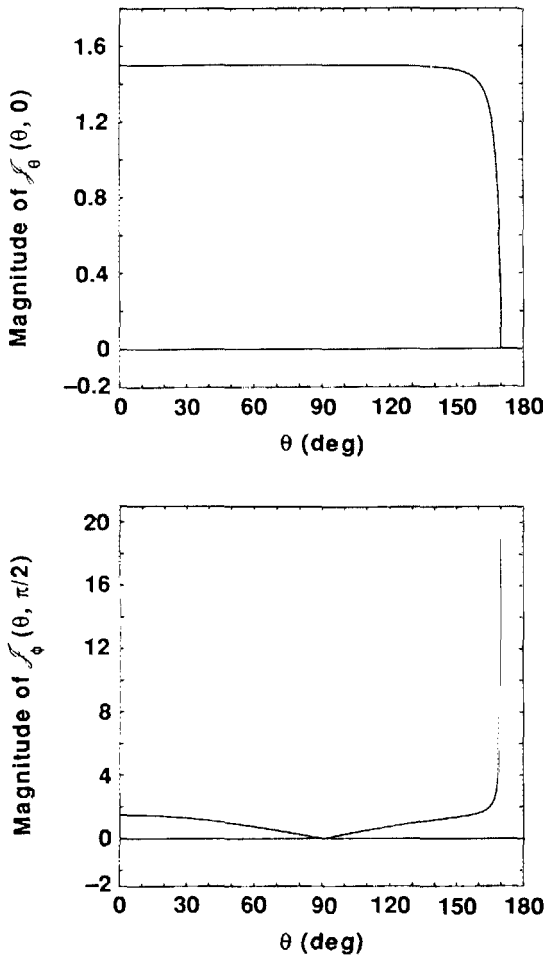

FIG. 6. The magnitudes of the current terms $\mathcal{J}_\theta(\theta,0)$ and $\mathcal{J}_\phi(\theta,\pi/2)$ induced on an open spherical shell with $\theta_0 = 170°$ when $ka = 0.01$ and $\theta^{\text{inc}} = 0.0°$.

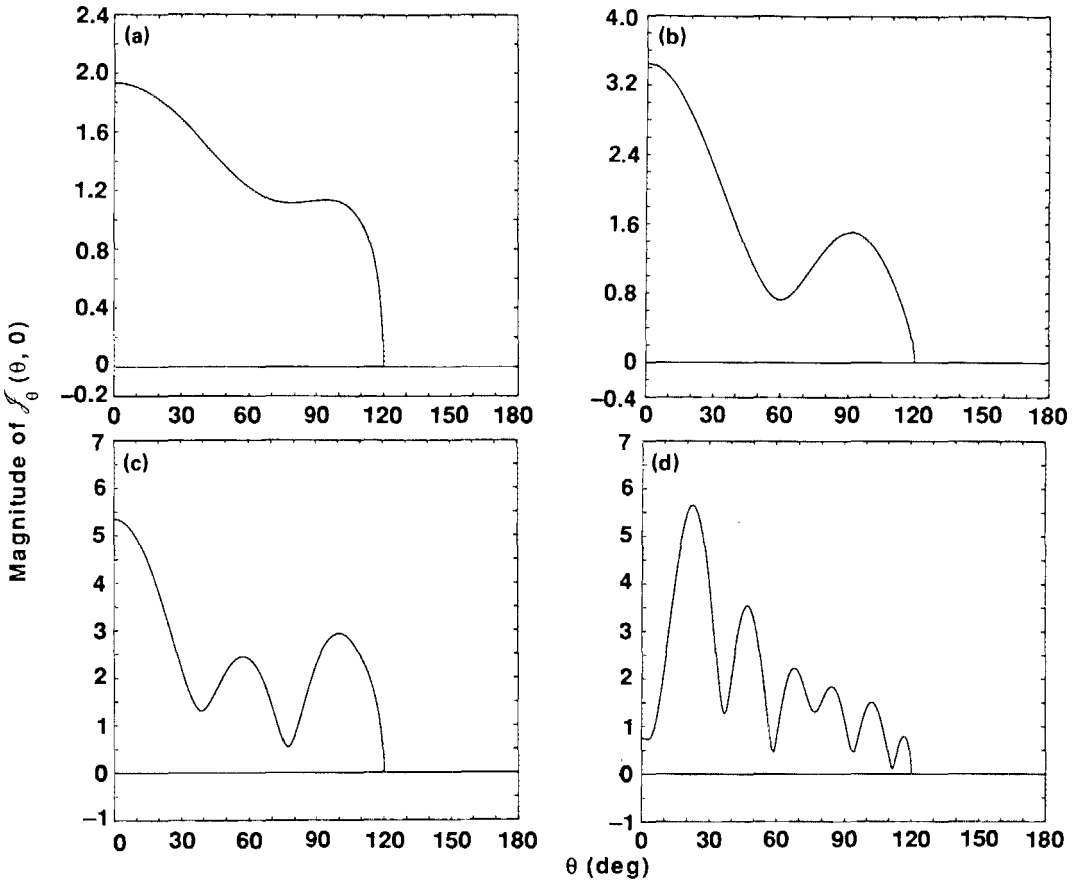

FIG. 7. The current term $\mathscr{J}_\theta(\theta,0)$ induced on an open spherical shell with $\theta_0 = 120°$ when $\theta^{inc} = 0°$ and the ka of the incident plane wave is (a) 1.0, (b) 3.0, (c) 5.0, and (d) 10.0.

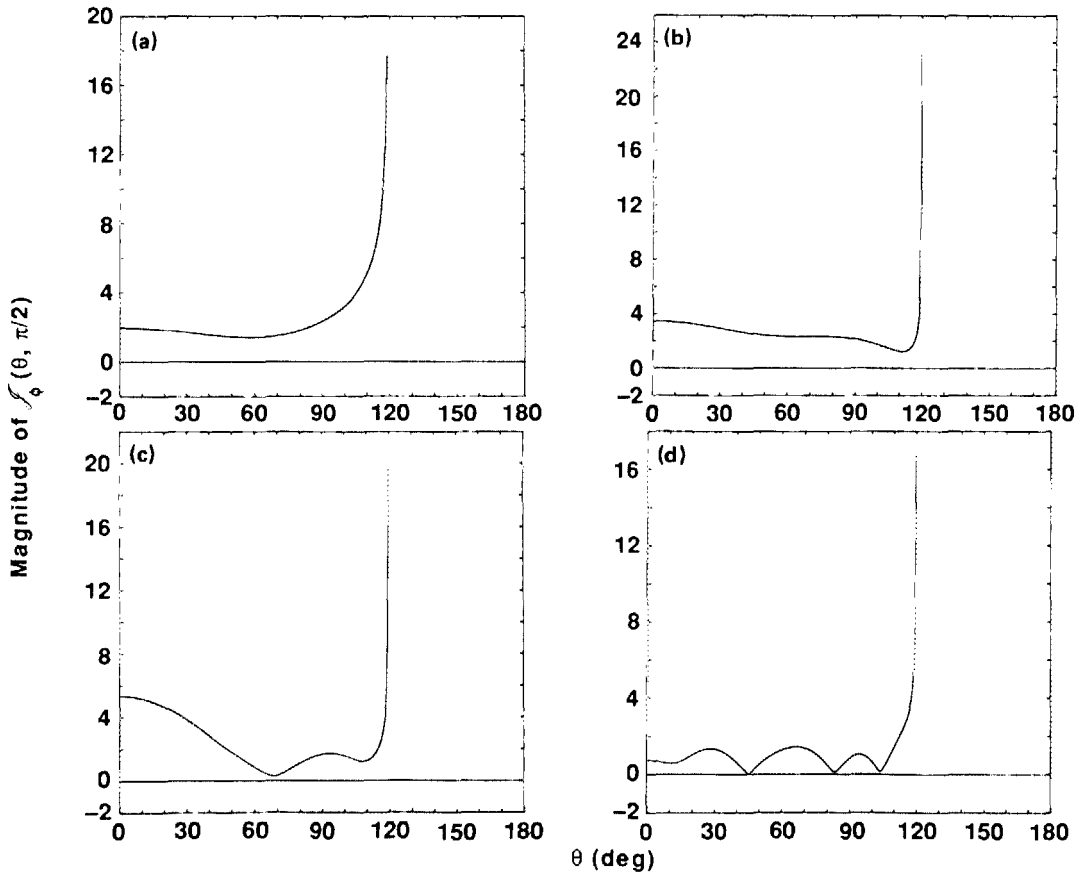

FIG. 8. The current term $\mathscr{J}_\phi(\theta,\pi/2)$ induced on an open spherical shell with $\theta_0 = 120°$ when $\theta^{inc} = 0°$ and the ka of the incident plane wave is (a) 1.0, (b) 3.0, (c) 5.0, and (d) 10.0.

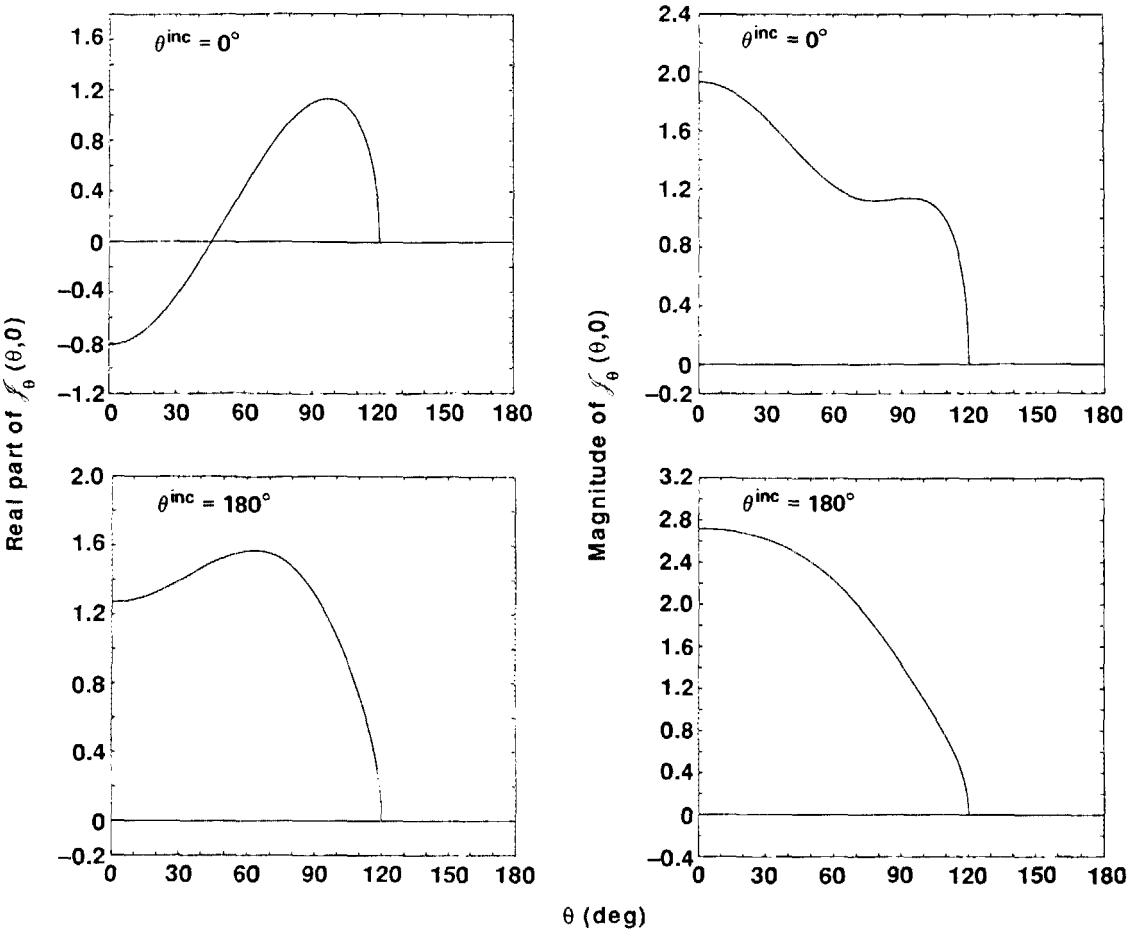

FIG. 9. A comparison of the real part and the magnitude of the current term $\mathcal{J}_\theta(\theta,0)$ induced on an open spherical shell with $\theta_0 = 120°$ when $ka = 1.0$ and the angle of incidence $\theta^{inc} = 0°$ and $\theta^{inc} = 180°$.

($\theta^{inc} = 180°$), $|\mathcal{J}_\theta|$ is much more uniformly distributed over the shell. The hump near the aperture edge which appeared when $\theta^{inc} = 0°$ is no longer present. In Fig. 10, $|\mathcal{J}_\theta|$ is given for $\theta^{inc} = 0°$ and $ka = 1.0$ when $\theta_0 = 120°$ and $\theta_0 = 170°$. The latter case exhibits a more pronounced hump near the aperture edge.

The distributions of J_θ and J_ϕ over the entire spherical shell for $ka = 3$, $\theta^{inc} = 0°$, and $\theta_0 = 120°$ are shown in Figs. 11 and 12. In Fig. 11 the values of $|\mathcal{J}_\theta|$ and $|\mathcal{J}_\phi|$ are replotted in more detail to provide a reference for Figs. 12. In Figs. 12(a) and 12(b) $\mathcal{J}_\theta$ is viewed from the directions ($\theta = 0$, $\phi = 0$) and ($\theta = 76°$, $\phi = 125°$). Dark red represents the largest values; dark blue the smallest ones. The characteristic cosine pattern and null at the aperture edge are very apparent. The corresponding views of $\mathcal{J}_\phi$ are given in Figs. 12(c) and 12(d). The associated sine pattern and edge singularity are nicely reproduced.

The current results have been validated with a totally independent method[19]: a completely numerical solution based upon a method of moments (MoM) analysis of the problem. It has been demonstrated that the MoM solution converges to the dual series results when the former is applicable.

VII. ENERGY DENSITIES

To provide some measure of the degree of coupling of the incident field into the spherical cavity, the energy density at the center of the shell normalized to the incident field energy density there was calculated. This also allows a direct comparison with Senior–Desjardins results.[12,13]

Consider the normal incidence field expressions given in Table II for $r = 0$. With the small argument relations in the Appendix, one obtains ($n \neq 0$)

$$j_n(0) \equiv 0, \quad \{[xj_n(x)]'/x\}_{x=0} = \tfrac{2}{3}\delta_{n1},$$
$$[j_n(x)/x]_{x=0} = \tfrac{1}{3}\delta_{n1}.$$

Moreover, $P_1^{-1}(\cos\theta)/\sin\theta = -\frac{1}{2}$ and $-\partial_\theta P_1^{-1}(\cos\theta) = \cos(\theta/2)$. Therefore, the general electric and magnetic field vectors at the origin are

$$(E_r, E_\theta, E_\phi)(r=0) = (i/3)E_0\tau_{11}(\sin\theta\cos\phi, \cos\theta\cos\phi, -\sin\phi), \quad (7.1)$$

$$(H_r, H_\theta, H_\phi)(r=0) = (i/3)Y_0E_0\sigma_{11}(\sin\theta\sin\phi, \cos\theta\sin\phi, +\cos\phi), \quad (7.2)$$

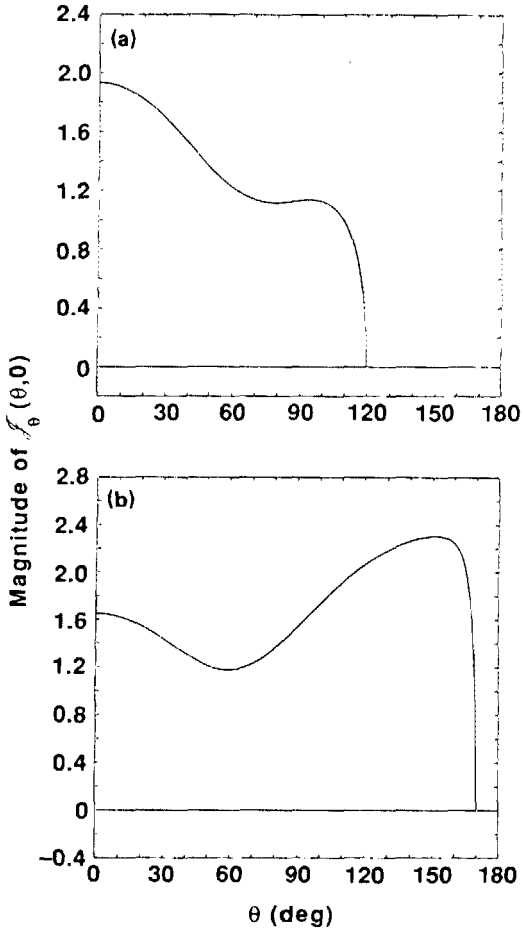

FIG. 10. A comparison of the magnitudes of the current term $\mathcal{J}_\theta(\theta,0)$ induced on open spherical shells with (a) $\theta_0 = 120°$ and (b) $\theta_0 = 170°$, when $ka = 1.0$ and the angle of incidence $\theta^{\text{inc}} = 0°$.

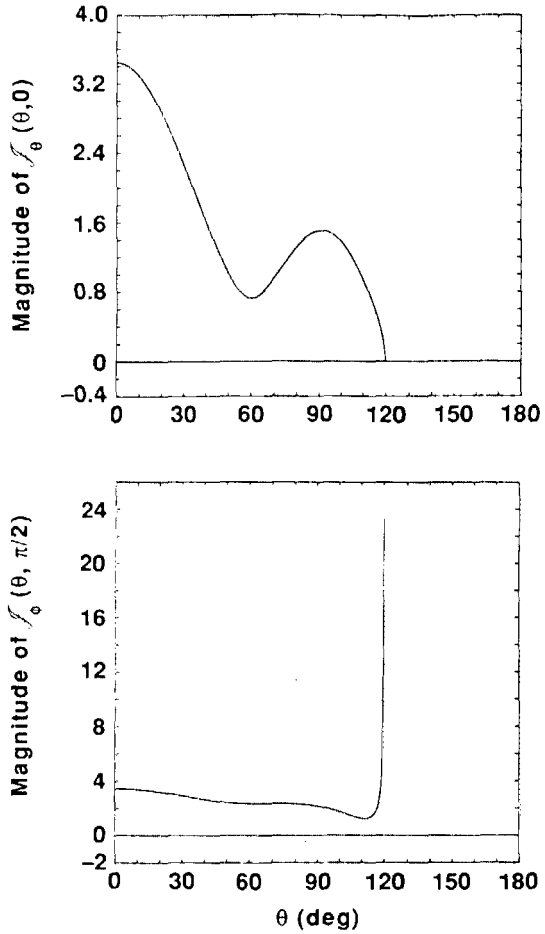

FIG. 11. The magnitudes of the current terms $\mathcal{J}_\theta(\theta,0)$ and $\mathcal{J}_\theta(\theta,\pi/2)$ induced on an open spherical shell with $\theta_0 = 120°$ when $ka = 3.0$ and the angle of incidence $\theta^{\text{inc}} = 0°$.

where

$$\sigma_{11} = \sigma_{11}^{\text{inc}} + \sigma_{11}^{<} = \sigma_{11}^{\text{inc}} + A_{11}h_1(ka), \tag{7.3}$$

$$\tau_{11} = \tau_{11}^{\text{inc}} + \tau_{11}^{<} = \tau_{11}^{\text{inc}} + B_{11}[kah_1(ka)]', \tag{7.4}$$

the incident field terms being

$$\sigma_{11}^{\text{inc}} = \begin{cases} 3i, & \theta^{\text{inc}} = 0, \\ 3i, & \theta^{\text{inc}} = \pi, \end{cases} \tag{7.5}$$

$$\tau_{11}^{\text{inc}} = \begin{cases} -3i, & \theta^{\text{inc}} = 0, \\ 3i, & \theta^{\text{inc}} = \pi. \end{cases} \tag{7.6}$$

Consequently, the associated energy density relation is simply

$$\begin{aligned} U(r=0) &= \tfrac{1}{2}(|\mathbf{E}|^2 + |Z_0\mathbf{H}|^2)(r=0) \\ &= E_0^2(|\sigma_{11}|^2 + |\tau_{11}|^2)/18, \end{aligned} \tag{7.7}$$

which leads to the desired energy density ratio

$$\begin{aligned} &\frac{U_{\text{tot}}(r=0)}{U_{\text{inc}}(r=0)} \\ &= \frac{|\sigma_{11}^{\text{inc}} + \sigma_{11}^{<}|^2 + |\tau_{11}^{\text{inc}} + \tau_{11}^{<}|^2}{|\sigma_{11}^{\text{inc}}|^2 + |\tau_{11}^{\text{inc}}|^2} \\ &= \tfrac{1}{18}\{|3i + A_{11}h_1(ka)|^2 + |3i \mp B_{11}[kah_1(ka)]'^2\} \\ &= \tfrac{1}{2}\{|1 - (i/3)A_{11}h_1(ka)|^2 + |1 \\ &\quad \pm (i/3)B_{11}[kah_1(ka)]'|^2\}. \end{aligned} \tag{7.8}$$

The upper sign is appropriate for $\theta^{\text{inc}} = 0$; the lower sign for $\theta^{\text{inc}} = \pi$.

The quantity $10\log_{10}[U_{\text{tot}}(r=0)/U_{\text{inc}}(r=0)]$ is plotted in Figs. 13–15 for the aperture angles 10°, 30°, and 60° (i.e., for $\theta_0 = 170°$, 150°, and 120°) and for the angle of incidence $\theta^{\text{inc}} = 0°$. Several interesting features are apparent immediately. For $\theta_0 = 170°$ the open spherical shell acts very similarly to a spherical cavity of the same size. The peaks in the data at $ka = 4.49$, 2.74, 3.87, and 4.97 closely correspond, respectively, to the lowest TE and TM modes of a closed cavity (see Ref. 36, pp. 268–271); i.e., to the lowest-order zero x_{11} of $[xj_1(x)]$ and to the zeros x'_{11}, x'_{12}, and x'_{13} of $[xj_1(x)]'$. They are slightly offset (detuned) from the closed cavity values because of the presence of the aperture. Extensions of the discussion in Sec. V for θ_0 near π and for ka small lead to the approximate coefficient expressions in this ka region,

$$A_{11}h_1(ka) \sim -\frac{\Sigma_{n=1}^{N} f_{1n}\Gamma_{n1}^{E}}{j_1(ka)}, \tag{7.9a}$$

$$B_{11}[kah_1(ka)]' \sim -\frac{\Sigma_{n=1}^{N}(3g_{1n}\Gamma_{n1}^{H}/2n+1)}{[kaj_1(ka)]'}, \tag{7.9b}$$

which readily explain the locations of the observed features. At higher ka peaks corresponding to the roots x_{np} of

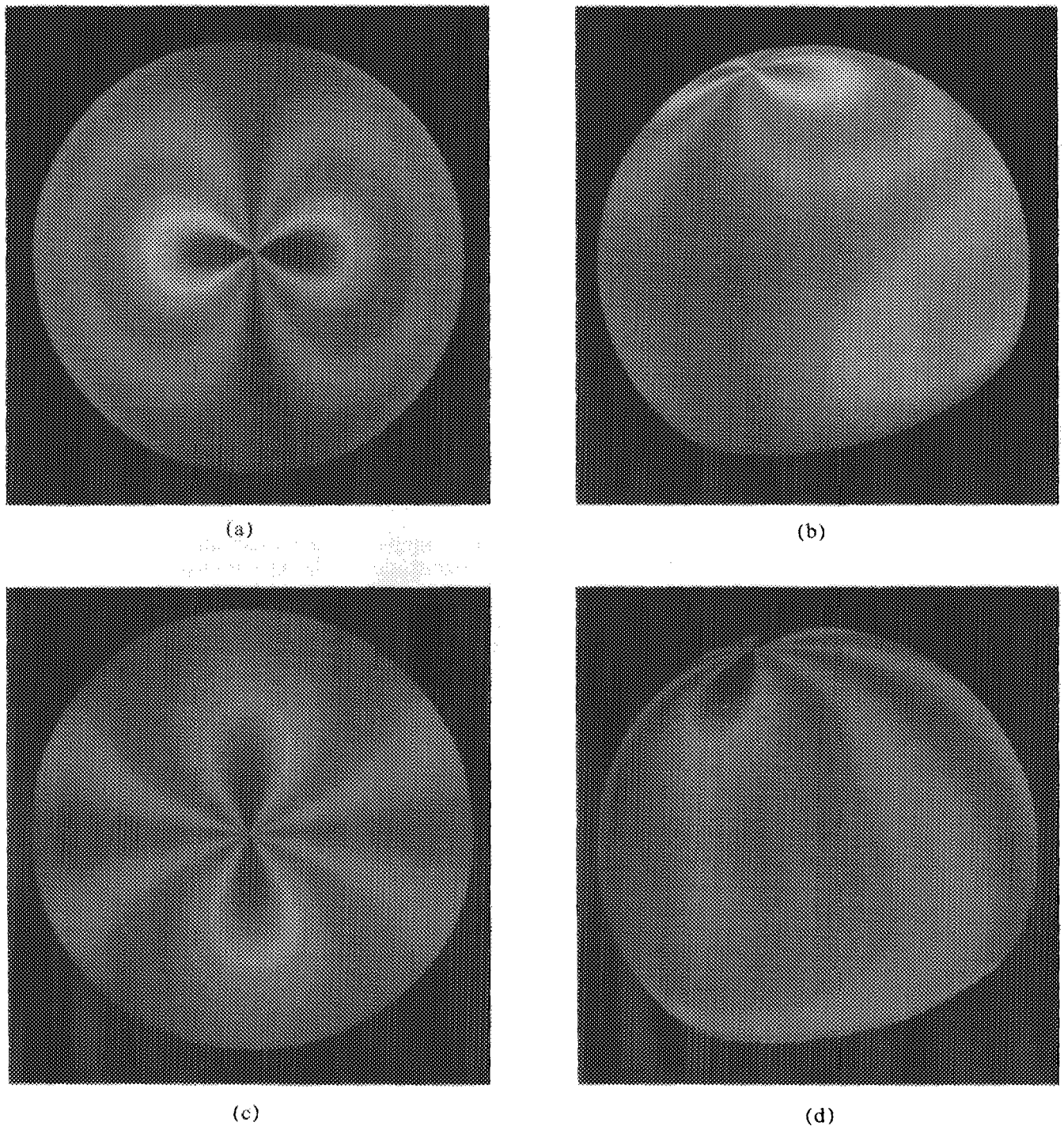

FIG. 12. The modal structure of the currents induced on an open spherical shell with $\theta_0 = 120°$ when $ka = 3.0$ and the angle of incidence $\theta^{inc} = 0°$ is revealed with three dimensional graphics: (a) a top view of the magnitude of the current term $\mathscr{J}_\theta(\theta,\phi)$; (b) a side view of the magnitude of the current term $\mathscr{J}_\theta(\theta,\phi)$; (c) a top view of the magnitude of the current term $\mathscr{J}_\phi(\theta,\phi)$; (d) a side view of the magnitude of the current term $\mathscr{J}_\phi(\theta,\phi)$.

$[xj_n(x)] = 0$ and x'_{np} of $[xj_n(x)]' = 0$ appear. The antiresonance form of the peaks at $ka = 3.87$ and 4.97 was not anticipated. In fact, only the TE_{n1} and TM_{n1} modes ($n = 1,2,...$) develop the resonance form of the peaks; all others have the antiresonant form. This behavior is a result of (1) the modal patterns induced in the open cavity—all TE_{np} and TM_{np} ($p \neq 1$) modes have nulls at $r = 0$, and (2) a reradiation effect that occurs because the aperture is backed by a resonant cavity.[40] The important features in the ka scans of (7.8) for larger apertures are associated with the modes effecting the antiresonant behavior.

Detuning of the cavity by the larger aperture is noticeable in the $\theta_0 = 150°$ data. The resonance peaks are broadened and the antiresonance peaks have become broad depressions. The resonance locations are downshifted to lower ka values (lower frequency); the antiresonance locations are

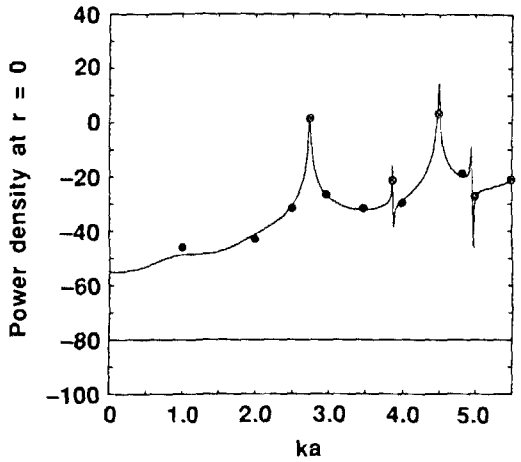

FIG. 13. A scan in ka of the total energy density of the field at the origin normalized to the energy density of the incident field there for an open spherical shell having an aperture angle of 170° and a plane wave incident at $\theta^{inc} = 0°$. The solid line is generated by the dual series solution, the dots by a MoM surface patch code.

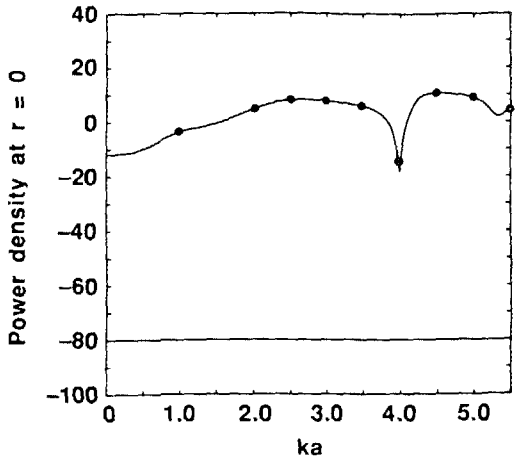

FIG. 15. A scan in ka of the total energy density of the field at the origin normalized to the energy density of the incident field there for an open spherical shell having an aperture angle of 120° and a plane wave incident at $\theta^{inc} = 0°$. The solid line is generated by the dual series solution, the dots by a MoM surface patch code.

upshifted to higher ka values (higher frequency). The depressions at the antiresonance locations indicate that this slightly open cavity may have poor energy storage characteristics, hence a large scattering cross section at those points. Energy storage and cross-section calculations are also in progress to study this (for instance see Ref. 40).

The largest aperture ($\theta_0 = 120°$) data shows nearly a complete detuning of the cavity. The observed depressions are shallower and broadened. They correspond to the original antiresonance locations $ka = 3.87$ and 4.97, thus demonstrating the considerable upshift in ka of their locations as the aperture size increases. The data also indicates a focusing of the energy near $r = 0$ over a large range of ka. This is expected since the shell is beginning to look largely like a spherical reflector when $\theta_0 = 120°$.

Comparing these results with those of Senior and Desjardins, very distinct dissimilarities are evident. Although the resonance peaks at $ka = 2.74$ and 4.49 are present in their results, the antiresonance peaks at $ka = 3.87$ and 4.97 are not. Similarly, the antiresonance depressions in the $\theta_0 = 150°$ data are absent. Also, the levels they predicted for small ka are close to 50 dB smaller in the $\theta_0 = 170°$ data and 20 dB smaller in the $\theta_0 = 150°$ data than ours. Analogously, their resonance peak levels are higher than those we predict. Our results again have been validated with a method of moments calculation.[19] Sample data points from those checks have been included in Figs. 13–15. Agreement is very good.

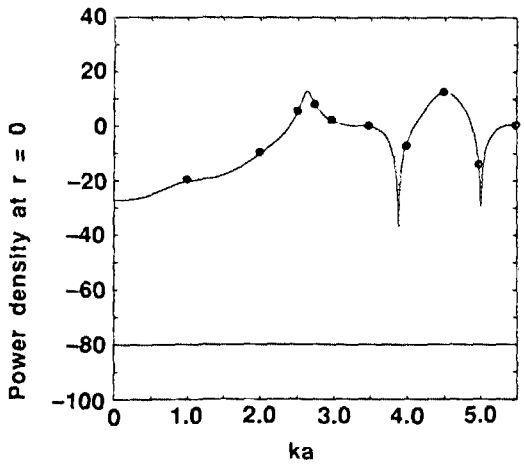

FIG. 14 A scan in ka of the total energy density of the field at the origin normalized to the energy density of the incident field there for an open spherical shell having an aperture angle of 150° and a plane wave incident at $\theta^{inc} = 0°$. The solid line is generated by the dual series solution, the dots by a MoM surface patch code.

VIII. SUMMARY

A complete solution of the scattering of a plane wave from a spherical shell having a circular aperture was developed in this paper. The angle of incidence and the polarization of the plane wave were arbitrary. This solution was constructed with a dual series equations approach and was validated in several different ways. Numerical results were given for the case of normal incidence. Induced currents on the open spherical shell were presented and it was demonstrated that they satisfy Meixner's edge conditions. Energy density scans in ka were also given; they were dominated by resonance features characteristic of the open spherical cavity.

Several new concepts and techniques were reported. The associated Legendre functions P_n^{-m} for $0 \leqslant n < m$ and their duals $\bar{P}_n^{-m}$ were introduced to produce a system of pseudodecoupled TE and TM dual series equations and to insure satisfaction of Meixner's edge conditions. Procedures were described in detail that generated an analytical solution of the resulting, previously untreated dual series systems and a numerical solution of the resulting infinite system of linear equations for the modal coefficients. Analytical preconditioning of the current sums led to results free of any Gibbs oscillations. The resonance features in the ka scans of an energy density ratio at the origin were observed to be predominantly of an antiresonant form.

Cross-section and stored energy calculations are currently in progress. Preliminary cross-section results are also dominated by antiresonance features and suggest that they

are characteristic of a cavity-backed aperture. These studies are summarized in Ref. 40.

ACKNOWLEDGMENTS

The authors would like to extend special thanks to Art Gautesen of the Department of Mathematics at Iowa State University and to Larry Warne at Sandia National Laboratories for a variety of thoughts and discussions that led to the successful conclusion of our work. The authors would also like to thank Brian Cabral of the Computer Systems Research Group at Lawrence Livermore National Laboratory for developing the graphics package that created the three-dimensional color figures.

This work was performed by the Lawrence Livermore National Laboratory and Sandia National Laboratories under the auspices of the U. S. Department of Energy under Contracts No. W-7405-ENG-48 and DE-AC04-76DP00789, respectively.

APPENDIX: ASYMPTOTIC BEHAVIOR OF THE TERMS χ_n^ϕ AND χ_n^ψ

The small argument and large index behavior of the terms χ_n^ϕ and χ_n^ψ will be developed for $n>0$. The spherical Bessel function expansions (see Ref. 37, Eqs. 10.1.2 and 10.1.3)

$$j_n(x)=\frac{(\pi/4)^{1/2}(x/2)^n}{\Gamma(n+\frac{3}{2})}\times\left[1-\frac{x^2}{2(2n+3)}+\mathcal{O}\left(\frac{x^4}{n^2}\right)\right], \tag{A1}$$

$$h_n(x)=\frac{-i}{(4\pi)^{1/2}}\left(\frac{x}{2}\right)^{-(n+1)}\Gamma\left(n+\frac{1}{2}\right)\times\left[1+\frac{x^2}{2(2n-1)}+\mathcal{O}\left(\frac{x^4}{n^2}\right)\right], \tag{A2}$$

and the identity

$$\Gamma(n+\tfrac{1}{2})=\{[(2n-1)\cdot\cdots\cdot5\cdot3\cdot1]/2^n\}\pi^{1/2} \tag{A3}$$

provide the necessary expressions. Equations (A1) and (A2) yield

$$[xj_n(x)]'=\left(\frac{\pi}{4}\right)^{1/2}\frac{(n+1)}{\Gamma(n+\frac{3}{2})}\left(\frac{x}{2}\right)^n\times\left[1-\frac{(n+3)x^2}{2(n+1)(2n+3)}+\mathcal{O}\left(\frac{x^4}{n^2}\right)\right], \tag{A4}$$

$$[xh_n(x)]'=\frac{i}{(4\pi)^{1/2}}n\Gamma\left(n+\frac{1}{2}\right)\left(\frac{x}{2}\right)^{-(n+1)}\times\left[1+\frac{(n-2)x^2}{2n(2n-1)}+\mathcal{O}\left(\frac{x^4}{n^2}\right)\right]. \tag{A5}$$

Combining (A1)–(A3) gives

$$j_n(x)h_n(x)=\frac{1}{i(2n+1)x}\times\left[1+\frac{2x^2}{(2n-1)(2n+3)}+\mathcal{O}\left(\frac{x^4}{n^4}\right)\right]; \tag{A6}$$

combining (A3)–(A5) gives

$$[xj_n(x)]'[xh_n(x)]'=\frac{n(n+1)}{-i(2n+1)x}\times\left[1-\frac{(2n^2+2n+3)x^2}{n(n+1)(2n-1)(2n+3)}+\mathcal{O}\left(\frac{x^4}{n^4}\right)\right]. \tag{A7}$$

Consequently, for small arguments

$$\lim_{x\to0}\chi_n^\phi=\lim_{x\to0}\{[i(2n+1)xj_n(x)h_n(x)]-1\}\sim0, \tag{A8}$$

$$\lim_{x\to0}\chi_n^\psi=\lim_{x\to0}\left[\left\{\frac{-i(2n+1)x}{n(n+1)}\times[xj_n(x)]'[xh_n(x)]'\right\}-1\right]\sim0, \tag{A9}$$

and for indices larger than the argument

$$\lim_{n\to\infty}\chi_n^\phi(x)=\lim_{n\to\infty}\mathcal{O}(x^2/n^2)\sim0, \tag{A10}$$

$$\lim_{n\to\infty}\chi_n^\psi(x)=\lim_{n\to\infty}\mathcal{O}(x^2/n^2)\sim0. \tag{A11}$$

Note that this limiting behavior is responsible for the number of terms required for convergence of the solution. In particular, for large enough N, the terms

$$\chi_N^\phi(ka)\sim\chi_N^\psi(ka)\sim(ka/N)^2$$

and the elements of the matrix $\mathcal{M}_{ij}$ in (4.49) are small. This explains the choice $N=10\,ka$ for the examples.

[1] A. M. Radin and V. P. Shestopalov, USSR Comp. Math. Math. Phys. **14**(5), 137 (1974).
[2] A. M. Radin and V. P. Shestopalov, Sov. Phys. Dokl. **18**, 642 (1974).
[3] A. M. Radin, V. A. Rezunenko, and V. P. Shestopalov, USSR Comp. Math. Math. Phys. **17**(2), 104 (1977).
[4] S. S. Vinogradov and V. P. Shestopalov, Sov. Phys. Dokl. **22**, 638 (1977).
[5] S. S. Vinogradov, Yu A. Tuchkin, and V. P. Shestopalov, Sov. Phys. Dokl. **23**, 650 (1978).
[6] S. S. Vinogradov, Yu A. Tuchkin, and V. P. Shestopalov, Sov. Phys. Dokl. **25**, 531 (1980).
[7] S. S. Vinogradov, Yu A. Tuchkin, and V. P. Shestopalov, Sov. Phys. Dokl. **26**, 169 (1981).
[8] S. S. Vinogradov and V. P. Shestopalov, Sov. Phys. Dokl. **26**, 314 (1981).
[9] S. S. Vinogradov, Radiophys. Quantum Electron. **26**, 78 (1983).
[10] K. F. Casey, Proceedings of the National Radio Science Meeting, June 1981, p. 69.
[11] S. Chang and T. B. A. Senior, U. S. Air Force Weapons Lab. Interaction Note 141, Albuquerque, NM, April 1969.
[12] T. B. A. Senior and G. A. Desjardins, U. S. Air Force Weapons Lab. Interaction Note 142, Alburquerque, NM, August 1973.
[13] T. B. A. Senior and G. A. Desjardins, IEEE Trans. Electromag. Compat. **EMC-16**, 205 (1974).
[14] T. B. A. Senior, U. S. Air Force Weapons Lab. Interaction Note 220, Albuquerque, NM, January 1975.
[15] R. K. Jones and T. H. Shumpert, IEEE Trans. Antennas Propagat. **AP-28**, 128 (1980).
[16] J. Meixner, Nachr. Akad. Wiss. Gottingen **2**, 74 (1946).
[17] J. Meixner, A. Naturforsch. **3a**, 506 (1948).
[18] R. W. Ziolkowski and W. A. Johnson, Radio Sci. **22**, 169 (1987).
[19] W. A. Johnson and R. W. Ziolkowski, Proceedings of the National Radio Science Meeting, June 1984, p. 174.
[20] W. A. Johnson and R. W. Ziolkowski, Radio Sci. **19**, 275 (1984).
[21] R. W. Ziolkowski, W. A. Johnson, and K. F. Casey, Radio Sci. **19**, 1425 (1984).

[22]R. W. Ziolkowski, SIAM J. Math. Anal. **16**, 358 (1985).
[23]R. W. Ziolkowski and J. B. Grant, to be published in IEEE Trans. Antennas Propaga.
[24]R. W. Ziolkowski and J. B. Grant, IEEE Trans. Microwave Theor. Tech. **MTT-34**, 1164 (1986).
[25]R. W. Ziolkowski and W. A. Johnson, UCRL-92940, Lawrence Livermore National Laboratory, Livermore, CA, July 1985.
[26]D. S. Jones, *The Theory of Electromagnetism* (Pergamon, New York, 1964).
[27]M. Born and E. Wolf, *Principles of Optics* (Macmillan, New York, 1964), 2nd ed.
[28]C. H. Wilcox, J. Math. Mech. **6**, 167 (1957).
[29]C. J. Bouwkamp and H. B. G. Casimir, Physica **20**, 539 (1954).
[30]J. H. Bruning and Y. T. Lo, IEEE Trans. Antennas Propagat. **AP-19**, 378 (1971); **AP-19**, 391 (1971).
[31]C. Liang and Y. T. Lo, Radio Sci. **2**, 1481 (1967).
[32]I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1965).
[33]T. M. MacRobert, *Spherical Harmonics* (Pergamon, New York, 1967), 3rd ed.
[34]E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics* (Chelsea, New York, 1955).
[35]L. Robin, *Fonctions Spheriques de Legendre et Fonctions Spheroidales* (Gauthier-Villars, Paris, 1958).
[36]F. Oberhettinger, *Fourier Expansions: A Collection of Formulas* (Academic, New York, 1973).
[37]*Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1970).
[38]R. F. Harrington, *Time-Harmonic Electromagnetic Fields* (McGraw-Hill, New York, 1977).
[39]J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1962).
[40]R. W. Ziolkowski, to be published in Radio Sci.

A geometric setting for internal motions of the quantum three-body system

Toshihiro Iwai
Department of Applied Mathematics and Physics, Kyoto University, Kyoto 606, Japan

(Received 15 May 1986; accepted for publication 7 January 1987)

Quantum mechanics for internal motions of the three-body system is set up on the basis of the complex vector bundle theory. The three-body system is called a triatomic molecule in the Born–Oppenheimer approximation. The internal states of the molecule are described as cross sections in the complex vector bundle assigned by an eigenvalue of the square of the total angular momentum operator. This bundle is equipped with a linear connection, which is a natural consequence of a geometric interpretation of the so-called Eckart condition. The coupling of the internal motion with the rotation is understood naturally in terms of this connection. The internal Hamiltonian operator is obtained which includes the internal motion–rotation coupling and a centrifugal potential. The complex vector bundle for the triatomic molecule proves to be a trivial bundle, though the geometric setting for the internal motion is independent of whether the bundle is trivial or not.

I. INTRODUCTION

Quantum mechanics of a few-body system has been drawing increasing interest in quantum chemistry.[1] A few-body system is called a molecule in the Born–Oppenheimer approximation. Theoretical treatment of nonrigid molecules has been related more or less with the Eckart frame.[2] Tachibana and the author[3] have discussed quantum mechanics of nonrigid molecules without using the Eckart frame, but they did not refer to the geometric setting in the large. This paper is a continuation of the previous paper[3] with particular interest in an application of the theory of connections in complex vector bundles.[4]

It is Guichardet[5] who observed that the center-of-mass system is made into a principal fiber bundle with a rotation group as structure group, and is endowed with a connection by the Eckart condition. Using the holonomy theorem,[4] he proved that the vibration cannot, in general, be separated from the rotation.

On the basis of Guichardet's observation, the present author[6] showed that a moving frame, called the Eckart frame, exists relative to which the molecule moves without rotation, but it depends on a choice of the molecular motion and is not unique for any molecular configuration. For this reason the Eckart frame is not suited for a description of quantum mechanics of nonrigid molecules.

The organization of this article is outlined in the following way: Section II deals with the center-of-mass system. It is shown that the center-of-mass system for a triatomic molecule is made into a principal SO(3) bundle $\dot{Q}$ and that the base manifold $M := \dot{Q}/\mathrm{SO}(3)$, called the internal space, is diffeomorphic to $\mathbb{R}^3_+ = \{(x, y, z) \in \mathbb{R}^3; z > 0\}$. Since the base manifold is contractible to a point, this SO(3) bundle becomes a trivial bundle[7]; $\dot{Q} \cong \mathbb{R}^3_+ \times \mathrm{SO}(3)$. Any point of $\dot{Q}$ can then be assigned uniquely by the internal coordinates in $\mathbb{R}^3_+$ and the Euler angles in SO(3). A coordinate system, called the Dragt coordinate system, is introduced on $\dot{Q}$ so that the Euler angles and the internal coordinates may assign, respectively, the principal axes of moment of inertia and the molecular configuration relative to the principal axes.

Section III is concerned with the connection due to Guichardet. The connection form and the curvature form are defined on the SO(3) bundle $\dot{Q}$ and expressed in the Dragt coordinate system. It is to be noted that even though the $\dot{Q}$ is a trivial bundle, the connection has nonvanishing curvature, i.e., the $\dot{Q}$ is not flat with respect to this connection.

Section IV is a review of the total angular momentum operator. The angular momentum operator is obtained as the infinitesimal generator of the SO(3) action on $\dot{Q}$. The right and left actions of SO(3) are strictly distinguished. The left and right actions give rise to the angular momentum operator with respect to a fixed frame and to the principal axis frame, respectively.

In Sec. V, the connection is discussed again in terms of vector fields. The connection is by definition an assignment of the vibrational (or horizontal) subspace of the tangent space at every point of $\dot{Q}$. Rotational and vibrational vectors are discussed to demonstrate that the connection theory naturally fits mechanics of several-particle systems. The connection form is in fact dual to the angular momentum.

Section VI discusses the metric and the volume element induced on the internal space. The results will be used for constructing the internal Hamiltonian operator in Sec. VIII.

In Sec. VII, the complex vector bundles V_l, l being nonnegative integers, are introduced in association with the SO(3) bundle $\pi: \dot{Q} \to M$. The linear connection and the curvature are defined in V_l and expressed in the Dragt coordinate system. Like the SO(3) bundle $\dot{Q}$, the V_l is a trivial bundle.

In Sec. VIII, quantum mechanics for internal states is established on the vector bundle V_l together with the assigned internal Hamiltonian operator H_l which acts on cross sections in V_l. The energy functional is used to derive the operator H_l from the usual Hamiltonian operator acting on wave functions on the center-of-mass system. The H_l is a matrix-valued second-order differential operator which contains both covariant derivation operators with respect to the linear connection introduced in Sec. VII and a matrix-valued centrifugal potential arising from the conservation of the total angular momentum. This section contains also remarks about internal Hamiltonian operators.

 0022-2488/87/061315-12$02.50

The result obtained can be interpreted in terms of gauge theory as follows: The rotation of the molecule induces a gauge field (the curvature introduced in Sec. VII). The internal motion is coupled with the gauge field through the gauge potential (the connection introduced in Sec. VII). The gauge field plays the role of a magnetic field on the internal space, and may be called the Coriolis field. The internal Hamiltonian describes the internal motion in coupling with the Coriolis field under the presence of the centrifugal potential.

The three-body problem in $\mathbb{R}^3$ is in marked contrast with the same problem in $\mathbb{R}^2$. In fact, in the former case, the SO(3) bundle and the associated complex vector bundles are both trivial, but in the latter case, the SO(2) bundle and the associated complex line bundles are neither trivial.[8]

II. SETTINGS IN THE CENTER-OF-MASS SYSTEM

A. An orthonormal system

Consider a three-body system in $\mathbb{R}^3$, which we call a triatomic molecule (or a molecule in short) in the Born–Oppenheimer approximation. Let y_k and m_k, $k=1,2,3$, denote the position vector and the mass of each particle, respectively. Then the configuration space Q_0 of the molecule is the linear space of all the triples $y=(y_1,y_2,y_3)$. The center-of-mass system Q is defined as the linear subspace of Q_0 defined by

$$\sum_{k=1}^{3} m_k y_k = 0 . \tag{2.1}$$

We define the inner product K on Q_0 by

$$K(x,y) = \sum m_k (x_k \mid y_k) , \tag{2.2}$$

where the round brackets denote the standard inner product in $\mathbb{R}^3$. The induced inner product in Q will be also denoted by K without any confusion.

The rotation group SO(3) has a natural action on the configuration space Q_0;

$$y=(y_1,y_2,y_3) \to gy = (gy_1, gy_2, gy_3) , \tag{2.3}$$

where $g \in SO(3)$. Since Eq. (2.1) is invariant under the SO(3) action, the SO(3) also acts on the center-of-mass system Q.

We start with the following proposition.

Proposition 1: The following system $\{c_k, f_k, f_{k+3}\}$, $k=1,2,3$, is an orthonormal system in Q_0 with respect to the inner product K:

$$\begin{aligned} c_k &= N_0(e_k,e_k,e_k) , \quad k=1,2,3 , \\ f_k &= N_1(m_3 e_k, 0, -m_1 e_k) , \quad k=1,2,3 , \\ f_{3+k} &= N_2(-m_2 e_k, (m_1+m_3)e_k, -m_2 e_k) , \quad k=1,2,3 , \end{aligned} \tag{2.4}$$

where N_j, $j=0,1,2$, are normalization constants given by

$$\begin{aligned} N_0 &= (m_1+m_2+m_3)^{-1/2} , \\ N_1 &= (m_1 m_3 (m_1+m_3))^{-1/2} , \\ N_2 &= (m_2(m_1+m_3)(m_1+m_2+m_3))^{-1/2} . \end{aligned} \tag{2.5}$$

It is a matter of calculation to verify that $\{c_k, f_k, f_{3+k}\}$, $k=1,2,3$, form an orthonormal system. The vectors $\{f_j\}$, $j=1,2,\ldots,6$, span the center-of-mass system Q, and the $\{c_k\}$, $k=1,2,3$, are the basis of $Q^{\perp}$, the orthogonal complement of Q. The space $Q^{\perp}$ is identified with the space of center-of-mass vectors. In fact, let B denote the center-of-mass vector

$$B = \sum_{k=1}^{3} m_k y_k \Big(\sum_{k=1}^{3} m_k \Big)^{-1} = \sum_{k=1}^{3} B^k e_k . \tag{2.6}$$

Then any triple (y_1,y_2,y_3) is as usual broken up into

$$(y_1,y_2,y_3) = (B,B,B) + (x_1,x_2,x_3) \tag{2.7}$$

with $\Sigma m_k x_k = 0$. It is now easy to see that

$$(B,B,B) = \sum_{k=1}^{3} \frac{B^k c_k}{N_0} , \tag{2.8}$$

which assures the above assertion. The triple $x=(x_1,x_2,x_3)$ is of course expressed in terms of $\{f_j\}$, $j=1,\ldots,6$;

$$x = \sum_{j=1}^{6} q^j f_j , \quad q^j = K(x,f_j) . \tag{2.9}$$

Thus we may consider $(B^k/N_0, q^j)$, $k=1,2,3$, $j=1,\ldots,6$, as the Cartesian coordinates in Q_0;

$$y = \sum_{k=1}^{3} \frac{B^k c_k}{N_0} + \sum_{j=1}^{6} q^j f_j . \tag{2.10}$$

It is now an easy matter to calculate $K(gf_i, f_j)$, $i,j=1,2,\ldots,6$, etc., for $g \in SO(3)$ in order to express the SO(3) action in the block diagonal form

$$\begin{pmatrix} g & & \\ & g & \\ & & g \end{pmatrix} , \quad g \in SO(3) , \tag{2.11}$$

with respect to the basis $\{c_k, f_k, f_{3+k}\}$, $k=1,2,3$, where missing matrix entries are all zero.

In view of (2.11), we are allowed to think of the center-of-mass system $Q \simeq \mathbb{R}^6$ as a product space $\mathbb{R}^3 \times \mathbb{R}^3$, the first factor spanned by f_k, and the second by f_{3+k}, $k=1,2,3$. The SO(3) action on $Q \simeq \mathbb{R}^3 \times \mathbb{R}^3$ is then a diagonal one. Moreover, the vectors $\Sigma q^k f_k$ and $\Sigma q^{3+k} f_{3+k}$ can be represented in the original space $\mathbb{R}^3$. In effect, on setting $r^k = q^k$ and $s^k = q^{3+k}$, $k=1,2,3$, to define the vectors $r = \Sigma r^k e_k$ and $s = \Sigma s^k e_k$ in $\mathbb{R}^3$, we obtain, after a straightforward calculation,

$$\begin{aligned} r &= ((m_1 m_3)/(m_1+m_3))^{1/2}(x_1 - x_3) , \\ s &= \left(\frac{m_2(m_1+m_3)}{m_1+m_2+m_3} \right)^{1/2} \left(x_2 - \frac{m_1 x_1 + m_3 x_3}{m_1+m_3} \right) . \end{aligned} \tag{2.12}$$

These are Jacobi vectors used often in the three-body problem.[9,10] Thus the points x of the center-of-mass system can be thought of as the pairs of vectors (r,s).

B. The internal space

Following Guichardet,[5] we make the center-of-mass system Q into a principal fiber bundle with structure group SO(3). To this end, we have to investigate whether the SO(3) action on Q is free or not. Suppose that $gr = r$ and $gs = s$ for some vectors r and s. If r and s are linearly independent, g must be the identity. If they are linearly dependent, g need not be the identity. Therefore setting

$$D = \{(r,s) \in \mathbb{R}^3 \times \mathbb{R}^3 ; \ \lambda r + \mu s = 0 \text{ with } (\lambda,\mu) \neq (0,0)\} ,$$

we find that the SO(3) acts freely on

$$\dot{Q} := (\mathbb{R}^3 \times \mathbb{R}^3) - D , \tag{2.13}$$

so that the quotient space $M := \dot{Q}/\mathrm{SO}(3)$ is a manifold, called the internal space. Thus we obtain a principal fiber bundle $\pi: \dot{Q} \to M$ with structure group SO(3). We note here that the excluded set D corresponds to the configurations in which three particles lie in the same line, and two or three of them may happen to collide.

Now we wish to study the topology of the internal space. For this purpose we consider the mapping

$$(r,s) \to (|r|^2 - |s|^2, 2(r|s), 2|r \times s|) . \tag{2.14}$$

Note that the quantities in the mapping image are invariant under the SO(3) action. Since the set D is excluded, $|r \times s|$ must be positive. The quantities $|r|^2 - |s|^2$ and $(r|s)$ range over all the real numbers. Hence, by (2.14), $\dot{Q}$ is mapped onto $\mathbb{R}^3_+$, the half-space of $\mathbb{R}^3$. Furthermore, it is easy to verify that

$$(|r|^2 - |s|^2)^2 + 4(r|s)^2 + 4|r \times s|^2 = (|r|^2 + |s^2|)^2 . \tag{2.15}$$

Conversely, when given a point (w^k) of $\mathbb{R}^3_+$, the inverse image of the mapping (2.14) is the set of solutions to the coupled equations $|r|^2 - |s|^2 = w^1$, $2(r|s) = w^2$, $2|r \times s| = w^3 > 0$. In solving these equations, we see from (2.15) that $|r|^2 + |s|^2$ is determined in terms of w^k, $k = 1,2,3$, so that $|r|$ and $|s|$ become known. Moreover, the angle made by two vectors r and s is determined by $2(r|s) = w^2$. Thus we can obtain a triangle formed by a certain pair of linearly independent vectors r_0 and s_0 in $\mathbb{R}^3$. The solutions are then the SO(3) orbit of r_0 and s_0. This shows that the inverse image of $(w^k) \in \mathbb{R}^3_+$ under the mapping (2.14) is topologically SO(3). Therefore we obtain the following theorem.

Theorem 2: The internal space $M = \dot{Q}/\mathrm{SO}(3)$ is topologically $\mathbb{R}^3_+$. Since $\mathbb{R}^3_+$ is contractible to a point, the principal fiber bundle $\pi: \dot{Q} \to M$ becomes a trivial bundle,[7] where π is equivalently given by (2.14).

The triviality of the bundle $\pi: \dot{Q} \to M$ implies the existence of a cross section[7] $\sigma_0: M \to \dot{Q}$, $\pi \circ \sigma_0 = \mathrm{id}$. This fact was known tacitly by Dragt,[11] and Levy-Leblond and Levy-Nahas,[12] and used effectively in analyzing the states of noninteracting three particles. According to them, r and s can be expressed in the form

$$\begin{aligned} r &= \rho(\cos(\psi/2)\cos(\chi/2)u_1 - \sin(\psi/2)\sin(\chi/2)u_2) , \\ s &= \rho(\sin(\psi/2)\cos(\chi/2)u_1 + \cos(\psi/2)\sin(\chi/2)u_2) , \end{aligned} \tag{2.16}$$

where (ρ,χ,ψ) are internal coordinates subject to

$$0 < \rho < +\infty , \quad 0 < \chi \leqslant \pi/2 , \quad 0 \leqslant \psi \leqslant 2\pi ,$$

and $u_k = g e_k$, $k = 1,2,3$, are a moving frame such that the molecule is set on the plane spanned by u_1 and u_2. We will refer to these coordinates as Dragt's coordinates. In the succeeding section we will see that the u_k are chosen so as to lie in the direction of the principal axes of moment of inertia for the molecule. It is also to be noted that the determinant of the coefficient matrix of (2.16) is $\frac{1}{2}\sin\chi$, so that $\chi \neq 0$. Now Eq. (2.16) with $u_k = e_k$, i.e., $g = \mathrm{id}$, provides a cross section σ_0: $M \to \dot{Q}$. Every element of $\dot{Q}$ is then given in the form $g\sigma_0(w)$, $g \in \mathrm{SO}(3)$, $w \in M$. Here we remark that the structure group acts on $\dot{Q}$ to the left, contrary to the usual convention according to which the structure group acts on the principal fiber bundle to the right.

From (2.16), the projection π given by (2.14) takes the form

$$\begin{aligned} w^1 &= |r|^2 - |s|^2 = \rho^2 \cos\psi \cos\chi , \\ w^2 &= 2(r|s) = \rho^2 \sin\psi \cos\chi , \\ w^3 &= 2|r \times s| = \rho^2 \sin\chi > 0 , \end{aligned} \tag{2.17}$$

so that $(\rho^2, \psi, \pi/2 - \chi)$ can be thought of as the spherical coordinates in $\mathbb{R}^3_+$.

We conclude this section by fixing the Euler angles as follows:

$$\begin{aligned} u_1 &= e_1(\cos\beta\cos\alpha\cos\gamma - \sin\alpha\sin\gamma) \\ &\quad + e_2(\cos\beta\sin\alpha\cos\gamma + \cos\alpha\sin\gamma) \\ &\quad + e_3(-\sin\beta\cos\gamma) , \\ u_2 &= e_1(-\cos\beta\cos\alpha\sin\gamma - \sin\alpha\cos\gamma) \\ &\quad + e_2(-\cos\beta\sin\alpha\sin\gamma + \cos\alpha\cos\gamma) \\ &\quad + e_3(\sin\beta\sin\gamma) , \\ u_3 &= e_1\sin\beta\cos\alpha + e_2\sin\beta\sin\alpha + e_3\cos\beta , \end{aligned} \tag{2.18}$$

where $0 \leqslant \alpha \leqslant 2\pi$, $0 \leqslant \beta \leqslant \pi$, $0 \leqslant \gamma \leqslant 2\pi$.

III. THE CONNECTION DUE TO GUICHARDET

A. A review of the connection

In Ref. 6, using the connection form due to Guichardet,[5] we have shown that there exists the Eckart frame, a frame relative to which the molecule moves without rotation, along any curve $x(t)$ in the center-of-mass system, but it depends inevitably on the choice of $x(t)$. Hence the Eckart frame is not suitable for description of molecular motions in quantum as well as classical mechanics.

The connection form is defined as follows[5,6]: Let $\Lambda^2\mathbb{R}^d$ denote the space of antisymmetric tensors of order 2 on $\mathbb{R}^d$, and so(d) the Lie algebra of SO(d). A linear isomorphism R of $\Lambda^2\mathbb{R}^d$ to so(d) is defined for $\xi = \Sigma_{i<j}\,\xi_{ij}\, e_i \wedge e_j$ and $x = \Sigma x^i e_i$ by

$$R_\xi(x) = \sum\Big(\sum \xi_{ij} x^j\Big) e_i . \tag{3.1}$$

The inertia operator $A_x: \Lambda^2\mathbb{R}^d \to \Lambda^2\mathbb{R}^d$ is defined on the center-of-mass system by

$$A_x(\xi) = -\sum m_k x_k \wedge R_\xi(x_k) , \tag{3.2}$$

which is symmetric and positive definite, and k ranges from 1 to N, the number of particles. The connection form ω on the center-of-mass system is then given by

$$\omega = R\Big(-A_x^{-1} \sum m_k x_k \wedge dx_k\Big) , \tag{3.3}$$

where $R(\xi)$ stands for R_ξ for notational convenience.

In our case, $d = 3$. If we set $\xi_{12} = \phi^3$, $\xi_{23} = \phi^1$, $\xi_{31} = \phi^2$, the two-vector $\xi = \Sigma_{i<j}\,\xi_{ij} e_i \wedge e_j$ is identified with $\phi = \Sigma \phi^i e_i$. Put another way, $\Lambda^2\mathbb{R}^3$ is identified with $\mathbb{R}^3$ by

$e_1 \wedge e_2 \to e_3$ and the cyclic permutations. Accordingly, R becomes a linear isomorphism of $\mathbb{R}^3$ to so(3);

$$R_\xi(x) = R_\phi(x) = -\phi \times x, \quad \text{for } x \in \mathbb{R}^3.$$

Alternatively, $R(e_1)$ is the matrix (ξ_{ij}) with nonzero elements $\xi_{23} = -\xi_{32} = 1$ only, and so on. It should be noted that R is Ad-equivariant in the sense that

$$R(g\phi) = \mathrm{Ad}_g R(\phi) = gR(\phi)g^{-1}. \tag{3.4}$$

With the above identification in mind, we treat A_x and ω in the form

$$A_x(\phi) = \sum m_k x_k \times (\phi \times x_k), \tag{3.5}$$

$$\omega = R\left(-A_x^{-1} \sum m_k x_k \times dx_k\right). \tag{3.6}$$

We note again that A_x is symmetric and positive definite, because

$$(\phi' | A_x(\phi)) = \sum m_k (\phi' \times x_k | \phi \times x_k). \tag{3.7}$$

Moreover, using (3.4), we can verify that

$$A_{gx}(\phi) = gA_x(g^{-1}\phi), \tag{3.8}$$

$$\lambda_g^* \omega = \mathrm{Ad}_g \omega = g\omega g^{-1}, \tag{3.9}$$

where λ_g^* denotes the pullback by the g action; $\lambda_g(x) = gx$.

A vector field $v = (v_k)$, $k = 1,2,3$, on $\dot{Q}$, which is subject to $\Sigma m_k v_k = 0$, is called vibrational if it satisfies $\omega(v) = 0$. From (3.6) this condition becomes equivalent to

$$\sum m_k x_k \times v_k = 0, \tag{3.10}$$

because $dx_k(v) = v_k$, and R and A_x are nonsingular. Equation (3.10) means that the total angular momentum vanishes for the vibrational vector fields. In this sense Eq. (3.10) is viewed as a generalization of the Eckart condition to any configuration of the molecule. Rotational vector fields are defined as infinitesimal generators of the SO(3) action on $\dot{Q}$, which are given by $R_\phi(x) := (R_\phi(x_k))$, $k = 1,2,3$. We notice that for rotational vector fields $R_\phi(x)$ one has

$$\omega(R_\phi(x)) = R\left(-A_x^{-1} \sum m_k x_k \times (-\phi \times x_k)\right)$$
$$= R(A_x^{-1} A_x(\phi)) = R_\phi. \tag{3.11}$$

Equations (3.9) and (3.11) are characteristic of the connection form.[4] We notice here that because of the left action Eq. (3.9) is expressed in a form different from the usual one. In what follows we will express A_x and ω in terms of r and s and of Dragt's coordinates.

B. The inertia operator

We start by calculating $\Sigma m_k x_k \times dx_k$. Since r and s are given by (2.12), the vectors x_k, $k = 1,2,3$, with $\Sigma m_k x_k = 0$, are expressed as linear combinations of r and s. Consequently, a straightforward calculation gives

$$\sum m_k x_k \times dx_k = r \times dr + s \times ds. \tag{3.12}$$

Applying (3.12) to a rotational vector $R_\phi(x)$, one obtains

$$\sum m_k x_k \times R_\phi(x_k) = r \times R_\phi(r) + s \times R_\phi(s). \tag{3.13}$$

We note here that since SO(3) acts on $\dot{Q}$ in the form (gr, gs), rotational vector fields are expressed as $(R_\phi(r), R_\phi(s))$.

From (3.5) and (3.13) it follows that

$$A_x(\phi) = r \times (\phi \times r) + s \times (\phi \times s). \tag{3.14}$$

The matrix elements $(e_j | A_x(e_k))$ of A_x are then easy to calculate in terms of r and s:

$$\begin{aligned} (e_j | A_x(e_k)) &= -(r^j r^k + s^j s^k) \quad (j \neq k), \\ (e_k | A_k(e_k)) &= |r|^2 + |s|^2 - (r^k)^2 - (s^k)^2. \end{aligned} \tag{3.15}$$

Applying (3.14) to $u_k = ge_k$, $k = 1,2,3$, together with (2.16), we obtain

$$\begin{aligned} A_x(u_1) &= \rho^2 \sin^2(\chi/2) u_1, \\ A_x(u_2) &= \rho^2 \cos^2(\chi/2) u_2, \\ A_x(u_3) &= \rho^2 u_3. \end{aligned} \tag{3.16}$$

These equations show that the u_k lie in the direction of the principal axes, and at the same time give the principal moments of inertia.

C. The connection form

The connection form (3.6) is now given from (3.12) as

$$\omega = R(-A_x^{-1}(r \times dr + s \times ds)), \tag{3.17}$$

where A_x^{-1} is the inverse of the matrix given in (3.15).

In what follows we are going to calculate ω in Dragt's coordinates. For this purpose, the following formula, easy to prove, is of great help.

Proposition 3: The frame (u_k), $k = 1,2,3$, satisfies the relation

$$(du_1, du_2, du_3) = (u_1, u_2, u_3) \begin{pmatrix} 0 & -\Theta^3 & \Theta^2 \\ \Theta^3 & 0 & -\Theta^1 \\ -\Theta^2 & \Theta^1 & 0 \end{pmatrix}, \tag{3.18}$$

where Θ^k, $k = 1,2,3$, are left-invariant one-forms on SO(3) which are expressed as

$$\begin{aligned} \Theta^1 &= \sin\gamma\, d\beta - \sin\beta \cos\gamma\, d\alpha, \\ \Theta^2 &= \cos\gamma\, d\beta + \sin\beta \sin\gamma\, d\alpha, \\ \Theta^3 &= \cos\beta\, d\alpha + d\gamma. \end{aligned} \tag{3.19}$$

Remark: When we set $g = (u_k)$ and $\Theta = -\Sigma R(e_k)\Theta^k$, a matrix-valued one-form, Eq. (3.18) is written as $dg = g\Theta$, or $\Theta = g^{-1}\, dg$.

Proposition 4: The right-invariant one-forms Ψ^j, $j = 1,2,3$, are given by $\Psi^j = \Sigma g_{jk} \Theta^k$, g_{jk} being the components of g. From Eqs. (2.18) and (3.19) one has

$$\begin{aligned} \Psi^1 &= -\sin\alpha\, d\beta + \sin\beta \cos\alpha\, d\gamma, \\ \Psi^2 &= \cos\alpha\, d\beta + \sin\beta \sin\alpha\, d\gamma, \\ \Psi^3 &= d\alpha + \cos\beta\, d\gamma. \end{aligned} \tag{3.20}$$

For the proof we use (3.4). Since the right-invariant matrix one-form is defined as $\Psi = dg\, g^{-1} = g\Theta g^{-1}$, one obtains

$$g\Theta g^{-1} = -g \sum R(e_k \Theta^k) g^{-1}$$
$$= -\sum R(ge_k \Theta^k) = -\sum R\left(\sum g_{jk} e_j \Theta^k\right).$$

Thus $\Psi = -\Sigma R(e_j)\Psi^j$ with $\Psi^j = \Sigma g_{jk}\Theta^k$. In the course of the calculation, we have also obtained

$$\Psi = -\sum R(e_j)\Psi^j = -\sum R(u_k)\Theta^k . \tag{3.21}$$

Now we proceed to the calculation of ω. Using (2.16) and (3.18), we are ready to compute $r\times dr + s\times ds$ in a straightforward manner to obtain

$$r\times dr + s\times ds = \rho^2 \sin^2(\chi/2)u_1\Theta^1 + \rho^2\cos^2(\chi/2)u_2\Theta^2 + \rho^2 u_3(\Theta^3 - \tfrac{1}{2}\sin\chi\, d\psi) . \tag{3.22}$$

Since A_x^{-1} is given from (3.16), the connection form (3.17) turns out to have the form

$$\omega = \Psi + R(u_3)\tfrac{1}{2}\sin\chi\, d\psi . \tag{3.23}$$

We denote the components of ω as follows:

$$\omega = \sum R(u_k)\sigma^k = \sum R(e_k)\omega^k . \tag{3.24}$$

On account of Eqs. (3.21) and (3.23), one has

$$\begin{aligned} \sigma^1 &= -\Theta^1 , \\ \sigma^2 &= -\Theta^2 , \\ \sigma^3 &= -\Theta^3 + \tfrac{1}{2}\sin\chi\, d\psi , \end{aligned} \tag{3.25}$$

$$\begin{aligned} \omega^1 &= -\Psi^1 + \tfrac{1}{2}\sin\beta\cos\alpha\sin\chi\, d\psi , \\ \omega^2 &= -\Psi^2 + \tfrac{1}{2}\sin\beta\sin\alpha\sin\chi\, d\psi , \\ \omega^3 &= -\Psi^3 + \tfrac{1}{2}\cos\beta\sin\chi\, d\psi . \end{aligned} \tag{3.26}$$

In summary, we have the following.

Theorem 5: The connnection form ω takes the form (3.23) or (3.24) with (3.25) and (3.26) in Dragt's coordinates (2.16).

Since the principal fiber bundle π: $\dot{Q}\to M$ is trivial (Theorem 2), the connection form ω is pulled back on the internal space M through the cross section σ_0: $M\to\dot{Q}$. In fact, setting $\alpha=\beta=\gamma=0$ in (3.23), we obtain

$$\sigma_0^*\omega = R(e_3)\tfrac{1}{2}\sin\chi\, d\psi . \tag{3.27}$$

D. The curvature form

The curvature form Ω is given by the structure equation

$$\Omega = d\omega - \omega\wedge\omega . \tag{3.28}$$

Note here that SO(3) acts on $\dot{Q}$ to the left, so that the structure equation takes a different form from the usual one.[4] We wish to express the Ω in Dragt's coordinates, as we have done for ω. To this end, the following propositions are of great use.

Proposition 6: The left- and right-invariant one-forms Θ and Ψ on SO(3) are subject to the following Maurer–Cartan equation,[13] respectively:

$$d\Theta = -\Theta\wedge\Theta , \quad d\Psi = \Psi\wedge\Psi . \tag{3.29}$$

Proposition 7: The matrices $R(u_k)$, $k=1,2,3$, satisfy the commutation relations

$$[R(u_1),R(u_2)] = -R(u_3) \quad \text{(cycl.)} . \tag{3.30}$$

To prove these equations, we note that $[R(e_1), R(e_2)] = -R(e_3)$ (cycl.), which are easy to see. On using (3.4) with $ge_k = u_k$, Eq. (3.30) proves to be a consequence of these commutation relations.

We are now in a position to write out (3.28) in Dragt's coordinates. From (3.23) we obtain

$$d\omega = d\Psi + R(u_3)\tfrac{1}{2}\cos\chi\, d\chi\wedge d\psi + (R(u_1)\Theta^2 - R(u_2)\Theta^1)\wedge\tfrac{1}{2}\sin\chi\, d\psi ,$$

where we have used (3.18) to get $R(du_3) = R(u_1)\Theta^2 - R(u_2)\Theta^1$. On the other hand, using (3.21) and (3.30) to get

$$R(u_3)\Psi = \Psi R(u_3) + R(u_2)\Theta^1 - R(u_1)\Theta^2 ,$$

one obtains

$$\omega\wedge\omega = \Psi\wedge\Psi + (R(u_1)\Theta^2 - R(u_2)\Theta^1)\wedge\tfrac{1}{2}\sin\chi\, d\psi .$$

Thus the curvature form (3.28) turns out to be

$$\Omega = R(u_3)\tfrac{1}{2}\cos\chi\, d\chi\wedge d\psi . \tag{3.31}$$

Theorem 8: The curvature form Ω is expressed as (3.31) in Dragt's coordinates (2.16).

The Ω is pulled back on the internal space through the cross section σ_0 to give

$$\sigma_0^*\Omega = R(e_3)\tfrac{1}{2}\cos\chi\, d\chi\wedge d\psi . \tag{3.32}$$

From (3.27) and (3.32) we have $\sigma_0^*\Omega = d\sigma_0^*\omega$.

IV. THE TOTAL ANGULAR MOMENTUM OPERATOR

The total angular momentum operator is defined through the infinitesimal generator of the SO(3) action. The operator $\hat{J}_j$ will be defined as i times the infinitesimal generator J_j of the action of $\exp tR(e_j)$, $j=1,2,3$. While we have treated the left action of SO(3), the right action is also of great importance in dealing with the total angular momentum operator. Recall Eq. (2.16); r and s are assigned by the moving frame u_k and the internal coordinates. The left action of $h\in$SO(3) maps $u_k = \Sigma u_k^l e_l$ to $u_k' = hu_k = \Sigma(\Sigma h_{jl}u_k^l)e_j$ with the Euler angles (α',β',γ'), keeping the internal coordinates fixed. This means that we are observing the rotation of the molecule in the fixed fame $\{e_k\}$. Contrary to this, under the right action of h, the u_k are mapped to $\Sigma u_j h_{jk}$, while the internal coordinates are left invariant. This implies that we are observing the rotation of the molecule in the moving frame $\{u_k\}$. In effect, the right action is related to a left action as follows: Let the matrix g be viewed as a set of the column vectors $u_k = ge_k$. Then under the left action of $\exp tR(u_j)$, the moving frame $g=(u_k)$ is mapped as a whole to $(\exp tR(u_j))g$, which are put into, by using (3.4),

$$(\exp tR(u_j))g = g(\exp tR(e_j))g^{-1}g = g(\exp tR(e_j)) . \tag{4.1}$$

Thus the left action of $\exp tR(u_j)$, the rotation of the molecule around the u_j axis, becomes equivalent to the right action of $\exp tR(e_j)$ on the frame $g=(u_k)$.

Though the total angular momentum operator is already well known, we wish to derive them in order to realize the difference between right and left actions.

A. The left action

We start with the left SO(3) action on a triple $x\in Q$. Let $h(t) = \exp tR(\phi)$ be a one-parameter subgroup of SO(3).

Then its infinitesimal generator is given by

$$\sum\left(R(\phi)x_k\left|\frac{\partial}{\partial x_k}\right.\right)=\sum\left(-\phi\times x_k\left|\frac{\partial}{\partial x_k}\right.\right)$$
$$=\left(-\phi\left|\sum x_k\times\frac{\partial}{\partial x_k}\right.\right), \tag{4.2}$$

where $\partial/\partial x_k$, $k=1,2,3$, denote the gradient vectors. Let

$$J=\sum x_k\times\frac{\partial}{\partial x_k}=-\sum e_jJ_j\,. \tag{4.3}$$

Then the infinitesimal generator (4.2) is written as $(-\phi|J)$. Since $J_j=(-e_j|J)$, each J_j turns out to be the infinitesimal generator of $\exp tR(e_j)$, $j=1,2,3$.

We turn to the expression of J in r and s. Since the action of $h(t)$ has the form $(h(t)r,h(t)s)$, in the same manner as (4.2) we have

$$J=r\times\frac{\partial}{\partial r}+s\times\frac{\partial}{\partial s}=-\sum e_jJ_j\,, \tag{4.4}$$

where $J_j=(-e_j|J)$ is expressed in r and s.

Following the same procedure as the above, we can express J_j, $j=1,2,3$, in the Euler angles. For any r and s, we set $r(t)=h(t)r$, $s(t)=h(t)s$, and $u_k(t)=h(t)u_k$, where $u_k(t)$ has the Euler angles $(\alpha(t),\beta(t),\gamma(t))$ with $\alpha(0)=\alpha$, $\beta(0)=\beta$, $\gamma(0)=\gamma$. Then, making use of (3.21), we have

$$\frac{dr(t)}{dt}=\left(\sum\Theta^k\left(\frac{d}{dt}\right)u_k\right)\times r(t)\,, \tag{4.5}$$

and the same equation for $s(t)$, where d/dt stands for the tangent vector to the curve $r(t)$. In fact

$$\frac{dr(t)}{dt}=\frac{dh(t)}{dt}h(t)^{-1}r(t)=\Psi\left(\frac{d}{dt}\right)r(t)$$
$$=-\sum\Theta^k\left(\frac{d}{dt}\right)R(u_k)r(t)$$
$$=\left(\sum\Theta_k\left(\frac{d}{dt}\right)u_k\right)\times r(t)\,.$$

The same calculation is good for $s(t)$.

Let $h(t)=\exp tR(e_j)$ for a fixed j. Then $(dr/dt)(0)=R(e_j)r=-e_j\times r$, and $(ds/dt)(0)=R(e_j)s=-e_j\times s$, so that we have

$$-e_j=\sum\Theta^k\left(\frac{d}{dt}\right)u_k(0)\,.$$

For each j, this is the equation which determines $(\dot\alpha(0),\dot\beta(0),\dot\gamma(0))$ and hence J_j. A straightforward calculation then results in

$$J_1=\cos\alpha\cot\beta\frac{\partial}{\partial\alpha}+\sin\alpha\frac{\partial}{\partial\beta}-\frac{\cos\alpha}{\sin\beta}\frac{\partial}{\partial\gamma}\,,$$
$$J_2=\sin\alpha\cot\beta\frac{\partial}{\partial\alpha}-\cos\alpha\frac{\partial}{\partial\beta}-\frac{\sin\alpha}{\sin\beta}\frac{\partial}{\partial\gamma}\,, \tag{4.6}$$
$$J_3=-\frac{\partial}{\partial\alpha}\,.$$

The J_j, $j=1,2,3$, are known as the right-invariant vector fields on SO(3), which satisfy, for the right-invariant one-forms (3.20),

$$\Psi^j(J_k)=-\delta^j{}_k\,. \tag{4.7}$$

The total angular momentum operators are defined as $\hat J_k=iJ_k$ to satisfy

$$[\hat J_1,\hat J_2]=i\hat J_3\quad\text{(cycl.)}\,. \tag{4.8}$$

B. The right action

The right action of $h(t)$ is expressed as $u_k(t)=\Sigma\,u_jh_{jk}(t)$. Differentiation gives $\dot u_k(0)=\Sigma\,u_j\dot h_{jk}(0)$. On the other hand, Eq. (3.18) provides

$$(\dot u_1(0),\dot u_2(0),\dot u_3(0))=(u_1,u_2,u_3)\Theta\left(\frac{d}{dt}\right)\Big|_{t=0}\,.$$

Hence, for each $h(t)=\exp tR(e_j)$, this equation reads

$$\left(\sum u_l\dot h_{lk}(0)\right)=(u_k)\Theta\left(\frac{d}{dt}\right)\Big|_{t=0}\,.$$

This determines $(\dot\alpha(0),\dot\beta(0),\dot\gamma(0))$, and therefore L_j, the infinitesimal generator of the right action of $\exp tR(e_j)$. To obtain L_j is now a matter of calculation:

$$L_1=\frac{\cos\gamma}{\sin\beta}\frac{\partial}{\partial\alpha}-\sin\gamma\frac{\partial}{\partial\beta}-\cos\gamma\cot\beta\frac{\partial}{\partial\gamma}\,,$$
$$L_2=-\frac{\sin\gamma}{\sin\beta}\frac{\partial}{\partial\alpha}-\cos\gamma\frac{\partial}{\partial\beta}+\sin\gamma\cot\beta\frac{\partial}{\partial\gamma}\,, \tag{4.9}$$
$$L_3=-\frac{\partial}{\partial\gamma}\,.$$

These are known as the left-invariant vector fields on SO(3), which satisfy, for the left-invariant one-forms (3.17),

$$\Theta^k(L_j)=-\delta^k{}_j\,. \tag{4.10}$$

The total angular momentum operators are defined as $\hat L_k=-iL_k$ to satisfy

$$[\hat L_1,\hat L_2]=i\hat L_3\quad\text{(cycl.)}\,. \tag{4.11}$$

It is worth mentioning that J_k and L_k are related by

$$\sum_{k=1}^{3}g_{jk}L_k=J_j\,,\quad j=1,2,3\,, \tag{4.12}$$

as the right- and left-invariant one-forms are related in the same fashion (Proposition 4). From (4.3) and (4.12) it follows that

$$J=-\sum e_jJ_j=-\sum u_kL_k\,. \tag{4.13}$$

This equation is consistent with Eq. (4.1). In fact, from Eq. (4.1) the infinitesimal generator L_k of the right action of $\exp tR(e_j)$ must be the infinitesimal generator of the left action of $\exp tR(u_k)$. Hence the L_k must equal $(-u_k|J)$ on account of (4.2) and (4.3). However, this is also a consequence of (4.13). We note in conclusion that in a different manner the operators J_k and L_k are derived in Ref. 14 in the coordinates $(\psi,\theta,\phi)=(-\alpha,-\beta,-\gamma)$.

In summary we have the following.

Proposition 9: The total angular momentum operator $\hat J$ is expressed as

$$\hat J=-i\sum x_k\times\frac{\partial}{\partial x_k}=-i\left(r\times\frac{\partial}{\partial r}+s\times\frac{\partial}{\partial s}\right)$$
$$=\sum e_j\hat J_j=-\sum u_k\hat L_k\,, \tag{4.14}$$

where $\hat{J}_j = iJ_j$ and $\hat{L}_k = -iL_k$ are given by (4.6) and (4.9), respectively, in the Euler angles.

V. ROTATIONS AND VIBRATIONS

A. Rotational and vibrational vectors

We have introduced the connection in Sec. III in terms of differential one-forms, and defined vibrational vector fields as those vector fields for which the connection form vanishes. Rotational vector fields are the infinitesimal generators of the SO(3) action. In this section, we review again the connection in terms of vector fields. Let K_x denote the inner product induced from K into the tangent space $T_x(\dot{Q})$;

$$K_x(u,v) = \sum m_k (u_k | v_k) . \tag{5.1}$$

We show that the vibrational vector fields are orthogonal to the rotational vector fields with respect to K_x. In fact, for a vibrational vector field v, subject to (3.10), and a rotational vector field $R_\phi(x)$, we have

$$\begin{aligned} K_x(v,R_\phi(x)) &= \sum m_k (v_k | -\phi \times x_k) \\ &= \left(\sum m_k v_k \times x_k | \phi \right) = 0 . \end{aligned} \tag{5.2}$$

Thus every tangent space to $\dot{Q}$ is decomposed into an orthogonal direct sum of the rotational and vibrational subspaces. This is indeed the alternative definition of the connection that Guichardet[5] first gave.

Let P_x denote the projection such that $P_x(v)$, $v \in T_x(\dot{Q})$, is the rotational component of v. Then we have, for $v = \Sigma (v_k | \partial / \partial x_k)$,

$$P_x(v)_j = \left(A_x^{-1} \sum m_k x_k \times v_k \right) \times x_j , \quad j = 1,2,3 . \tag{5.3}$$

Because the complement $v - P_x(v)$ is vibrational;

$$\sum m_j x_j \times \left(v_j - \left(A_x^{-1} \sum m_k x_k \times v_k \right) \times x_j \right) = 0 .$$

The rotational vector fields J_j and L_j have of course the rotational components only. Indeed, for

$$\begin{aligned} J_j &= \sum \left(-e_j \times x_k \left| \frac{\partial}{\partial x_k} \right. \right) , \\ L_j &= \sum \left(-u_j \times x_k \left| \frac{\partial}{\partial x_k} \right. \right) , \end{aligned} \tag{5.4}$$

we have, from (5.3) and (5.4), $P_x(J_j)_k = -e_j \times x_k$ and $P_x(L_j)_k = -u_j \times x_k$.

It is of great importance to recognize that the rotational vector field is dual to the connection form. Ineed, from (3.5), (3.6), (3.24), and (5.4), it follows that

$$\omega^k(J_j) = \delta^k{}_j , \quad \sigma^k(L_j) = \delta^k{}_j . \tag{5.5}$$

Each of these equations is a specialization of Eq. (3.11). These are also verified by (3.25), (3.26), (4.7), and (4.10).

We now deal with vibrational vector fields. For a vector field X on the internal space M, its vibrational (or horizontal) lift[4,5] is defined as the vibrational vector field on $\dot{Q}$ which projects to X; $\pi_* X^*_x = X_{\pi(x)}$, where π_* is the tangent map of π: $\dot{Q} \to M$. Alternatively, the X^* is determined by $\omega(X^*) = 0$ and $\pi_* X^* = X$. We are going to obtain vibrational lifts in Dragt's coordinates. For $\partial / \partial \zeta^j$, $(\zeta^j) = (\rho, \chi, \psi)$, their vibrational lifts $(\partial / \partial \zeta^j)^*$ are found, by using (5.5), to be

$$\begin{aligned} \left(\frac{\partial}{\partial \rho} \right)^* &= \frac{\partial}{\partial \rho} , \quad \left(\frac{\partial}{\partial \chi} \right)^* = \frac{\partial}{\partial \chi} , \\ \left(\frac{\partial}{\partial \psi} \right)^* &= \frac{\partial}{\partial \psi} - \frac{1}{2} \sin \chi \, L_3 . \end{aligned} \tag{5.6}$$

In summary, the J_j (or L_j) and $(\partial / \partial \zeta^j)^*$ constitute a basis of the space of vector fields on $\dot{Q}$.

B. Rotational and vibrational covectors

In Sec. V A we have seen that the connection is an assignment of the vibrational subspace of the tangent space at every point of $\dot{Q}$. In this section we deal again with the connection in the cotangent space. Let $T^*_x(\dot{Q})$ denote the cotangent space to $\dot{Q}$ at x. Then the inner product K_x provides an isomorphism K^b_x of $T_x(\dot{Q})$ to $T^*_x(\dot{Q})$;

$$K^b_x(u) \cdot v := K_x(u,v) , \quad u,v \in T_x(\dot{Q}) . \tag{5.7}$$

Let $K^b_x(u) = p = (p_k)$. Then from (5.7) one has $p_k = m_k u_k$. Further, every cotangent space is equipped with the inner product K^*_x; for $p,q \in T^*_x(\dot{Q})$, K^*_x is defined through $K^\#_x = (K^b_x)^{-1}$ as

$$K^*_x(p,q) := K_x(K^\#_x(p), K^\#_x(q)) = \sum \frac{(p_k | q_k)}{m_k} . \tag{5.8}$$

An assignment of the vibrational subspace of the cotangent space is made as follows: With the infinitesimal generator $R_\phi(x)$ of the SO(3) action, we can associate the rotational covector $K^b_x(R_\phi(x)) = (m_k R_\phi(x_k))$, a triple. Then, like vibrational vectors in the tangent space, vibrational covectors are defined as those which are orthogonal to rotational covectors with respect to K^*_x. For a covector $p = (p_k)$, the orthogonality condition takes the form

$$\sum x_k \times p_k = 0 , \tag{5.9}$$

which is similar to Eq. (3.10), and interpreted as a generalization of the Eckart condition.

Now, in a dual manner to (5.3), we can obtain the rotational component $P^*_x(p)$ of a covector p as follows:

$$P^*_x(p)_j = \left(A_x^{-1} \sum x_k \times p_k \right) \times m_j x_j . \tag{5.10}$$

For the proof, we have only to show that $p - P^*_x(p)$ is vibrational;

$$\sum x_k \times \left(p_k - \left(A_x^{-1} \sum x_j \times p_j \right) \times m_k x_k \right) = 0 .$$

This is, however, easy to verify.

The components of the connection form, ω^j or σ^j, have indeed the rotational components only. To see this, we note that ω^j and σ^j are expressed as

$$\omega^j = \left(-e_j | A_x^{-1} \sum m_k x_k \times dx_k \right) , \quad j = 1,2,3 , \tag{5.11}$$

$$\sigma^j = \left(-u_j | A_x^{-1} \sum m_k x_k \times dx_k \right) , \quad j = 1,2,3 . \tag{5.12}$$

These are easy consequences of (3.6) and (3.24). Because of these expressions of ω^j and σ^j, we have dealt with the connection form ω in the vector-valued form $\omega = A_x^{-1}\Sigma m_k x_k \times dx_k$ in Ref. 3. Equations (5.11) and (5.12) are put into the form

$$\omega^j = \sum (-A_x^{-1}(e_j)\times m_k x_k | dx_k), \tag{5.13}$$

$$\sigma^j = \sum (-A_x^{-1}(u_j)\times m_k x_k | dx_k), \tag{5.14}$$

respectively. Thinking of ω^j and σ^j as covectors and applying (5.10), we have $P_x^*(\omega^j)_k = -A_x^{-1}(e_j)\times m_k x_k$ and $P_x^*(\sigma^j) = -A_x^{-1}(u_j)\times m_k x_k$.

We turn to vibrational one-forms. On account of (5.9), a one-form $\nu = \Sigma(p_k|dx_k)$ is vibrational if and only if it vanishes for any rotational vector $R_\phi(x)$;

$$\nu(R_\phi(x)) = \sum(p_k| -\phi\times x_k) = -\left(\sum x_k\times p_k|\phi\right) = 0.$$

For $d\xi^j$, $(\xi^j) = (\rho,\chi,\psi)$, the pullbacks $\pi^* d\xi^j$ are vibrational, as is easily seen. In summary, we have found the basis ω^j (or σ^j) and $d\xi^j$, viewed as one-forms on $\dot{Q}$, in the space of one-forms on $\dot{Q}$. This basis is dual to the basis J_j (or L_j) and $(\partial/\partial\xi^j)^*$ obtained in the last section.

Proposition 10: The one-forms ω^j (or σ^j) and $d\xi^j$ and vector fields J_j (or L_j) and $(\partial/\partial\xi^j)^*$ constitute dual bases on $\dot{Q}$, where ω^j and σ^j are the components of the connection form ω [see (3.24)], and J_j and L_j the components of the infinitesimal rotational vector J [see (4.13)], respectively, and $(\partial/\partial\xi^j)^*$ are the horizontal lifts of $\partial/\partial\xi^j$, $(\xi^j) = (\rho,\chi,\psi)$.

Remark: In a local sense, this proposition holds true for any internal coordinates.

VI. THE METRIC AND THE VOLUME ELEMENT ON THE INTERNAL SPACE

A. The metric

We have defined the inner product K on Q, which induces the metric tensor K_x given in (5.1). In a symbolic way, we can write K_x as

$$K_x = \sum m_k (dx_k|dx_k) = (dr|dr) + (ds|ds). \tag{6.1}$$

Here the last equality holds because of the fact the (r^k) and (s^k) are the components of $x\in Q$ with respect to the orthonormal system $\{f_k, f_{3+k}\}$, $k = 1,2,3$. Our purpose in this section is to express K_x in terms of ω^j (or σ^j) and $d\xi^j$, the basis obtained in the last section.

Using (5.4) and (3.7), we obtain, for J_j and L_j,

$$K_x(J_j,J_l) = \sum m_k (e_j\times x_k | e_l\times x_k) = (e_j|A_x(e_l)), \tag{6.2}$$

and similarly

$$K_x(L_j,L_l) = (u_j|A_x(u_l)). \tag{6.3}$$

We turn to $K_x((\partial/\partial\xi^i)^*, (\partial/\partial\xi^j)^*)$, $(\xi^j) = (\rho,\chi,\psi)$. A metric tensor (b_{ij}) is defined on the internal space M by

$$b_{ij} = B_{\pi(x)}\left(\frac{\partial}{\partial\xi^i}, \frac{\partial}{\partial\xi^j}\right) := K_x\left(\left(\frac{\partial}{\partial\xi^i}\right)^*, \left(\frac{\partial}{\partial\xi^j}\right)^*\right). \tag{6.4}$$

The b_{ij} are well defined, that is, the right-hand side of (6.4) is independent of the x chosen, because every vibrational subspace at x in the same SO(3) orbit is isomorphic with the tangent space to M at $\pi(x)$, and because K_x is SO(3) invariant. Thus on account of the orthogonality of J_j (or L_j) and $(\partial/\partial\xi^j)^*$, and of the duality (Proposition 10), we have

$$K_x = \sum (e_i|A_x(e_j))\omega^i\omega^j + \sum b_{ij}\, d\xi^i\, d\xi^j = \sum (u_i|A_x(u_j))\sigma^i\sigma^j + \sum b_{ij}\, d\xi^i\, d\xi^j. \tag{6.5}$$

In order to get (b_{ij}) in an explicit form, we refer to Eq. (6.1). Using (2.16), (3.18), and (3.25), we can work out K_x to get

$$\begin{aligned}(dr|dr) + (ds|ds) &= d\rho^2 + \tfrac{1}{4}\rho^2(d\chi^2 + \cos^2\chi\, d\psi^2) + \rho^2\sin^2(\chi/2)(\sigma^1)^2 \\ &\quad + \rho^2\cos^2(\chi/2)(\sigma^2)^2 + \rho^2(\sigma^3)^2.\end{aligned} \tag{6.6}$$

From (6.5) and (6.6) together with (3.16), the induced metric on the internal space turns out to be

$$B_{\pi(x)} = d\rho^2 + \tfrac{1}{4}\rho^2(d\chi^2 + \cos^2\chi\, d\psi^2). \tag{6.7}$$

If we set $\rho^2 = \tau$ and $\chi = \pi/2 - \theta$, Eq. (6.7) goes over into

$$B_{\pi(x)} = (1/4\tau)(d\tau^2 + \tau^2(d\theta^2 + \sin^2\theta\, d\psi^2)). \tag{6.8}$$

This shows that $B_{\pi(x)}$ is a conformally flat metric, because $4\tau B_{\pi(x)}$ is the standard flat metric expressed in the spherical coordinates (τ,θ,ψ) in $\mathbb{R}^3_+$.

Theorem 11: The internal space is endowed with the conformally flat metric which is expressed as Eq. (6.7) in Dragt's coordinates (2.16).

For the sake of use in Sec. VII, we discuss also K_x^* in a dual manner to K_x. From the definition (5.8) of K_x^* together with (6.2), (6.3), and (6.4), we obtain

$$K_x^*(\omega^i,\omega^j) = (e_i|A_x^{-1}(e_j)), \tag{6.9}$$

$$K_x^*(\sigma^i,\sigma^j) = (u_i|A_x^{-1}(u_j)), \tag{6.10}$$

$$K_x^*(d\xi^i,d\xi^j) = b^{ij}, \tag{6.11}$$

where (b^{ij}) is the inverse of (b_{ij}). Equation (6.11) can serve as a generalized definition of Wilson's G matrix.[15]

Using (6.9)–(6.11), we obtain, for a real-valued function f on Q,

$$\begin{aligned}K_x^*(df,df) &= \sum \frac{1}{m_k}\left(\frac{\partial f}{\partial x_k}\middle|\frac{\partial f}{\partial x_k}\right) \\ &= \sum (e_i|A_x^{-1}(e_j))J_i f J_j f + \sum b^{ij}\left(\frac{\partial}{\partial\xi^i}\right)^* f\left(\frac{\partial}{\partial\xi^j}\right)^* f \\ &= \sum (u_i|A_x^{-1}(u_j))L_i f L_j f + \sum b^{ij}\left(\frac{\partial}{\partial\xi^i}\right)^* f\left(\frac{\partial}{\partial\xi^j}\right)^* f.\end{aligned} \tag{6.12}$$

This equation is dual to Eq. (6.5), and the coefficients are

ready to be known from (6.6) in Dragt's coordinates;

$$K_x^*(df,df) = \frac{1}{\rho^2 \sin^2(\chi/2)}(L_1 f)^2 + \frac{1}{\rho^2 \cos^2(\chi/2)}(L_2 f)^2 + \frac{1}{\rho^2}(L_3 f)^2 + \left(\frac{\partial f}{\partial \rho}\right)^2 + \frac{4}{\rho^2}\left(\frac{\partial f}{\partial \chi}\right)^2 + \frac{4}{\rho^2 \cos^2\chi}\left(\left(\frac{\partial}{\partial \psi}\right)^* f\right)^2 . \quad (6.13)$$

B. The volume element

In order to perform integration on the internal space, we have to fix the volume element which is to be deduced from that on the configuration space Q_0. The volume element on Q_0 is of course $dQ_0 := dy_1 \wedge dy_2 \wedge dy_3$, where dy_k stands for $dy_k^1 \wedge dy_k^2 \wedge dy_k^3$, $k = 1,2,3$. The volume element dV_0 defined by the inner product K is related to dQ_0 by

$$dV_0 = (m_1 m_2 m_3)^{3/2}\, dQ_0 . \quad (6.14)$$

We recall that $(B^k/N_0, q^j)$, $k = 1,2,3$, $j = 1,\ldots,6$, serve as the Cartesian coordinates in Q_0, so that we have

$$dV_0 = N_0^{-3} dB^1 \wedge \cdots \wedge dB^3 \wedge dq^1 \wedge \cdots \wedge dq^6 . \quad (6.15)$$

Separating off the center-of-mass coordinates (B^k) from dQ_0, we obtain, from (6.14) and (6.15), the volume element on Q

$$dQ = \mu\, dq^1 \wedge \cdots \wedge dq^6 , \quad = \mu\, dV , \quad \mu = \left(\frac{\Sigma m_k}{\Pi m_k}\right)^{3/2} , \quad (6.16)$$

where $dV = dq^1 \wedge \cdots \wedge dq^6$ equals the volume element defined by the metric K_x on Q.

Expressing dV in Dragt's coordinates is straightforward. From (6.5) the volume element dV takes the form

$$dV = (\det A_x \det(b_{ij}))^{1/2} \omega^1 \wedge \omega^2 \wedge \omega^3 \wedge d\xi^1 \wedge d\xi^2 \wedge d\xi^3 = (\det A_{g^{-1}x} \det(b_{ij}))^{1/2} \sigma^1 \wedge \sigma^2 \wedge \sigma^3 \wedge d\xi^1 \wedge d\xi^2 \wedge d\xi^3 . \quad (6.17)$$

Here we have used the equality $(u_i | A_x(u_j)) = (e_i | A_{g^{-1}x}(e_j))$. We note that $\det A_x = \det A_{g^{-1}x}$ because of Eq. (3.8). Put another way, $\det A_x$ depends only on $\pi(x)$. Further, inserting (3.25) and (3.26) into (6.17), we obtain

$$dQ = \mu\, dV = dG \wedge dM , \quad (6.18)$$

where

$$dG := -\Theta^1 \wedge \Theta^2 \wedge \Theta^3 = -\Psi^1 \wedge \Psi^2 \wedge \Psi^3 , \quad (6.19)$$

$$dM := \mu(\det A_x \det(b_{ij}))^{1/2}\, d\xi^1 \wedge d\xi^2 \wedge d\xi^3 . \quad (6.20)$$

We remark that dM is $\mu(\det(A_x))^{1/2}$ times the volume element defined by the metric (b_{ij}) on M. Expressing dG and dM in Dragt's coordinates is an easy matter. From (3.19) and (3.20), and from (3.16) and (6.7), it follows that

$$dG = \sin\beta\, d\alpha \wedge d\beta \wedge d\gamma , \quad (6.21)$$

$$dM = (\mu/16)\, \rho^5 \sin 2\chi\, d\rho \wedge d\chi \wedge d\psi , \quad (6.22)$$

respectively.

Proposition 12: The volume element dM on the internal space M is given by (6.20) or (6.22), which is $\mu(\det A_x)^{1/2}$, $\mu = (\Sigma m_k / \Pi m_k)^{3/2}$, times the volume element defined by the induced metric B on M.

VII. ASSOCIATED COMPLEX VECTOR BUNDLES

A. The bundles V_l

Our goal is to set up quantum mechanics for internal molecular motions. To accomplish the task, we have to answer the question of how the Euler angles should be eliminated from the wave function $f(x)$ on Q to give wave functions on M. To do so, we invoke the vector bundle theory. Recall that the center-of-mass system is made into the principal SO(3) bundle π: $\dot{Q} \to M$. With this bundle, complex vector bundles are associated as follows: Fix a non-negative integer l. Let D^l denote the l th unitary irreducible representation of SO(3) and $\mathbb{C}^{2l+1}$ its representation space. By $D^l_{jk}(g)$, $g \in SO(3)$, we mean the matrix elements; for $z = (z_j) \in \mathbb{C}^{2l+1}$ (Ref. 16), one has

$$(D^l(g)z)_k = \sum_{j=-l}^{l} D^l_{kj}(g) z_j . \quad (7.1)$$

For a basis $|lm\rangle$ with $\hat{J}_3 |lm\rangle = m|lm\rangle$, this equation is often written as

$$D^l(g)|lm\rangle = \sum |lm'\rangle D^l_{m'm}(g) .$$

It is also well known[14] that the matrix $D^l(g)$ satisfies

$$\hat{J}^2 D^l(g) = l(l+1) D^l(g) , \quad \hat{J}^2 = \sum \hat{J}_k^2 = \sum \hat{L}_k^2 , \quad (7.2)$$

$$\hat{J}_k D^l(g) = -[\hat{J}_k] D^l(g) , \quad k = 1,2,3 , \quad (7.3)$$

$$\hat{L}_k D^l(g) = D^l(g)[\hat{L}_k] , \quad k = 1,2,3 , \quad (7.4)$$

where $[\hat{J}_k] = [\hat{L}_k]$ denote the representation matrices of $\hat{J}_k$ and $\hat{L}_k$, respectively. Especially, one has

$$[\hat{J}_3] = [\hat{L}_3] = \mathrm{diag}(l, l-1, \ldots, -l) . \quad (7.5)$$

We notice that Eqs. (7.3) and (7.4) result from the very definition of J_k and L_k. Indeed, J_k and L_k are infinitesimal generators of the left and right actions of SO(3), respectively.

Define a left action of SO(3) on the product space $\dot{Q} \times \mathbb{C}^{2l+1}$ by

$$(x, z) \to (gx, D^l(g)z) . \quad (7.6)$$

This action gives an equivalence relation in $\dot{Q} \times \mathbb{C}^{2l+1}$. The quotient manifold, denoted by $\dot{Q} \times_{SO(3)} \mathbb{C}^{2l+1}$ is made into a complex vector bundle $V_l = (\dot{Q} \times_{SO(3)} \mathbb{C}^{2l+1}, \pi_l, M)$ via the commutative diagram

$$\begin{array}{ccc} \dot{Q} \times \mathbb{C}^{2l+1} & \xrightarrow{\quad q_l \quad} & \dot{Q} \times_{SO(3)} \mathbb{C}^{2l+1} \\ {\scriptstyle \mathrm{pr}}\downarrow & & \downarrow{\scriptstyle \pi_l} \\ \dot{Q} & \xrightarrow{\quad \pi \quad} & M \end{array} , \quad (7.7)$$

where pr denotes the projection onto the first factor, and q_l is the natural projection. The π_l is the projection in V_l such that $\pi_l \circ q_l = \pi \circ \mathrm{pr}$. A map σ: $M \to \dot{Q} \times_{SO(3)} \mathbb{C}^{2l+1}$ such that $\pi \circ \sigma = \mathrm{id}$ is called a cross section in V_l. The internal states of

the molecule are then described as the cross sections in the complex vector bundle V_l. The "quantum" number l will prove to assign the total angular momentum of the molecule; $\hat{J}^2 = l(l+1)$.

A $\mathbb{C}^{2l+1}$-valued function F on $\dot{Q}$ is said to be D^l-equivariant if it satisfies

$$F(gx) = D^l(g)F(x)\,. \tag{7.8}$$

To any D^l-equivariant function, there corresponds a cross section in the complex vector bundle V_l, and vice versa.[4] We denote by $q_l^{\#}$ the one-to-one correspondence from the cross section to the D^l-equivariant function.

In our case, the correspondence $q_l^{\#}$ is quite simple on account of Theorem 2. Let σ_0 denote the cross section given in Sec. II B. Then any point x of $\dot{Q}$ is of the form $g\sigma_0(w)$, $w \in M$. Let F be a D^l-equivariant function on $\dot{Q}$. Then from (7.8) it follows that

$$F(x) = D^l(g)F(\sigma_0(w))\,. \tag{7.9}$$

Thus, setting $\Phi = F \circ \sigma_0$, one can identify Φ with a cross section and has $q_l^{\#}\Phi = F$. This means that any cross section in V_l becomes a $\mathbb{C}^{2l+1}$-valued function on M because of the triviality of the SO(3) bundle $\pi: \dot{Q} \to M$. We note also that Eq. (7.9) is also expressed as

$$F_k(x) = \sum_{j=l}^{-l} D^l_{kj}(g)\Phi_j(w)$$

in components. This form of wave functions are frequently used in the three-body problem.[10,12] If $\Phi_j(w) = \text{const}$, the F_k's are those used by van Winter for the asymmetric rotator.[17] In what follows, we treat D^l-equivariant functions in the form $D^l(g)\Phi(w)$. Now, the quantum number l is easy to understand. Indeed, from (7.2) and (7.3) the D^l-equivariant function $D^l(g)\Phi(w)$ is a set of simultaneous eigenstates of $\hat{J}^2$ and $\hat{J}_3$. The cross section Φ in V_l is therefore thought of as a set of the internal states of the molecule with a specified eigenvalue $\hat{J}^2 = l(l+1)$. This understanding goes well irrespective of whether the complex vector bundle V_l is trivial or not.

Theorem 13: The internal states of the triatomic molecule with an eigenvalue $l(l+1)$ of the square of the total angular momentum operator are described as the cross sections in the complex vector bundle V_l given in (7.7). The V_l is, however, a trivial bundle, and hence the cross sections become $\mathbb{C}^{2l+1}$-valued functions on the internal space M.

The inner product for cross sections in V_l is defined as follows: Let Φ and Ψ be cross sections in V_l. Then the Hermitian metric in V_l is defined for Φ and Ψ by

$$(\Phi|\Psi)(w) := ((q_l^{\#}\Phi)(x)|(q_l^{\#}\Psi)(x))\,, \tag{7.10}$$

where the round brackets in the right-hand side indicate the Hermitian inner product in $\mathbb{C}^{2l+1}$;

$$(u|v) = \sum_{k=l}^{-l} \bar{u}_k v_k\,, \quad u,v \in \mathbb{C}^{2l+1}\,.$$

We note that the right-hand side of (7.10) depends only on $\pi(x) = w$. The inner product of Φ and Ψ is then defined as

$$\langle\Phi|\Psi\rangle_M := \int_M (\Phi|\Psi)\,dM\,. \tag{7.11}$$

The L^2 space of cross sections with respect to this inner product is considered the state space of the internal molecular motions. From (6.18) and (7.10), the inner product is expressed as

$$\begin{aligned}\int_Q ((q_l^{\#}\Phi)(x)|(q_l^{\#}\Psi)(x))\,dQ &= \int_{SO(3)} dG \int_M (\Phi|\Psi)\,dM \\ &= 8\pi^2\langle\Phi|\Psi\rangle_M\,.\end{aligned} \tag{7.12}$$

B. The linear connection on V_l

The vector bundle V_l is equipped with the linear connection associated with the connection on $\dot{Q}$ (Ref. 4). Let X be a vector field on M and X^* its horizontal lift. Then for a cross section σ in V_l, its covariant derivative with respect to X is defined by

$$\nabla_X\sigma = q_l^{\#\,-1}X^*(q_l^{\#}\sigma)\,. \tag{7.13}$$

The operator ∇ is called the associated linear connection, which is linear in X and σ, and satisfies for arbitrary functions f on M the conditions

$$\nabla_{fX}\sigma = f\nabla_X\sigma\,, \tag{7.14}$$

$$\nabla_X f\sigma = (Xf)\sigma + f\nabla_X\sigma\,. \tag{7.15}$$

The curvature of ∇ is defined by

$$R(X,Y)\sigma = [\nabla_X,\nabla_Y]\sigma - \nabla_{[X,Y]}\sigma\,. \tag{7.16}$$

From (7.13), the curvature is also written as

$$R(X,Y)\sigma = q_l^{\#\,-1}([X^*,Y^*] - [X,Y]^*)q_l^{\#}\sigma\,. \tag{7.17}$$

It is easy to express the associated linear connection and its curvature in Dragt's coordinates. From (5.6), (7.4), and (7.13), it follows for a cross section σ with $q_l^{\#}\sigma = D^l(g)\Phi(w)$ that

$$\begin{aligned}&\nabla_{\partial/\partial\rho} = \frac{\partial}{\partial\rho}\otimes I\,, \quad \nabla_{\partial/\partial\chi} = \frac{\partial}{\partial\chi}\otimes I\,, \\ &\nabla_{\partial/\partial\psi} = \frac{\partial}{\partial\psi}\otimes I - \frac{i}{2}\sin\chi[\hat{L}_3]\,,\end{aligned} \tag{7.18}$$

where I is the $(2l+1)\times(2l+1)$ identity matrix. We note here that the connection form $\sigma_0^*\omega$ given in (3.27) is represented in the linear connection ∇ so as to couple with $\partial/\partial\xi^j$, $(\xi^j) = (\rho,\chi,\psi)$. The curvature is computed by using (7.17) to give

$$R\left(\frac{\partial}{\partial\psi},\frac{\partial}{\partial\chi}\right) = \frac{i}{2}\cos\chi[\hat{L}_3]\,, \tag{7.19}$$

and the other components vanishing. From (3.32) this curvature R proves to be a representation of $\sigma_0^*\Omega$.

Theorem 14: The linear connection and its curvature on the complex vector bundle V_l are expressed as (7.18) and (7.19), respectively, in Dragt's coordinates.

VIII. QUANTUM MECHANICS FOR INTERNAL STATES

We are in the final stage to set up quantum mechanics for internal states of the triatomic molecule. What we have to do is to obtain the internal Hamiltonian operator acting on cross sections in V_l.

A. The Laplacian on Q

We start with the kinetic energy of the molecule in Q_0, which is expressed as

$$\frac{1}{2}\int_{Q_0}\sum\frac{1}{m_k}\left(\overline{\frac{\partial f}{\partial x_k}}\,\middle|\,\frac{\partial f}{\partial x_k}\right)dQ_0\,, \tag{8.1}$$

where $\partial/\partial x_k$, $k=1,2,3$, are gradient vectors. When integrated by parts, this functional yields the Laplacian Δ_0 with respect to the inner product K;

$$\Delta_0=\sum\frac{1}{m_k}\left(\frac{\partial}{\partial x_k}\right)^2. \tag{8.2}$$

Owing to the fact that $(B^k/N_0,q^j)$ serve as the Cartesian coordinates in Q_0, the Laplacian Δ_0 are put into the form

$$\Delta_0=\sum_{i=1}^{6}\left(\frac{\partial}{\partial q^i}\right)^2+N_0^2\sum_{k=1}^{3}\left(\frac{\partial}{\partial B^k}\right)^2. \tag{8.3}$$

Separating off the center-of-mass coordinates (B^k), $k=1,2,3$, or assuming that the wave functions under consideration are independent of (B^k), we obtain the Laplacian $\Delta=\Sigma(\partial/\partial q^i)^2$ on Q. This operator can be also derived from the functional

$$\int_Q\sum\left|\frac{\partial f}{\partial q^i}\right|^2dQ=\int_Q K_x^*(\overline{df},df)\,dQ\,. \tag{8.4}$$

From (6.12), this expression turns into

$$\begin{aligned}&\int_Q\sum\left(e_i\left|A_x^{-1}(e_j)\right.\right)\overline{J_if}\,J_jf\,dQ\\&\quad+\int_Q\sum b^{ij}\overline{\left(\frac{\partial}{\partial\xi^i}\right)^*f}\left(\frac{\partial}{\partial\xi^j}\right)^*f\,dQ\\&\quad=\int_Q\sum\left(u_i\left|A_x^{-1}(u_j)\right.\right)\overline{L_if}\,L_jf\,dQ\\&\qquad+\int_Q\sum b^{ij}\overline{\left(\frac{\partial}{\partial\xi^i}\right)^*f}\left(\frac{\partial}{\partial\xi^j}\right)^*f\,dQ\,.\end{aligned} \tag{8.5}$$

In both sides, the first term is twice the rotational energy, and the last twice the vibrational energy.

We are ready to express Δ in Dragt's coordinates, using (6.13), (6.18), and (6.22). After integration by parts we can obtain the following.

Theorem 15: The Laplacian Δ on the center-of-mass system Q with respect to the metric K_x, given in (6.6), takes the form

$$\begin{aligned}\Delta=&\frac{\partial}{\partial\rho^2}+\frac{5}{\rho}\frac{\partial}{\partial\rho}+\frac{4}{\rho^2}\left[\frac{\partial^2}{\partial\chi^2}+2\cot 2\chi\frac{\partial}{\partial\chi}+\frac{1}{\cos^2\chi}\right.\\&\times\left(\frac{\partial}{\partial\psi}-\frac{i}{2}\sin\chi\,\hat{L}_3\right)^2\Big]-\left(\frac{1}{\rho^2\sin^2(\chi/2)}(\hat{L}_1)^2\right.\\&\left.+\frac{1}{\rho^2\cos^2(\chi/2)}(\hat{L}_2)^2+\frac{1}{\rho^2}(\hat{L}_3)^2\right),\end{aligned} \tag{8.6}$$

where $L_k=i\hat{L}_k$, $k=1,2,3$, are given in (4.9).

Remark: The last term including $(\hat{L}_k)^2$ is derived from the rotational energy and the rest from the vibrational energy. The vibrational part of Δ includes the differential operators coming from the horizontal lifts only [see (5.6)]. If the vibrational part is separated off and (ρ,χ,ψ) are fixed, the operator $-\Delta/2$ reduces to the well-known Hamiltonian for the rigid rotator. The Laplacian Δ was given in Ref. 12 in the expanded form.

B. The internal Hamiltonian operator

We are now in a position to obtain the internal Hamiltonian operator H_l acting on cross sections in V_l. Let H denote the Hamiltonian operator on Q;

$$H=-\tfrac{1}{2}\Delta+U\,, \tag{8.7}$$

where U is the potential function invariant under the SO(3) action. Let Φ be a cross section in V_l and F the corresponding D^l-equivariant function on $\dot{Q}$; $q_l^{\#}\Phi=F$. Then, on account of Eq. (7.12), the internal Hamiltonian operator H_l is defined through

$$\int_Q(F|HF)\,dQ=8\pi^2\langle\Phi|H_l\Phi\rangle_M\,. \tag{8.8}$$

We note that the left-hand side of (8.8) is thought of as a sum of the kinetic energy for a wave function F_k. Written out for $F(x)=D^l(g)\Phi(w)$, and integrated on SO(3), the left-hand side will yield the operator H_l. For the sake of convenience we treat the left-hand side in a form similar to (8.5). Then we have

$$\begin{aligned}&\frac{1}{2}\int_Q\sum b^{ij}\left(\left(\frac{\partial}{\partial\xi^i}\right)^*F\,\middle|\left(\frac{\partial}{\partial\xi^j}\right)^*F\right)dQ\\&\quad+\frac{1}{2}\int_Q\sum A_{ij}^{-1}(L_iF|L_jF)\,dQ\\&\quad=\frac{1}{2}\int_Q\sum b^{ij}\left(\left(\frac{\partial}{\partial\xi^i}\right)^*F\,\middle|\left(\frac{\partial}{\partial\xi^j}\right)^*F\right)dQ\\&\qquad+\frac{1}{2}\int_Q\sum a_{ij}^{-1}(J_iF|J_jF)\,dQ\,,\end{aligned} \tag{8.9}$$

where the round brackets in each integrand indicate the inner product in $\mathbb{C}^{2l+1}$, and

$$A_{ij}^{-1}=(u_i|A_x^{-1}(u_j)),\quad a_{ij}^{-1}=(e_i|A_x^{-1}(e_j))\,.$$

Using the definition of the covariant derivative and the fact that $D^l(g)$ and $[\hat{L}_k]=[\hat{J}_k]$ are unitary and Hermitian matrices, respectively, together with (7.3) and (7.4), we obtain, from (8.9),

$$\begin{aligned}&\frac{1}{2}8\pi^2\int_M\sum b^{ij}(\nabla_i\Phi|\nabla_j\Phi)\,dM+\frac{1}{2}8\pi^2\\&\quad\times\int_M\sum A_{ij}^{-1}(\Phi|[\hat{L}_i][\hat{L}_j]\Phi)\,dM\\&\quad=\frac{1}{2}8\pi^2\int_M\sum b^{ij}(\nabla_i\Phi|\nabla_j\Phi)\,dM\\&\qquad+\frac{1}{2}8\pi^2\int_M\sum a_{ij}^{-1}(\Phi|D^l(g^{-1})\\&\qquad\times[\hat{J}_i][\hat{J}_j]D^l(g)\Phi)\,dM\,,\end{aligned} \tag{8.10}$$

where ∇_i, $i=1,2,3$, stand for $\nabla_{\partial/\partial\xi^i}$. Integration by parts gives the operator

$$T_l = -\frac{1}{2}\frac{1}{J_M}\sum \nabla_i (J_M b^{ij}\nabla_j) + \frac{1}{2}\sum A_{ij}^{-1}[\hat{L}_i][\hat{L}_j]$$
$$= -\frac{1}{2}\frac{1}{J_M}\sum \nabla_i (J_M b^{ij}\nabla_j)$$
$$+\frac{1}{2}\sum a_{ij}^{-1} D^l(g^{-1})[\hat{J}_i][\hat{J}_j]D^l(g)\,, \qquad (8.11)$$

where $J_M = \mu(\det A_x \det(b_{ij}))^{1/2}$ is the volume density on M [see (6.20)]. The internal Hamiltonian H_l is then the sum $T_l + U\otimes I$, where I is $(2l+1)\times(2l+1)$ identity matrix. Writing out (8.11) in Dragt's coordinates, we obtain

$$H_l = -\frac{1}{2}\left[\left(\frac{\partial}{\partial\rho^2}+\frac{5}{\rho}\frac{\partial}{\partial\rho}\right.\right.$$
$$\left.+\frac{4}{\rho^2}\left(\frac{\partial^2}{\partial\chi^2}+2\cot 2\chi\frac{\partial}{\partial\chi}\right)\right)\otimes I$$
$$+\frac{4}{\rho^2\sin^2\chi}\left(\frac{\partial}{\partial\psi}\otimes I-\frac{i}{2}\sin\chi[\hat{L}_3]\right)^2\Bigg]$$
$$+\frac{1}{2}\left[\frac{1}{\rho^2\sin^2(\chi/2)}[\hat{L}_1]^2+\frac{1}{\rho^2\cos^2(\chi/2)}[\hat{L}_2]^2\right.$$
$$\left.+\frac{1}{\rho^2}[\hat{L}_3]^2\right]+U\otimes I\,. \qquad (8.12)$$

Theorem 16: The internal Hamiltonian operator acting on cross sections in V_l is given by $H_l = T_l + U\otimes I$, where T_l is defined in (8.11). In Dragt's coordinates, the H_l takes the form of (8.12).

We conclude this section with some remarks on the Schrödinger operator H_l. Zickendraht[10] has already derived the system of coupled equations $H_l\Phi = E\Phi$ without reference to the internal motion. The quantum three-body problem without interaction was also analyzed in Refs. 11 and 12 by using an SU(3) action on the center-of-mass system $Q\simeq\mathbb{R}^3\times\mathbb{R}^3\simeq\mathbb{C}^3$, but the internal motion was received little attention. In Refs. 18 and 19, the same problem was studied through the hyperspherical functions on $S^5\subset\mathbb{R}^6\cong Q$. For several-particle systems, the internal Hamiltonian operator will be obtained in the same manner as was done in this paper, though the topology of the internal space is difficult to study.

Remark: For the four-body problem, the internal space can be shown to be homeomorphic to $\mathbb{R}\times(S^5 - P)$, where P is a submanifold of S^5 homeomorphic to the projective plane $\mathbb{R}P^2$. (For the proof, see Ref. 20, Lemma 6.3.)

Further, we notice that when the angular momentum vanishes, the internal Hamiltonian reduces to the one we have presented in the previous paper.[3] That is, for $l = 0$, we have

$$H_0 = -\frac{1}{2}J_M^{-1}\sum \partial_i (J_M b^{ij}\partial_j) + U,\ \partial_i = \frac{\partial}{\partial\xi^i}.$$

It is worth mentioning that the first term is not $-\frac{1}{2}$ times the Laplacian on the Riemannian space M endowed with the metric B, because J_M is $(\Sigma m_k/\Pi m_k)^{3/2}(\det A_x)^{1/2}$ times $(\det(b_{ij}))^{1/2}$, and $\det A_k$ is not constant.

The advantage of the use of the connection theory is that in terms of connection theory one can understand how the internal motion is coupled with the rotation. The rotation of the molecule induces on the internal space the matrix-valued gauge field or the curvature (7.19), with which the internal motion is coupled through the gauge potential or the connection (7.18). The curvature of this connection may be called the Coriolis field in the sense that it is induced by the rotation and plays the role of a matrix-valued magnetic field on the internal space. The internal motion is driven also by the matrix-valued centrifugal potential, each of the second terms in the right-hand side of (8.11). The internal Hamiltonian operator is a second order matrix-valued differential operator including both the internal motion-Coriolis field coupling and the centrifugal potential. The ambiguity in an early paper[21] by Hirschfelder and Wigner is thus cleared up in terms of vector bundle theory.

The last but very important point to make is the fact that for a quantum system one can describe the internal motion of a molecule (Theorems 13 and 16) but for a classical system one cannot. This is explained by Jacobi's celebrated "elimination of nodes." According to Wintner,[22] the dimensions of the classical Hamiltonian system are reduced by $3 + 1 = 4$ by eliminating the angular momentum. However, in order to get a classical Hamiltonian system for the internal motion this reduction of dimensions must be $3\times 2 = 6$ in number. If so, the reduced Hamiltonian system would be of dimension $3\times 2 = 6$, twice the dimension of the internal space.

ACKNOWLEDGMENTS

The author would like to thank Dr. A. Tachibana at Kyoto University for helpful discussions.

This work was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education in Japan.

[1] J. Tennyson and B. T. Sutcliffe, J. Chem. Phys. **77**, 4061 (1982), and references therein.
[2] B. T. Sutcliffe, *Quantum Dynamics of Molecules*, edited by R. G. Woolley (Plenum, New York, 1980).
[3] A. Tachibana and T. Iwai, Phys. Rev. A **33**, 2262 (1986).
[4] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963), Vol. I.
[5] A. Guichardet, Ann. Inst. H. Poincaré **40**, 329 (1984).
[6] T. Iwai, "A geometric setting for a classical molecular dynamics," Technical Report #85009, Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan.
[7] N. Steenrod, *The Topology of Fiber Bundles* (Princeton U.P., Princeton, NJ, 1951).
[8] T. Iwai, J. Math. Phys. **28**, 964 (1987).
[9] F. T. Smith, Phys. Rev. **120**, 1058 (1960); J. Math. Phys. **3**, 735 (1962).
[10] W. Zickendraht, Ann. Phys. (NY) **35**, 18 (1965).
[11] A. J. Dragt, J. Math. Phys. **6**, 533 (1965).
[12] J.-M. Levy-Leblond and M. Levy-Nahas, J. Math. Phys. **6**, 1571 (1965).
[13] Y. Matsushima, *Differentiable Manifolds* (Marcel Dekker, New York, 1972).
[14] J.-M. Normand, *A Lie Group: Rotations in Quantum Mechanics* (North-Holland, Amsterdam, 1980).
[15] E. B. Wilson Jr., J. C. Decius, and P. C. Cross, *Molecular Vibrations* (McGraw-Hill, New York, 1955).
[16] As long as D is concerned, the components of vectors are designated by the subscript indices.
[17] C. van Winter, Physica **20**, 274 (1954).
[18] H. Mayer, J. Phys. A **8**, 1562 (1982).
[19] F. del Aguila, J. Math. Phys. **21**, 2327 (1980).
[20] M. S. Narasimhan and T. R. Ramadas, Commun. Math. Phys. **67**, 121 (1979).
[21] J. O. Hirshfelder and E. Wigner, Proc. Natl. Acad. Sci. USA **21**, 113 (1935).
[22] A. Wintner, *The Analytical Foundations of Celestial Mechanics* (Princeton U.P., Princeton, NJ, 1941).

Quantum mechanics of a charged scalar boson with respect to an observer's past light cone

G. H. Derrick
School of Physics, University of Sydney, Sydney, New South Wales 2006, Australia

(Received 7 August 1986; accepted for publication 14 January 1987)

An observer whose instantaneous "here-now" has Minkowski coordinates $z^\lambda (\lambda = 0,1,2,3)$ can only be aware of events within or on the past light cone with vertex at z^λ. In conventional quantum mechanics his current quantum state would refer to some spacelike surface containing z^λ, for example, $x^0 = z^0$. This is, however, a region of space-time about which the observer can know nothing except the single event z^λ, his current here-now. The aim of the present paper is to give a version of quantum mechanics in which the intrinsically unknowable "quantum state at the present time" is replaced by the "quantum state on the past light cone." The theory is an extension and adaptation of Dirac's point mechanics [Rev. Mod. Phys. **21**, 392 (1949)].

I. INTRODUCTION

A previous paper[1] drew attention to the problems that stem from the finiteness of the velocity of light c in the conventional approaches both to classical and to quantum theory. At a certain time t_0 an observer simply cannot know all the initial data of a system if that data is specified on the spacelike surface $t = t_0$. Reference 1 considered an adaptation of Dirac's classical Hamiltonian point mechanics[2] in which dynamical variables are specified by their values on an observer's past light cones, progression from data on an initial light cone to that on the current light cone being achieved by a canonical transformation. The aim of this present paper is to give a corresponding quantum treatment.

We first need to recall the definition of the light cone coordinates[1] belonging to an observer. Let the observer's trajectory in four dimensional Minkowski space be given in parametric form by[3]

$$x^\lambda = z^\lambda(\tau), \tag{1}$$

where the parameter τ is taken as the proper interval $\int(\eta_{\lambda\mu}\, dz^\lambda\, dz^\mu)^{1/2}$ measured along the trajectory from some arbitrary event. Thus $\eta_{\lambda\mu} v^\lambda v^\mu = 1$, $v^0 \geqslant 1$, where $v^\lambda = dz^\lambda(\tau)/d\tau$ is the observer's four-velocity vector. We now define a change of coordinates $x^\lambda \to (\tau, y^1, y^2, y^3)$ by

$$x^\lambda = z^\lambda(\tau) + y^\lambda, \quad y^0 = -y, \tag{2}$$

where $y = |\mathbf{y}|$. Thus the three-surface $\tau = \tau_0$ is the past light cone with vertex $z^\lambda(\tau_0)$, while the past-pointing null vector y^λ serves to parametrize the cone. All compatible dynamical data on this cone can be known by the observer when his personal ideal clock reads τ_0/c, i.e., when his "here-now" is $z^\lambda(\tau_0)$.

Reference 1 considered in detail the classical motion of a spin-zero particle of mass m, which was either free or suffered electromagnetic interactions. For this system the evolution of any dynamical variable f was shown to be governed by

$$\frac{df}{d\tau} = v_\lambda \{f, p^\lambda\} + \frac{\partial f}{\partial \tau}. \tag{3}$$

Here the Poisson bracket is defined with respect to a set of conjugate generalized coordinate-momentum pairs of which f and p^λ are functions. The second term on the right of (3) arises if f is additionally an explicit function of τ. The four-vector p^λ is the energy-momentum vector of the system. In Ref. 1 it was shown that for a wide class of dynamical variables (3) is equivalent to

$$\frac{\partial f}{\partial z^\lambda} = \{f, p_\lambda\} + \left(\frac{\partial f}{\partial z^\lambda}\right)_{\text{explicit}}, \tag{4}$$

where we now have *four* independent variables z^λ. The second term on the right of (4) arises if f depends not only on the conjugate variables which define the Poisson brackets but also explicitly on z^λ. Dynamical variables which satisfy (4) depend only on the particle trajectory and where the latter cuts the past light cone with vertex z^λ. The route by which the observer arrived at z^λ is irrelevent.

In the present paper we seek Schrödinger picture quantum analogs of (3) and (4) of the form

$$i\hbar \frac{d}{d\tau}|\Psi(\tau)\rangle = v_\lambda p^\lambda_{\text{OP}} |\Psi(\tau)\rangle, \tag{5}$$

$$i\hbar \frac{\partial}{\partial z^\lambda}|\Psi(z^\kappa)\rangle = p_{\text{OP}\lambda} |\Psi(z^\kappa)\rangle. \tag{6}$$

The problem is to define a suitable Hilbert space $\mathscr{H}_{\text{phys}}$ of physical states $|\Psi\rangle$ and to specify the action of the four-momentum operator p^λ_{OP} on these states. In (5), τ is regarded as an external parameter whose interpretation is that an observer following trajectory (1) finds that his quantum state $|\Psi(\tau)\rangle$ evolves with his proper time according to (5). Likewise in (6), $|\Psi(z^\kappa)\rangle$ is the state belonging to an observer whose here-now is z^κ, with (6) governing how the state varies with z^κ.

As in Ref. 1 we shall focus our attention on a system of one spin-zero boson of mass m. For this system a suitable set of classical variables is the pair $\mathbf{y}$, $\boldsymbol{\pi}$, where the coordinate $\mathbf{y}$ specifies the particle position on the observer's past light cone according to (2), and the conjugate variable $\boldsymbol{\pi}$ is defined as in Ref. 1. In the case of a free particle, the appropri-

 0022-2488/87/061327-14$02.50

ate momentum vector $p^\lambda \equiv p^\lambda(\mathbf{y},\boldsymbol{\pi})$, which governs the classical motion via (3) and (4), is given by

$$p^0 = \tfrac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi})^{-1}(\boldsymbol{\pi}^2 + m^2c^2)y,$$
$$\mathbf{p} = \boldsymbol{\pi} - \tfrac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi})^{-1}(\boldsymbol{\pi}^2 + m^2c^2)\mathbf{y}. \tag{7}$$

When an electromagnetic field derived from a potential $A^\lambda \equiv A^\lambda(x^\kappa) \equiv A^\lambda(z^\kappa + y^\kappa)$ is present, and the particle has a charge e, (7) is changed to

$$p^0 - (e/c)A^0 = \tfrac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi}_E)^{-1}(\boldsymbol{\pi}_E^2 + m^2c^2)y,$$
$$\mathbf{p} - (e/c)\mathbf{A} = \boldsymbol{\pi}_E - \tfrac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi}_E)^{-1}(\boldsymbol{\pi}_E^2 + m^2c^2)\mathbf{y}, \tag{8}$$

where

$$\boldsymbol{\pi}_E = \boldsymbol{\pi} - (e/c)(\mathbf{A} + \hat{\mathbf{y}}A^0),$$
$$\hat{\mathbf{y}} = \mathbf{y}/y. \tag{9}$$

What we need are quantum analogs of (7) and (8), where $\mathbf{y}$ and $\boldsymbol{\pi}$ are replaced by operators in some Hilbert space. An obvious candidate for the latter is $\mathscr{H}_y = L^2(\mathbb{R}^3, d^3\mathbf{y}/y)$. This is defined as the Hilbert space of complex scalar functions $\psi(\mathbf{y})$ for which the norm

$$(\psi,\psi)_y = \int \frac{d^3\mathbf{y}}{y}\,|\psi(\mathbf{y})|^2$$

exists, and with the scalar product of two elements $\psi_1(\mathbf{y})$ and $\psi_2(\mathbf{y})$ defined by

$$(\psi_1,\psi_2)_y = \int \frac{d^3\mathbf{y}}{y}\psi_1^*(\mathbf{y})\psi_2(\mathbf{y}). \tag{10}$$

If ψ_1 and ψ_2 are SO(1,3) scalars then this scalar product is Lorentz invariant on account of the like invariance of $d^3\mathbf{y}/y$. This motivates using the measure $d^3\mathbf{y}/y$ rather than $d^3\mathbf{y}$.

In the Hilbert space $\mathscr{H}_y$ the operator analogs of the classical variables $\mathbf{y}$ and $\boldsymbol{\pi}$ act according to

$$\mathbf{y}_{\mathrm{OP}}\psi(\mathbf{y}) = \mathbf{y}\psi(\mathbf{y}),$$
$$\boldsymbol{\pi}_{\mathrm{OP}}\psi(\mathbf{y}) = -i\hbar y^{1/2}\left(\frac{\partial}{\partial \mathbf{y}}\right)[y^{-1/2}\psi(\mathbf{y})]. \tag{11}$$

The operators so defined are Hermitian with respect to (10) in domains which are dense subspaces of $\mathscr{H}_y$. From these operators our aim is to construct an appropriate four-momentum operator p^λ_{OP} which determines the evolution of quantum states with respect to the external parameters τ or z^λ according to (5) or (6).

By analogy with nonrelativistic quantum mechanics, we shall tentatively interpret $|\psi(\mathbf{y},\tau)|^2\, d^3\mathbf{y}/y$ as the probability that a measurement of the position of the particle on the past light cone with vertex $z^\lambda(\tau)$ will yield a value in the range $\mathbf{y}$ to $\mathbf{y} + d\mathbf{y}$ (normalizing ψ to unity). Such a measurement, of course, needs the collaboration of a large number of auxiliary observers spread throughout the current past light cone of the central observer, the measurement having been prearranged.

In a relativistic quantum theory we should anticipate the appearance of antiparticle states. This means that we should not expect $\mathscr{H}_y$ to be the same as $\mathscr{H}_{\mathrm{phys}}$, the Hilbert space of physical states $|\Psi\rangle$. We shall in fact resolve the elements $\psi(\mathbf{y}) \in \mathscr{H}_y$ into particle and antiparticle components, and then construct the states $|\Psi\rangle$ from the particle amplitude and the complex conjugate of the antiparticle amplitude.

In Sec. II we review the conventional theory of a charged scalar boson based on the Klein–Gordon equation. This will be written both in the usual Minkowski coordinates and in the light cone coordinates defined by (2). The latter formulation adds some insight to Sec. III, which addresses the difficult problem of finding energy momentum operators p^λ_{OP} which parallel the classical expressions (7) and (8).

II. KLEIN–GORDON THEORY

A. The Klein–Gordon equation in Minkowski coordinates

In the conventional, first-quantized theory of a noninteracting charged scalar boson we have a complex amplitude $\Phi \equiv \Phi(x^\lambda)$, which obeys the Klein–Gordon equation

$$\eta^{\lambda\mu}\frac{\partial^2\Phi}{\partial x^\lambda\,\partial x^\mu} + \left(\frac{mc}{\hbar}\right)^2\Phi = 0. \tag{12}$$

By taking a Fourier transform we obtain the general solution of (12) in the form[3]

$$\Phi = (2\pi)^{-3/2}\int dS_k\left[a(\mathbf{k})e^{-ik_\lambda x^\lambda} + \{b(\mathbf{k})e^{-ik_\lambda x^\lambda}\}^*\right]. \tag{13}$$

The integration in (13) is over all future-pointing vectors k^λ lying on the mass shell, i.e.,

$$\eta_{\lambda\mu}k^\lambda k^\mu = \kappa^2, \quad \kappa = mc/\hbar,$$
$$k^0 = \epsilon_k = (\mathbf{k}^2 + \kappa^2)^{1/2}, \tag{14}$$

with the standard Lorentz invariant measure[4]

$$dS_k = d^3\mathbf{k}/\epsilon_k. \tag{15}$$

The particle and antiparticle amplitudes for four-momentum $\hbar k^\lambda$ are, respectively, $a(\mathbf{k})$ and $b(\mathbf{k})$. A Lorentz invariant scalar product[4] may be introduced between any two solutions Φ and Φ' of (12):

$$(\Phi,\Phi')_{\mathrm{KG}} = \frac{i}{2}\int_S dS^\mu\left[\Phi^*\frac{\partial\Phi'}{\partial x^\mu} - \frac{\partial\Phi^*}{\partial x^\mu}\Phi'\right], \tag{16}$$

where S is any unbounded spacelike three-surface. For sufficiently localized solutions (16) is independent of the choice of S. However the norm $(\Phi,\Phi)_{\mathrm{KG}}$ can be negative so that the scalar product (16) is not suitable for the definition of a Hilbert space. The Hilbert space of physical states $\mathscr{H}_{\mathrm{phys}}$ is constructed instead from the amplitudes $a(\mathbf{k})$ and $b(\mathbf{k})$. Let us write

$$\phi(\mathbf{k}) = \begin{bmatrix} a(\mathbf{k}) \\ b(\mathbf{k}) \end{bmatrix}, \quad \phi^\dagger(\mathbf{k}) = [a^*(\mathbf{k}), b^*(\mathbf{k})], \tag{17}$$

and define the Hilbert space $\mathscr{H}_{\mathrm{phys}} = \mathscr{H}_k$ to be the linear vector space of all $\phi(\mathbf{k})$ for which the norm (ϕ,ϕ) exists, based on the scalar product

$$(\phi,\phi') = \int dS_k\,\phi^\dagger(\mathbf{k})\phi'(\mathbf{k}),$$
$$= \int dS_k\,[a^*(\mathbf{k})a'(\mathbf{k}) + b^*(\mathbf{k})b'(\mathbf{k})]. \tag{18}$$

In $\mathscr{H}_k$ the momentum four-vector and position three-vector operators are given, respectively, by

$$p^\lambda_{\mathrm{OP}}\phi(\mathbf{k}) = \hbar k^\lambda\phi(\mathbf{k}), \tag{19}$$

$$\mathbf{x}_{\mathrm{OP}}\phi(\mathbf{k}) = i\epsilon_k^{1/2}\frac{\partial}{\partial \mathbf{k}}\left[\epsilon_k^{-1/2}\phi(\mathbf{k})\right], \tag{20}$$

while the charge operator is

$$Q = e\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}. \tag{21}$$

The Hermitian operators which generate infinitesimal rotations and boosts are defined, respectively, by

$$(j_k^{23}, j_k^{31}, j_k^{12})\phi(\mathbf{k}) = -i\hbar\mathbf{k}\times\frac{\partial\phi(\mathbf{k})}{\partial \mathbf{k}}, \tag{22}$$

$$(j_k^{01}, j_k^{02}, j_k^{03})\phi(\mathbf{k}) = -ih\epsilon_k\frac{\partial\phi(\mathbf{k})}{\partial \mathbf{k}}, \tag{23}$$

with the tensor $j_k^{\lambda\mu}$ being interpreted as the angular momentum operator in $\mathscr{H}_k$. The commutation relations of $p_{\mathrm{OP}}^{\lambda}$ and $j_k^{\lambda\mu}$ are those appropriate for the generators of the Poincaré group:

$$[p_{\mathrm{OP}}^{\kappa}, p_{\mathrm{OP}}^{\lambda}] = 0, \tag{24}$$

$$[j_k^{\mu\kappa}, p_{\mathrm{OP}}^{\lambda}] = i\hbar[\eta^{\kappa\lambda}p_{\mathrm{OP}}^{\iota} - \eta^{\iota\lambda}p_{\mathrm{OP}}^{\kappa}], \tag{25}$$

$$[j_k^{\mu\kappa}, j_k^{\lambda\mu}] = i\hbar[\eta^{\iota\mu}j_k^{\kappa\lambda} + \eta^{\kappa\lambda}j_k^{\mu\mu} - \eta^{\iota\lambda}j_k^{\kappa\mu} - \eta^{\kappa\mu}j_k^{\iota\lambda}]. \tag{26}$$

When an external electromagnetic field derived from the vector potential A^{λ} is present we modify (12) by the ansatz $\partial/\partial x^{\lambda} \to \partial/\partial x^{\lambda} + [ie/(\hbar c)]A_{\lambda}$:

$$\eta^{\lambda\mu}\left(\frac{\partial}{\partial x^{\lambda}} + \frac{ie}{\hbar c}A_{\lambda}\right)\left(\frac{\partial}{\partial x^{\mu}} + \frac{ie}{\hbar c}A_{\mu}\right)\Phi + \kappa^2\Phi = 0. \tag{27}$$

However, no longer is there in general any natural separation into particle and antiparticle amplitudes based on the sign of the frequency. Thus it is now difficult to know how to construct the Hilbert space of physical states $\mathscr{H}_{\mathrm{phys}}$, and to define therein suitable energy, momentum, and angular momentum operators. A satisfactory resolution of these problems requires second quantization and the use of the interaction picture.

B. The Klein–Gordon equation in light cone coordinates

It is instructive to rewrite the Klein–Gordon equation in terms of the light cone coordinates of an observer. The formalism so obtained is equivalent to the usual one, and does *not* yield the past light cone quantum theory that we are seeking. Nevertheless it does provide some useful pointers on how to attain our goal, and in particular leads to the valuable identities (49) and (50) below.

The metric tensor for the coordinates (τ, y^1, y^2, y^3) has been given in Ref. 1, and with its aid standard tensor analysis transforms the free Klein–Gordon equation (12) to the form

$$i\Sigma\frac{\partial\Phi}{\partial\tau} = v_{\lambda}R^{\lambda}\Phi. \tag{28}$$

In (28) Σ is a scalar operator and R^{λ} is a vector operator, with the definitions

$$\Sigma = -i\left(\mathbf{y}\cdot\frac{\partial}{\partial\mathbf{y}} + 1\right), \tag{29}$$

$$R^0 = \frac{1}{2}y\left(-\frac{\partial^2}{\partial\mathbf{y}^2} + \kappa^2\right), \tag{30}$$

$$\mathbf{R} = -\frac{\partial}{\partial\mathbf{y}}\left(\mathbf{y}\cdot\frac{\partial}{\partial\mathbf{y}}\right) - \frac{1}{2}\mathbf{y}\left(-\frac{\partial^2}{\partial y^2} + \kappa^2\right). \tag{31}$$

These operators are Hermitian with respect to the $\mathscr{H}_y$ scalar product (10), and are subject to the identity

$$\eta_{\lambda\mu}R^{\lambda}R^{\mu} = \kappa^2(\Sigma^2 - 1). \tag{32}$$

The indefinite Klein–Gordon scalar product (16) assumes a simple form in light cone coordinates. Choosing a sequence of spacelike three-surfaces which have the past light cone $\tau = \text{const}$ as their limit, we find that for sufficiently localized Φ, Φ',

$$(\Phi,\Phi')_{\mathrm{KG}} = \frac{i}{2}\int\frac{d^3\mathbf{y}}{y}\left[\Phi'\mathbf{y}\cdot\frac{\partial\Phi^*}{\partial\mathbf{y}} - \Phi^*\mathbf{y}\cdot\frac{\partial\Phi'}{\partial\mathbf{y}}\right] = (\Phi,\Sigma\Phi')_y, \tag{33}$$

where the latter is an $\mathscr{H}_y$ scalar product [see (10)]. As we shall see later in this section, Σ has positive, negative, and zero eigenvalues, consistent with the indefinite character of the Klein–Gordon scalar product.

In terms of our light cone coordinates the general free particle solution (13) becomes

$$\Phi = \int dS_k\left[a(\mathbf{k})u_k(\mathbf{y})e^{-ik_{\lambda}z^{\lambda}} + \{b(\mathbf{k})u_k(\mathbf{y})e^{-ik_{\lambda}z^{\lambda}}\}^*\right], \tag{34}$$

where

$$u_k(\mathbf{y}) = (2\pi)^{-3/2}e^{-ik_{\lambda}y^{\lambda}} = (2\pi)^{-3/2}e^{i(\epsilon_k y + \mathbf{k}\cdot\mathbf{y})}. \tag{35}$$

Substituting (34) into (28) yields the eigenvalue equations

$$\begin{aligned} R^{\lambda}u_k(\mathbf{y}) &= k^{\lambda}\Sigma u_k(\mathbf{y}),\\ R^{\lambda}u_k^*(\mathbf{y}) &= -k^{\lambda}\Sigma u_k^*(\mathbf{y}). \end{aligned} \tag{36}$$

In Appendix A we prove the following orthogonality and completeness properties:

$$(u_k, \Sigma u_{k'})_y = \epsilon_k\delta(\mathbf{k} - \mathbf{k}'), \tag{37}$$

$$(u_k^*, \Sigma u_{k'}^*)_y = -\epsilon_k\delta(\mathbf{k} - \mathbf{k}'), \tag{38}$$

$$(u_k^*, \Sigma u_{k'})_y = 0, \tag{39}$$

$$2\,\mathrm{Re}\left\{\int dS_k\ \Sigma u_k(\mathbf{y})u_k^*(\mathbf{y}')\right\} = y\delta(\mathbf{y} - \mathbf{y}'). \tag{40}$$

Using the orthogonality relations (37)–(39) we can project out the particle and antiparticle amplitudes from (34):

$$\begin{aligned} a(\mathbf{k}) &= e^{ik_{\lambda}z^{\lambda}}(u_k, \Sigma\Phi)_y,\\ b(\mathbf{k}) &= e^{ik_{\lambda}z^{\lambda}}(u_k, \Sigma\Phi^*)_y. \end{aligned} \tag{41}$$

Thus the value of $\Phi(\mathbf{y})$ on any initial light cone suffices to determine $a(\mathbf{k})$ and $b(\mathbf{k})$, and hence the value of $\Phi(\mathbf{y})$ on all later light cones. Contrast this behavior with that of the usual theory, where both Φ and $\partial\Phi/\partial x^0$ are needed as initial data. We must make one proviso, however. Integrating (28) with respect to y yields

$$i\frac{\partial\Phi}{\partial\tau} = v_{\lambda}\int_0^1 (R^{\lambda}\cdot\Phi(\alpha\mathbf{y}))d\alpha + \frac{C(\tau,\hat{\mathbf{y}})}{y}. \tag{42}$$

(The notation in the integrand means that $R^{\lambda}\Phi$ is to be evaluated at $\alpha\mathbf{y}$ before integrating over α.) An arbitrary function $C(\tau,\hat{\mathbf{y}})$ appears in (42), indicating that the evolution is not unique if one allows solutions which fall off with

distance like y^{-1}. Such solutions would involve infinite energies and would not be normalizable.

The significance of the operators R^λ may be seen from the expression for the expectation value of the momentum operator $p^\lambda_{\rm OP}$ defined by (19). We have

$$\langle p^\lambda_{\rm OP}\rangle = \int dS_k\,[\,|a(\mathbf{k})|^2 + |b(\mathbf{k})|^2]\hbar k^\lambda = (\Phi, R^\lambda \Phi)_y, \tag{43}$$

where use has been made of (36) and (40). Despite the appearance of (43), R^λ cannot be regarded as a momentum operator in $\mathscr{H}_y$ because its components do not mutually commute. Further, the identity (32) shows that the eigenvalues of $\eta_{\lambda\mu}R^\lambda R^\mu$ can have either sign.

The form of (28)–(31) is quite suggestive. Comparison with the expressions (7) for the classical momentum vector p^λ shows $\hbar^2 R^\lambda$ is the Hermitian part of the operator obtained from $(\mathbf{y}\cdot\boldsymbol{\pi})p^\lambda$ by the replacement $\boldsymbol{\pi}\to\boldsymbol{\pi}_{\rm OP}$ [see (11)]. The operator $\hbar\Sigma$ is likewise obtained from the classical quantity $D\equiv\mathbf{y}\cdot\boldsymbol{\pi}$. Consider now a solution Φ_+ of (28) which has no antiparticle content. We can heuristically write (28) as

$$i\hbar\frac{\partial\Phi_+}{\partial\tau} = v_\lambda \hbar\Sigma^{-1}R^\lambda\Phi_+, \tag{44}$$

ignoring the fact that Σ can have zero eigenvalues. Equation (44) looks like (5), with $p^\lambda_{\rm OP} = \hbar\Sigma^{-1}R^\lambda$. Unfortunately this operator is not Hermitian with respect to the scalar product (10), but only with respect to the indefinite product (33), and further, its components do not mutually commute. On these grounds $\hbar\Sigma^{-1}R^\lambda$ is not acceptable as the momentum operator, even if we manage to give it a rigorous meaning in some domain.

In subsequent sections the eigenvalues and eigenvectors of the operator Σ, defined in (29), will be of importance. A complete, orthonormal set of simultaneous eigenvectors of the commuting Hermitian operators Σ and $\hat{\mathbf{y}}$, with respective eigenvalues σ and $\mathbf{w}$, is

$$v_{\sigma\mathbf{w}}(\mathbf{y}) = (2\pi)^{-1/2}y^{i\sigma-1}\delta(\hat{\mathbf{y}},\mathbf{w}). \tag{45}$$

In (45), σ takes any real value in $(-\infty,\infty)$, $\mathbf{w}$ is any real unit vector, while $\delta(\hat{\mathbf{y}},\mathbf{w})$ is the surface Dirac delta function for a unit sphere. Strictly, these functions lie outside $\mathscr{H}_y$ because they are not normalizable, but nevertheless they comprise a useful expansion set for the elements of $\mathscr{H}_y$ on account of their orthogonality and completeness:

$$\begin{aligned}&(v_{\sigma\mathbf{w}}, v_{\sigma'\mathbf{w}'})_y = \delta(\sigma-\sigma')\delta(\mathbf{w},\mathbf{w}'),\\ &\int d\sigma\, d^2\mathbf{w}\, v_{\sigma\mathbf{w}}(\mathbf{y})v^*_{\sigma\mathbf{w}}(\mathbf{y}') = y\delta(\mathbf{y}-\mathbf{y}').\end{aligned} \tag{46}$$

In particular, the function $u_k(\mathbf{y})$ of (35) has the expansion

$$\begin{aligned}&u_k(\mathbf{y}) = \int d\sigma\, d^2\mathbf{w}\, u_k(\sigma,\mathbf{w})v_{\sigma\mathbf{w}}(\mathbf{y}),\\ &u_k(\sigma,\mathbf{w}) = \frac{ie^{1/2\pi\sigma}}{4\pi^2}\Gamma(1-i\sigma)(\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{i\sigma-1}.\end{aligned} \tag{47}$$

The integration in (47) is over the complete range of the eigenvalues, $-\infty$ to ∞ for σ and the surface of a unit sphere for $\mathbf{w}$. Substituting (47) into the orthogonality relation (37) yields

$$\begin{aligned}&\frac{1}{16\pi^3}\int \sigma^2\, d\sigma\, d^2\mathbf{w}\,\frac{e^{\pi\sigma}}{\sinh\pi\sigma}(\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{-i\sigma-1}\\ &\quad\times(\epsilon_{k'} + \mathbf{k}'\cdot\mathbf{w})^{i\sigma-1} = \epsilon_k\delta(\mathbf{k}-\mathbf{k}').\end{aligned} \tag{48}$$

If we take the complex conjugate of (48) and replace the variable of integration σ by $-\sigma$, the left-hand side assumes the same form except that $e^{\pi\sigma}$ is replaced by $-e^{-\pi\sigma}$. Averaging this integral with (48) then yields the important identity

$$\begin{aligned}&\frac{1}{16\pi^3}\int \sigma^2\, d\sigma\, d^2\mathbf{w}(\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{-i\sigma-1}(\epsilon_{k'} + \mathbf{k}'\cdot\mathbf{w})^{i\sigma-1}\\ &\quad = \epsilon_k\delta(\mathbf{k}-\mathbf{k}').\end{aligned} \tag{49}$$

In a similar manner (47) substituted into the completeness relation (40) leads to

$$\begin{aligned}&\frac{1}{16\pi^3}\int dS_k\,\sigma\sigma'(\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{i\sigma-1}(\epsilon_k + \mathbf{k}\cdot\mathbf{w}')^{-i\sigma'-1}\\ &\quad = \tfrac{1}{2}\delta(\sigma-\sigma')\delta(\mathbf{w},\mathbf{w}') + \delta(\sigma+\sigma')G(\sigma,\mathbf{w}\cdot\mathbf{w}').\end{aligned} \tag{50}$$

The form of the function G is left undetermined by this argument, but is proved in Appendix A to be

$$G(\sigma,\mathbf{w}\cdot\mathbf{w}') = -(i\sigma/4\pi)(\tfrac{1}{2}\kappa^2)^{i\sigma}(1-\mathbf{w}\cdot\mathbf{w}')^{+i\sigma-1}. \tag{51}$$

The identities (49) and (50) will play an important role in Sec. III.

III. QUANTIZATION OF THE FREE SYSTEM

A. Introduction

Passage from classical to quantum theory is notoriously ambiguous[5] because of the problem of how to order noncommuting factors. The classical system we are dealing with here is that of one free charged scalar boson whose dynamics are described in terms of the light cone coordinates (τ,y^1,y^2,y^3) belonging to an observer with trajectory (1). The energy momentum four-vector p^λ of this system is given by (7) as a function of the Hamiltonian conjugate variables $\mathbf{y}$ and $\boldsymbol{\pi}$. Suppose we attempt in (7) the substitutions $\mathbf{y}\to\mathbf{y}_{\rm OP}$ and $\boldsymbol{\pi}\to\boldsymbol{\pi}_{\rm OP}$ according to (11). How are we to order the various noncommuting factors, and how are we to interpret $(\mathbf{y}\cdot\boldsymbol{\pi})^{-1}$ when this denominator becomes an operator? Classically this denominator causes no problem, because $\mathbf{y}\cdot\boldsymbol{\pi}$ is intrinsically positive, and even when $y\to 0$ the ratio $\mathbf{y}\cdot\boldsymbol{\pi}/y$ stays finite.

Because of the above difficulties we adopt a different approach, that of defining the energy-momentum operator $p^\lambda_{\rm OP}$ by its eigenvectors and eigenvalues. Guided by Klein–Gordon theory, we seek a complete orthonormal set of states $|\Psi_{kq}\rangle\in\mathscr{H}_{\rm phys}$, labeled by a future pointing four-vector k^λ lying on the mass shell (14), and additionally by a charge index $q = \pm 1$. Thus

$$\begin{aligned}&\langle\Psi_{kq}|\Psi_{k'q'}\rangle = \epsilon_k\delta(\mathbf{k}-\mathbf{k}')\delta_{qq'},\\ &\sum_q\int dS_k\,|\Psi_{kq}\rangle\langle\Psi_{kq}| = I,\end{aligned} \tag{52}$$

where I is the unit operator in $\mathscr{H}_{\rm phys}$. Having found such a set we *define* the momentum and charge operators by

$$p^{\lambda}_{\mathrm{OP}} = \sum_q \int dS_k |\Psi_{kq}\rangle \hbar k^{\lambda} \langle \Psi_{kq}|, \tag{53}$$

$$Q = \sum_q \int dS_k |\Psi_{kq}\rangle eq \langle \Psi_{kq}|. \tag{54}$$

By construction, $|\Psi_{kq}\rangle$ is a simultaneous eigenstate of these operators with eigenvalues $\hbar k^{\lambda}$ and eq, respectively.

The above program is also subject to ambiguity. We must first decide how the Hilbert space of physical states $\mathscr{H}_{\mathrm{phys}}$ is related to $\mathscr{H}_y$. Then we must decide which of the infinite number of complete orthonormal sets satisfying (52) is the one that corresponds most closely in the classical limit to the classical system described by (7).

Section III B considers the specification of $\mathscr{H}_{\mathrm{phys}}$, based on resolution of $\psi(\mathbf{y}) \in \mathscr{H}_y$ into particle and antiparticle amplitudes. Section III C deals with the construction of complete orthonormal sets and then the remaining parts of Sec. III concern the application of these sets to (53) and (54). The modifications necessary for the incorporation of electromagnetic interactions are considered in Sec. IV.

B. Particle and antiparticle amplitudes and the specifications of $\mathscr{H}_{\mathrm{phys}}$

Let us first see how to accommodate antiparticles in the classical Hamilton theory. A free classical particle has a future-pointing timelike momentum vector p^{λ}, so that the quantity $D = \mathbf{y}\cdot\boldsymbol{\pi} = yp^0 + \mathbf{y}\cdot\mathbf{p}$ is necessarily positive. The same is true for an antiparticle. Thus classically we could treat particle and antiparticle as two disjoint systems, with conjugate variables $\mathbf{y}_p, \boldsymbol{\pi}_p$ and $\mathbf{y}_a, \boldsymbol{\pi}_a$, respectively. In this approach the phase spaces are restricted by $\mathbf{y}_p\cdot\boldsymbol{\pi}_p > 0$ and $\mathbf{y}_a\cdot\boldsymbol{\pi}_a > 0$, and the same functional form (7) applies to the evolution generators $p^{\lambda}(\mathbf{y}_p, \boldsymbol{\pi}_p)$ and $p^{\lambda}(\mathbf{y}_a, \boldsymbol{\pi}_a)$, which are interpreted, respectively, as the particle and antiparticle momentum vectors. However, particle and antiparticle can be treated alternatively as a single system by exploiting Feynman's idea that an antiparticle behaves like a particle traveling backwards in time. In this alternative approach we enlarge the classical phase space to allow all values of $\mathbf{y}\cdot\boldsymbol{\pi}$, both positive and negative. The variables $\mathbf{y}$ and $\boldsymbol{\pi}$ evolve according to

$$\frac{d\mathbf{y}}{d\tau} = v^{\lambda}\{\mathbf{y}, P_{\lambda}\}, \quad \frac{d\boldsymbol{\pi}}{d\tau} = v^{\lambda}\{\boldsymbol{\pi}, P_{\lambda}\}, \tag{55}$$

with the function $P^{\lambda}(\mathbf{y},\boldsymbol{\pi})$ defined by

$$\begin{aligned} P^0 &= \tfrac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi})^{-1}(\boldsymbol{\pi}^2 + m^2c^2)y, \\ \mathbf{P} &= \boldsymbol{\pi} - \tfrac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi})^{-1}(\boldsymbol{\pi}^2 + m^2c^2)\mathbf{y}. \end{aligned} \tag{56}$$

Here $P^{\lambda}(\mathbf{y},\boldsymbol{\pi})$ takes the same form as $p^{\lambda}(\mathbf{y},\boldsymbol{\pi})$ in (7), except that it is now defined for both signs of $\mathbf{y}\cdot\boldsymbol{\pi}$. If $\mathbf{y}\cdot\boldsymbol{\pi} > 0$ we interpret $\mathbf{y}_p = \mathbf{y}$, $\boldsymbol{\pi}_p = \boldsymbol{\pi}$ as the conjugate variables for a particle with momentum $p^{\lambda}_p = P^{\lambda}$. On the other hand, if $\mathbf{y}\cdot\boldsymbol{\pi} < 0$ then we have an antiparticle with conjugate variables $\mathbf{y}_a = \mathbf{y}$, $\boldsymbol{\pi}_a = -\boldsymbol{\pi}$ and momentum $p^{\lambda}_a = -P^{\lambda}$. In each case the momentum $p^{\lambda} = \mathrm{sgn}(\mathbf{y}\cdot\boldsymbol{\pi})P^{\lambda}$ is future pointing timelike, but the evolution generator is P^{λ} rather than p^{λ}. The charge is $e\,\mathrm{sgn}(\mathbf{y}\cdot\boldsymbol{\pi})$ (Ref. 3).

The above suggests that in quantum theory, particles and antiparticles should be associated, respectively, with the positive and negative eigenvalues of Σ, the operator analog of $D/\hbar$ defined by (29). As we saw in Sec. II B, Σ has a complete orthonormal set of eigenfunctions $v_{\sigma w}(\mathbf{y})$, given by (45), the eigenvalue σ taking all values in $(-\infty, \infty)$. Any element $\psi(\mathbf{y}) \in \mathscr{H}_y$ has the decomposition

$$\psi(\mathbf{y}) = \psi_+(\mathbf{y}) + \psi_-(\mathbf{y}), \tag{57}$$

where

$$\psi_+(\mathbf{y}) = \int_0^{\infty} d\sigma \int d^2\mathbf{w}\, v_{\sigma w}(\mathbf{y})(v_{\sigma w}, \psi)_y,$$

$$\psi_-(\mathbf{y}) = \int_{-\infty}^{0} d\sigma \int d^2\mathbf{w}\, v_{\sigma w}(\mathbf{y})(v_{\sigma w}, \psi)_y.$$

The above equations may be written

$$\psi_+(\mathbf{y}) = \Theta(\Sigma)\psi(\mathbf{y}), \quad \psi_-(\mathbf{y}) = \Theta(-\Sigma)\psi(\mathbf{y}),$$

where the step functions $\Theta(\Sigma)$ and $\Theta(-\Sigma)$ are projection operators with matrix elements

$$\begin{aligned} \langle \mathbf{y}|\Theta(\Sigma)|\mathbf{y}'\rangle &= \int_0^{\infty} d\sigma \int d^2\mathbf{w}\, v_{\sigma w}(\mathbf{y}) v^*_{\sigma w}(\mathbf{y}'), \\ &= \frac{i\delta(\hat{\mathbf{y}},\hat{\mathbf{y}}')}{2\pi y y' [\log(y/y') + i\epsilon]}, \\ &= \frac{1}{2} y\delta(\mathbf{y} - \mathbf{y}') + \frac{i}{2\pi} \mathscr{P} \frac{\delta(\hat{\mathbf{y}},\hat{\mathbf{y}}')}{yy' \log(y/y')}, \end{aligned} \tag{58}$$

$$\langle \mathbf{y}|\Theta(-\Sigma)|\mathbf{y}'\rangle = \langle \mathbf{y}|\Theta(\Sigma)|\mathbf{y}'\rangle^*. \tag{59}$$

In (58) $\epsilon \to +0$ and $\mathscr{P}$ denotes the principal value. An equivalent form is

$$\Theta(\pm\Sigma)\psi(\mathbf{y}) = \frac{1}{2}\psi(\mathbf{y}) \pm \frac{i}{2\pi} \mathscr{P} \int_0^{\infty} \frac{\psi(\alpha\mathbf{y})d\alpha}{-\log\alpha}. \tag{60}$$

Corresponding to (57), $\mathscr{H}_y$ has the decomposition

$$\mathscr{H}_y = \mathscr{H}^+_y \oplus \mathscr{H}^-_y, \tag{61}$$

where the two subspaces $\mathscr{H}^+_y$ and $\mathscr{H}^-_y$ are themselves Hilbert spaces, whose elements satisfy $\Theta(\Sigma)\psi_+ = \psi_+$ and $\Theta(-\Sigma)\psi_- = \psi_-$, respectively.

It remains now to relate $\psi(\mathbf{y}) \in \mathscr{H}_y$ to physical states $|\Psi\rangle \in \mathscr{H}_{\mathrm{phys}}$. By analogy with Klein–Gordon theory let us suppose that physical states $|\Psi\rangle$ can be represented by

$$\Psi(\mathbf{y}) = \begin{bmatrix} \psi_p(\mathbf{y}) \\ \psi_a(\mathbf{y}) \end{bmatrix}, \tag{62}$$

with $\psi_p(\mathbf{y})$ and $\psi_a(\mathbf{y})$ being the particle and antiparticle amplitudes. The appropriate scalar product is

$$(\Psi,\Psi')_{\mathrm{phys}} = \int \frac{d^3\mathbf{y}}{y} [\psi^*_p(\mathbf{y})\psi'_p(\mathbf{y}) + \psi^*_a(\mathbf{y})\psi'_a(\mathbf{y})]. \tag{63}$$

Let us make the identification

$$\begin{aligned} &\psi_p(\mathbf{y}) = \psi_+(\mathbf{y}) = \Theta(\Sigma)\psi(\mathbf{y}), \\ &\psi_a(\mathbf{y}) = [\psi_-(\mathbf{y})]^* = \Theta(\Sigma)\psi^*(\mathbf{y}), \\ &\mathscr{H}_{\mathrm{phys}} = \mathscr{H}^+_y \oplus \mathscr{H}^+_y. \end{aligned} \tag{64}$$

Thus both components of the physical state $\Psi(\mathbf{y})$ belong to the positive eigenspace of Σ:

$$\Theta(\Sigma)\Psi(\mathbf{y}) = \Psi(\mathbf{y}).$$

This is the quantum analog of the classical result that both $\mathbf{y}_p \cdot \boldsymbol{\pi}_p$ and $\mathbf{y}_a \cdot \boldsymbol{\pi}_a$ are positive.

The hypothesis given by (62)–(64) is similar to that of Klein–Gordon and Dirac theory in that particle and antiparticle amplitudes must be projected out from the coordinate space wave function in order to define a physical state. We have a one-to-one mapping, albeit nonlinear, between the points $\Psi(\mathbf{y}) \in \mathscr{H}_{\text{phys}}$ and the points $\Psi(\mathbf{y}) \in \mathscr{H}_y$:

$$\Psi(\mathbf{y}) = \begin{bmatrix} \psi_p(\mathbf{y}) \\ \psi_a(\mathbf{y}) \end{bmatrix} = \Theta(\Sigma) \begin{bmatrix} \psi(\mathbf{y}) \\ \psi^*(\mathbf{y}) \end{bmatrix}, \qquad \psi(\mathbf{y}) = \psi_p(\mathbf{y}) + \psi_a^*(\mathbf{y}). \tag{65}$$

The momentum operator p_{OP}^{λ}, whose explicit form we have yet to determine, governs the evolution of the physical states $\Psi(\mathbf{y})$ according to (5). Analogy with the classical evolution equation (55) suggests that there should also exist an operator P_{OP}^{λ} which determines the evolution of coordinate space wave functions $\psi(\mathbf{y})$ according to

$$i\hbar \frac{\partial \psi}{\partial \tau} = v_\lambda P_{\text{OP}}^{\lambda} \psi. \tag{66}$$

Note that the operators p_{OP}^{λ} and P_{OP}^{λ} act in different Hilbert spaces, viz $\mathscr{H}_{\text{phys}}$ and $\mathscr{H}_y$, respectively. We now consider the problem of how to give effect to the ansatz (53) for determining p_{OP}^{λ}, and how to specify P_{OP}^{λ} in terms of the latter.

C. Complete and orthonormal sets

We seek complete, orthonormal sets of states $|\Psi_{kq}\rangle \in \mathscr{H}_{\text{phys}}$ satisfying (52), which states are to be interpreted as energy-momentum-charge eigenstates according to (53) and (54). Let $\psi_{kq}(\mathbf{y})$ be the element of $\mathscr{H}_y$ from which $\Psi_{kq}(\mathbf{y})$is derived by the prescription (65). Following the ideas of the previous section, we assume that the particle states and the antiparticle states correspond, respectively, to the positive and to the negative eigenspaces of Σ. Thus

$$\Theta(\Sigma)\psi_{k1}(\mathbf{y}) = \psi_{k1}(\mathbf{y}), \qquad \Theta(-\Sigma)\psi_{k(-1)}(\mathbf{y}) = \psi_{k(-1)}(\mathbf{y}). \tag{67}$$

Motivated by the desire to achieve maximum particle–antiparticle symmetry we make the additional assumption $\psi_{k(-1)}(\mathbf{y}) = \psi_{k1}^*(\mathbf{y})$, which is consistent with (67). For brevity we write $\psi_k(\mathbf{y})$ for $\psi_{k1}(\mathbf{y})$ so that (67) becomes

$$\psi_{k1}(\mathbf{y}) = \psi_k(\mathbf{y}) = \psi_{k(-1)}^*(\mathbf{y}), \qquad \Theta(\Sigma)\psi_k(\mathbf{y}) = \psi_k(\mathbf{y}). \tag{68}$$

The orthogonality and completeness relations (52) now reduce to

$$(\psi_k, \psi_{k'})_y = \epsilon_k \delta(\mathbf{k} - \mathbf{k}'), \tag{69}$$

$$\int dS_k \, \psi_k(\mathbf{y}) \psi_k^*(\mathbf{y}') = \langle \mathbf{y} | \Theta(\Sigma) | \mathbf{y}' \rangle. \tag{70}$$

Since $\psi_k(\mathbf{y}) \in \mathscr{H}_y^+$, $\Theta(\Sigma)$ acts as the unit operator in the completeness relation (70). [See (58).] Having found a solution of (68)–(70) we can reconstruct $\Psi_{kq}(\mathbf{y})$ by

$$\Psi_{k1}(\mathbf{y}) = \begin{bmatrix} \psi_k(\mathbf{y}) \\ 0 \end{bmatrix}, \tag{71}$$

$$\Psi_{k(-1)}(\mathbf{y}) = \begin{bmatrix} 0 \\ \psi_k(\mathbf{y}) \end{bmatrix}. \tag{72}$$

[See (62), (64), and (65).]

The functions $\Psi_{kq}(\mathbf{y})$ define a unitary mapping between the points $\Psi(\mathbf{y}) \in \mathscr{H}_{\text{phys}}$ and $\phi(\mathbf{k}) \in \mathscr{H}_k$ [see (17)]:

$$a_k = (\Psi_{k1}, \Psi)_{\text{phys}} = (\psi_k, \psi)_y, \qquad b_k = (\Psi_{k(-1)}, \Psi)_{\text{phys}} = (\psi_k, \psi^*)_y. \tag{73}$$

The inverse mapping is

$$\Psi(\mathbf{y}) = \int dS_k \, [a_k \Psi_{k1}(\mathbf{y}) + b_k \Psi_{k(-1)}(\mathbf{y})], \qquad \psi(\mathbf{y}) = \int dS_k \, [a_k \psi_k(\mathbf{y}) + b_k^* \psi_k^*(\mathbf{y})]. \tag{74}$$

We now show that angular momentum considerations dictate that $\Psi_{kq}(\mathbf{y})$, and hence $\psi_k(\mathbf{y})$, be taken as a function solely of the SO(1,3) scalar

$$\zeta = -k_\lambda y^\lambda = \epsilon_k y + \mathbf{k} \cdot \mathbf{y}. \tag{75}$$

[This is analogous to the result of nonrelativistic quantum theory that the function which maps coordinate space onto momentum space, namely $(2\pi)^{-3/2} \exp(i\mathbf{k}\cdot\mathbf{x})$, depends solely on the SO(3) scalar $\mathbf{k}\cdot\mathbf{x}$.] Let us derive the form of the angular momentum tensor operator using Wigner's P_R prescription.[6] An infinitesimal Lorentz transformation parametrized by the antisymmetric matrix $\omega_{\nu\mu}$,

$$y'^\lambda = y^\lambda + \eta^{\lambda\nu}\omega_{\nu\mu} y^\mu,$$

changes the functional form of $\Psi(\mathbf{y})$ to

$$\Psi'(\mathbf{y}) = (I - \tfrac{1}{2} i\omega_{\lambda\mu} j_y^{\lambda\mu})\Psi(\mathbf{y}),$$

where the components of the angular momentum operator $j_y^{\lambda\mu}$ are

$$(j_y^{23}, j_y^{31}, j_y^{12}) = -i\hbar \mathbf{y} \times \frac{\partial}{\partial \mathbf{y}}, \qquad (j_y^{01}, j_y^{02}, j_y^{03}) = i\hbar y \frac{\partial}{\partial \mathbf{y}}. \tag{76}$$

We now demand that the unitary transformation (73) and (74) should transform $j_k^{\lambda\mu}$ given by (22) and (23) into $j_y^{\lambda\mu}$. This will be the case if and only if

$$(j_k^{\lambda\mu} + j_y^{\lambda\mu})\Psi_{kq}(\mathbf{y}) = 0.$$

The above equation is the condition that $\Psi_{kq}(\mathbf{y})$ is unchanged by a combined Lorentz transformation of the two vectors k^λ, y^λ, which implies that $\Psi_{kq}(\mathbf{y})$, and hence $\psi_k(\mathbf{y})$ are functions only of the scalar ζ defined by (75):

$$\psi_k(\mathbf{y}) = f(\zeta). \tag{77}$$

Let us now expand $\psi_k(\mathbf{y})$ in terms of the complete, orthonormal set $v_{\sigma w}(\mathbf{y})$ given in (45). Only terms with $\sigma \geqslant 0$ are needed on account of (68). The expansion coefficients are

$$\psi_{k\sigma w} = (v_{\sigma w}, \psi_k)_y, \qquad = F(\sigma)(\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{i\sigma - 1}, \tag{78}$$

where

$$F(\sigma) = (2\pi)^{-1/2} \int_0^\infty f(\zeta) \zeta^{-i\sigma} d\zeta.$$

Hence

$$\psi_k(\mathbf{y}) = \int_0^\infty d\sigma \int d^2\mathbf{w}\, \psi_{k\sigma\mathbf{w}} v_{\sigma\mathbf{w}}(\mathbf{y}),$$

$$= (2\pi)^{-1/2} \int_0^\infty d\sigma\, F(\sigma) \zeta^{i\sigma-1}. \tag{79}$$

The function $F(\sigma)$ must be chosen such that the orthogonality and completeness relations (69) and (70) are satisfied. Substituting (79) into these equations yields

$$\int_0^\infty |F(\sigma)|^2\, d\sigma \int d^2\mathbf{w} (\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{-i\sigma-1} (\epsilon_{k'} + \mathbf{k}'\cdot\mathbf{w})^{i\sigma-1}$$

$$= \epsilon_k \delta(\mathbf{k} - \mathbf{k}'), \tag{80}$$

$$\int dS_k\, F(\sigma) F^*(\sigma') (\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{i\sigma-1} (\epsilon_k + \mathbf{k}\cdot\mathbf{w}')^{-i\sigma'-1}$$

$$= \delta(\sigma - \sigma')\delta(\mathbf{w},\mathbf{w}'). \tag{81}$$

The form of these equations is similar to that of (49) and (50), except that here only positive values of σ and σ' are allowed.

We now prove a theorem derived from (49) and (50).

Theorem: A solution of (80) and (81) is given by any function $F(\sigma)$ whose modulus is $(2\pi)^{-3/2}\sigma$, i.e., we may write

$$F(\sigma) = (2\pi)^{-3/2} i\sigma e^{ig(\sigma)}, \quad \sigma \geqslant 0, \tag{82}$$

where $g(\sigma)$ is an arbitrary real function. The proof of this theorem depends on two lemmas.

Lemma 1: For any $F_1(\sigma)$ and $F_2(\sigma)$ defined in $-\infty < \sigma < \infty$, the integral

$$J = \int_{-\infty}^{\infty} F_1^*(\sigma) F_2(\sigma) d\sigma$$

$$\times \int d^2\mathbf{w} (\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{-i\sigma-1} (\epsilon_{k'} + \mathbf{k}'\cdot\mathbf{w})^{i\sigma-1}$$

is unchanged if $\mathbf{k}$ and $\mathbf{k}'$ are interchanged.

Lemma 2: The integral J of Lemma 1 is given by

$$J = \int_0^\infty [F_1^*(\sigma) F_2(\sigma) + F_1^*(-\sigma) F_2(-\sigma)] d\sigma$$

$$\times \int d^2\mathbf{w} (\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{-i\sigma-1} (\epsilon_{k'} + \mathbf{k}'\cdot\mathbf{w})^{i\sigma-1}.$$

Lemma 1 is proved by observing that J is the Lorentz invariant scalar product $(f_1, f_2)_y$, where

$$f_1(\mathbf{y}) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} d\sigma\, F_1(\sigma) (-k_\lambda y^\lambda)^{i\sigma-1},$$

$$f_2(\mathbf{y}) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} d\sigma\, F_2(\sigma) (-k'_\lambda y^\lambda)^{i\sigma-1}.$$

Since both k_λ and k'_λ lie on the mass shell (14), the only scalars we can form from these vectors are $k^\lambda k'_\lambda$ and $\kappa = mc/\hbar$ (Ref. 7). Hence J is a function only of these scalars, and is consequently symmetric in k_λ and k'_λ, proving the first lemma. To prove Lemma 2, divide the range of integration over σ into the two intervals $(0,\infty)$ and $(-\infty,0)$. If in the second integral we replace σ by $-\sigma$ and invoke Lemma 1, then we obtain the form for J given by Lemma 2.

With the aid of these lemmas we can now prove the theorem. Applying Lemma 2 to (49), with $F_1(\sigma) = F_2(\sigma) = \sigma$ we obtain

$$(2\pi)^{-3} \int_0^\infty \sigma^2\, d\sigma \int d^2\mathbf{w} (\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{-i\sigma-1}$$

$$\times (\epsilon_{k'} + \mathbf{k}'\cdot\mathbf{w})^{i\sigma-1} = \epsilon_k \delta(\mathbf{k} - \mathbf{k}'). \tag{83}$$

With the restriction to positive σ values (50) becomes

$$(2\pi)^{-3} \int dS_k\, \sigma\sigma' (\epsilon_k + \mathbf{k}\cdot\mathbf{w})^{i\sigma-1}$$

$$\times (\epsilon_k + \mathbf{k}\cdot\mathbf{w}')^{-i\sigma'-1} = \delta(\sigma - \sigma')\delta(\mathbf{w},\mathbf{w}'). \tag{84}$$

From (83) and (84) we see that any $F(\sigma)$ of the form (82) will satisfy (80) and (81), which establishes the theorem.

To summarize Sec. III C, a complete orthonormal set of states $\Psi_{kq}(\mathbf{y})$ is given by (71) and (72), with

$$\psi_k(\mathbf{y}) = f(-k_\lambda y^\lambda)$$

$$= (2\pi)^{-2} i \int_0^\infty d\sigma\, \sigma e^{ig(\sigma)} (-k_\lambda y^\lambda)^{i\sigma-1}, \tag{85}$$

$g(\sigma)$ being any arbitrary real function. Some examples for particular values of $g(\sigma)$ are given in Appendix B. The question of what function $g(\sigma)$ should be adopted will now be considered.

D. Choice of the phase function $g(\sigma)$

Equations (71), (72), and (85) together define a complete orthonormal set of states $\Psi_{kq}(\mathbf{y})$. Consider an arbitrary element $\Psi(\mathbf{y}) \in \mathscr{H}_{\text{phys}}$ derived from the element $\psi(\mathbf{y}) \in \mathscr{H}_y$ via the mapping (65). According to (53) the action of the momentum operator in $\mathscr{H}_{\text{phys}}$ is

$$p^\lambda_{\text{OP}} \Psi(\mathbf{y}) = \int dS_k\, \psi_k(\mathbf{y}) \hbar k^\lambda \begin{bmatrix} (\psi_k,\psi_p)_y \\ (\psi_k,\psi_a)_y \end{bmatrix}. \tag{86}$$

To be consistent with (66) the evolution operator in $\mathscr{H}_y$ has to act according to

$$P^\lambda_{\text{OP}} \psi(\mathbf{y}) = \int dS_k\, \hbar k^\lambda [\psi_k(\mathbf{y}) (\psi_k,\psi)_y$$

$$- \psi_k^*(\mathbf{y}) (\psi_k^*,\psi)_y]. \tag{87}$$

Now any choice of the real function $g(\sigma)$ gives rise to a unitary representation in $\mathscr{H}_{\text{phys}}$ of the Poincaré group, based on the infinitesimal generators p^λ_{OP} and $j_y^{\lambda\mu}$ [see (76)]. This representation is unitarily equivalent via (73) and (74) to the direct sum of two copies of the standard mass m spin-zero representation,[8] corresponding to particle and antiparticle. A representation is likewise generated in $\mathscr{H}_y$ by P^λ_{OP} and $j_y^{\lambda\mu}$.

However, we cannot leave $g(\sigma)$ arbitrary if we wish the quantum theory to correspond in the classical limit to the formalism based on the classical momentum vector of (7). We want $\mathbf{y}$ to correspond to the light cone coordinate defined by (2), with $|\psi(\mathbf{y})|^2\, d^3\mathbf{y}/y$ representing the probability that a measurement of this coordinate will yield a value in the interval $(\mathbf{y},\mathbf{y} + d\mathbf{y})$. In what follows we present a heuristic argument which makes plausible the particular choice

$$g(\sigma) = \arg[\Gamma(-i\sigma)],$$

$$\psi_k(\mathbf{y}) = (2\pi)^{-2} i \int_0^\infty d\sigma\, \sigma (\pi^{-1}\sigma \sinh \pi\sigma)^{1/2} \tag{88}$$

$$\times \Gamma(-i\sigma) (-k_\lambda y^\lambda)^{i\sigma-1},$$

where Γ denotes the usual gamma function.

Our starting point is the observation that there exists a classical canonical transformation which transforms the classical evolution generator P^λ of (56) into $-P^\lambda$. The transformed variables $\mathbf{\Pi}$, Y^λ, with $Y^0 = -|\mathbf{Y}|$, are derived from the generating function $mc(2y_\lambda Y^\lambda)^{1/2}$, and take the form

$$\begin{aligned} Y^0 &= -(mc)^{-2}\boldsymbol{\pi}^2 y, \\ \mathbf{Y} &= (mc)^{-2}[\boldsymbol{\pi}^2\mathbf{y} - 2(\mathbf{y}\cdot\boldsymbol{\pi})\boldsymbol{\pi}], \\ \mathbf{\Pi} &= (mc)^2\boldsymbol{\pi}/\boldsymbol{\pi}^2. \end{aligned} \tag{89}$$

Substitution into (56) leads to

$$P^\lambda(\mathbf{Y},\mathbf{\Pi}) = -P^\lambda(\mathbf{y},\boldsymbol{\pi}), \quad \mathbf{Y}\cdot\mathbf{\Pi} = -\mathbf{y}\cdot\boldsymbol{\pi}. \tag{90}$$

Thus in the $\mathbf{Y},\mathbf{\Pi}$ system, particle and antiparticle have interchanged roles. Double application of this canonical transformation leads back to the original variables.

These classical results suggest that there ought to be a quantum counterpart to (89) in the form of a unitary operator W which interchanges the role of particle and antiparticle wave functions in $\mathscr{H}_y$. Thus W should satisfy

$$\begin{aligned} &WP^\lambda_{\rm OP}W^\dagger = -P^\lambda_{\rm OP}, \\ &W\psi_k = \psi_k^*, \quad W = W^{-1} = W^\dagger. \end{aligned} \tag{91}$$

An operator equivalent to (89) in $\mathscr{H}_y$ is

$$Y^0_{\rm OP} = \kappa^{-2} y \frac{\partial^2}{\partial \mathbf{y}^2},$$

$$\mathbf{Y}_{\rm OP} = \kappa^{-2}\left[-\mathbf{y}\frac{\partial^2}{\partial \mathbf{y}^2} + 2\frac{\partial}{\partial \mathbf{y}}\left(\mathbf{y}\cdot\frac{\partial}{\partial \mathbf{y}}\right)\right],$$

where $\kappa = mc/\hbar$. These four operators are Hermitian in a dense subspace of $\mathscr{H}_y$, mutually commute, and satisfy $\eta_{\lambda\mu}Y^\lambda_{\rm OP}Y^\mu_{\rm OP} = 0$. A complete orthonormal set of simultaneous eigenfunctions with respective eigenvalues Y^λ ($Y^0 = -|\mathbf{Y}|$) is given by

$$w(y_\lambda Y^\lambda) = (\kappa^2/4\pi)J_0([2\kappa^2 y_\lambda Y^\lambda]^{1/2}). \tag{92}$$

The eigenvalue property may be proved readily by direct differentiation. The orthogonality and completeness relations are

$$\int \frac{d^3\mathbf{y}}{y} w(y_\lambda Y^\lambda)w(y_\lambda Y^{\lambda'}) = Y\delta(\mathbf{Y}-\mathbf{Y}'),$$

$$\int \frac{d^3\mathbf{Y}}{Y} w(y_\lambda Y^\lambda)w(y'_\mu Y^\mu) = y\delta(\mathbf{y}-\mathbf{y}'),$$

which may be proved by methods akin to those used in Appendix A.

The appropriate definition of the action of W on $\psi(\mathbf{y})\in\mathscr{H}_y$ is then

$$(W\psi)(\mathbf{y}) = \int \frac{d^3\mathbf{Y}}{Y} w(y_\lambda Y^\lambda)\psi(\mathbf{Y}). \tag{93}$$

Applying (93) to $\psi_k(\mathbf{y})$ given by (85) we find (see Appendix C)

$$(W\psi_k)(\mathbf{y}) = -(2\pi)^{-2}i\int_0^\infty d\sigma\, \sigma e^{ig(\sigma)} \times \frac{\Gamma(i\sigma)}{\Gamma(-i\sigma)}(-k_\lambda y^\lambda)^{-i\sigma-1}, \tag{94}$$

which coincides with $\psi_k^*(\mathbf{y})$ if

$$\begin{aligned} e^{ig(\sigma)} &= [\Gamma(-i\sigma)/\Gamma(i\sigma)]^{1/2}, \\ &= \pm(\pi^{-1}\sigma\sinh\pi\sigma)^{1/2}\Gamma(-i\sigma). \end{aligned}$$

Thus apart from a trivial assignment of sign, (88) is the only choice of the set $\psi_k(\mathbf{y})$ for which the classical symmetry property (90) goes over to the quantum symmetry property (91) under the unitary mapping W corresponding to (89).

It does not appear possible to express the momentum eigenfunctions $\psi_k(\mathbf{y})$ given by (88) in terms of elementary functions. However only the asymptotic behavior for large $\zeta \equiv -k_\lambda y^\lambda$ will be needed in this paper. The integrand contains the factor $\exp\{i\arg[\Gamma(-i\sigma)] + i\sigma\log\zeta\}$, which, for large ζ, oscillates rapidly except in the neighborhood of a stationary point of the exponent. The only stationary point occurs at the value of $\sigma = \zeta - (12\zeta)^{-1} + O(\zeta^{-3})$. The method of stationary phase then gives the asymptotic behavior for large ζ as

$$\psi_k(\mathbf{y}) \sim \frac{i\zeta^{1/2}e^{i\zeta}}{(2\pi)^{3/2}}\left[1 + \frac{1}{12\zeta} - \frac{1}{24\zeta^2} + O(\zeta^{-3})\right]. \tag{95}$$

At the other extreme, when ζ is small, the integral is dominated by values of σ of the order of $-1/\log\zeta$. The limiting behavior as $\zeta\to 0$ is

$$4\pi^2\zeta\psi_k(\mathbf{y}) \sim (\log\zeta)^{-2} - 2\gamma(\log\zeta)^{-3} + O((\log\zeta)^{-4}), \tag{96}$$

where $\gamma \equiv 0.577\,22$ is Euler's constant.

Further creditability will now be given to (88), by showing that this choice of $g(\sigma)$ leads to the correct nonrelativistic quantum mechanics limit ($c\to\infty$), and to the correct classical mechanics limit ($\hbar\to 0$).

E. The nonrelativistic limit

Consider an observer whose world line passes from $z^\lambda(\tau_0)$ to $z^\lambda(\tau)$. If at proper time τ_0/c the observer finds the particle with the momentum eigenfunction $\psi_k(\mathbf{y})$, then (66) predicts that at the later proper time τ/c his wave function will have evolved to

$$\psi_k(\mathbf{y},z) = \psi_k(\mathbf{y})\exp\{-ik_\lambda[z^\lambda(\tau) - z^\lambda(\tau_0)]\}. \tag{97}$$

These functions satisfy the orthogonality relations

$$(\psi_k(\mathbf{y},z),\ \psi_{k'}(\mathbf{y},z))_y = \epsilon_k\delta(\mathbf{k}-\mathbf{k}'),$$

which may be written

$$\int d^3\mathbf{y}\,\phi_k^*(\mathbf{y},z)\phi_{k'}(\mathbf{y},z) = \delta(\mathbf{k}-\mathbf{k}'), \tag{98}$$

with

$$\phi_k(\mathbf{y},z) = (\epsilon_k y)^{-1/2}\psi_k(\mathbf{y},z).$$

We will interpret $c\to\infty$ as meaning that both the dimensionless quantities κy and κ/k, with $\kappa = mc/\hbar$, become indefinitely large. In this limit $\zeta = \epsilon_k y + \mathbf{k}\cdot\mathbf{y}$ becomes large and we may use (95). The leading term is

$$\phi_k(\mathbf{y},z) \sim i(2\pi)^{-3/2}\exp\{-ik_\lambda[y^\lambda + z^\lambda(\tau) - z^\lambda(\tau_0)]\}$$
$$\sim i(2\pi)^{-3/2}\exp\{i\mathbf{k}\cdot\mathbf{x} - i\hbar^{-1}[mc^2 + \hbar^2k^2/(2m)]t + ik_\lambda z^\lambda(\tau_0)\},$$

where $x^\lambda = z^\lambda(\tau) + y^\lambda$ [see (2)], and $x^0 = ct$. Apart from an irrelevant phase factor, the above is the usual form for the energy-momentum eigenfunction in nonrelativistic quantum mechanics (including the rest energy mc^2). Note that in the nonrelativistic limit the integration $\int d^3\mathbf{y}$ in (98), taken over the past light cone $\tau = \text{const}$, reduces to $\int d^3\mathbf{x}$, taken over the hyperplane $t = \text{const}$.

F. The classical limit

Following Landau and Lifshitz,[9] we shall seek the classical limit by considering wave functions with large phases and large wavenumbers. A quasiclassical wave function for a particle is of the form

$$\psi(\mathbf{y},z) = \int dS_k\ \psi_k(\mathbf{y})A(\hbar\mathbf{k})\exp\{i\hbar^{-1}\phi(\hbar\mathbf{k}) - ik_\lambda z^\lambda(\tau)\}.$$

Here $\phi(\hbar\mathbf{k})$ is real, and we assume that $A(\hbar\mathbf{k})$ and $\phi(\hbar\mathbf{k})$ are functions which differ from zero in regions where $p^\lambda = \hbar k^\lambda$ has a macroscopic value, with k becoming indefinitely large as we let $\hbar \to 0$. Thus $\zeta = \epsilon_k y + \mathbf{k}\cdot\mathbf{y}$ is likewise large and once again the asymptotic form (95) is applicable. Changing variable to $\mathbf{p} = \hbar\mathbf{k}$ and using (95) yields

$$\psi(\mathbf{y},z) = \frac{i}{(2\pi)^{3/2}\hbar^2}\int \frac{d^3\mathbf{p}}{p^0}\zeta^{1/2}A(\mathbf{p}) \times\exp\{i\hbar^{-1}[\phi(\mathbf{p}) - p_\lambda x^\lambda]\}, \tag{99}$$

where $x^\lambda = z^\lambda(\tau) + y^\lambda$ and $p^0 = (\mathbf{p}^2 + m^2c^2)^{1/2}$. As $\hbar \to 0$ the exponent in (99) oscillates rapidly, and we may use the method of stationary phase. The stationary point is given by

$$\frac{\partial\phi}{\partial\mathbf{p}} = \frac{\mathbf{p}}{p^0}x^0 - \mathbf{x},$$

whose solution $\mathbf{p} = \mathbf{p}_s(x)$ is a function of x^λ. The stationary phase approximation to (99) then assumes the form

$$\psi = B\exp\{i\hbar^{-1}S(x)\},$$

where B is a relatively slowly varying function and

$$S(x) = [\phi(\mathbf{p}) - p_\lambda x^\lambda]_{\mathbf{p}=\mathbf{p}_s(x)}. \tag{100}$$

The differential of (100) is

$$dS(x) = -(p_s)_\lambda dx^\lambda.$$

Whence $\partial S(x)/\partial x^\lambda = -(p_s)_\lambda$, which substituted into the identity $p_s^0 = (\mathbf{p}_s^2 + m^2c^2)^{1/2}$ leads to the Hamilton–Jacobi equation

$$\frac{\partial S}{\partial x^0} + \left[\left(\frac{\partial S}{\partial \mathbf{x}}\right)^2 + m^2c^2\right]^{1/2} = 0.$$

Thus a wave packet of quasiclassical form, whose dimensions tend to zero with $\hbar$, will follow the classical trajectory of a free particle.

IV. ELECTROMAGNETIC INTERACTIONS

A. Classical treatment of particle and antiparticle

When an electromagnetic field derived from a potential $A^\lambda(x^\kappa) = A^\lambda(z^\kappa + y^\kappa)$ is present and the particle has charge e, the classical momentum $p^\lambda(\mathbf{y},\boldsymbol{\pi})$ is given by (8). This is obtained from the noninteracting form (7) by the prescription $p^\lambda \to p^\lambda - (e/c)A^\lambda$, $\boldsymbol{\pi} \to \boldsymbol{\pi}_E \equiv \boldsymbol{\pi} - (e/c)(\mathbf{A} + \hat{\mathbf{y}}A^0)$. Note the identity[1]

$$\mathbf{y}\cdot\boldsymbol{\pi}_E = -mcy_\lambda w^\lambda, \tag{101}$$

where w^λ is the four-velocity of the particle at the point where the past light cone with vertex z^λ intersects the particle trajectory. Thus (101) implies that $\mathbf{y}\cdot\boldsymbol{\pi}_E$ is Lorentz invariant and non-negative. To make the treatment more like that for the noninteracting case, let us effect a canonical transformation $\mathbf{y},\boldsymbol{\pi} \to \mathbf{y}',\boldsymbol{\pi}'$, which converts the relation $\mathbf{y}\cdot\boldsymbol{\pi}_E \geqslant 0$ into $\mathbf{y}'\cdot\boldsymbol{\pi}' \geqslant 0$. Consider the generating function $F(z^\lambda,\mathbf{y},\boldsymbol{\pi}') = \mathbf{y}\cdot\boldsymbol{\pi}' + (e/c)\chi(z^\lambda,\mathbf{y})$, where $\chi(z^\lambda,\mathbf{y})$ is to be chosen to make $\mathbf{y}\cdot\boldsymbol{\pi}_E = \mathbf{y}'\cdot\boldsymbol{\pi}'$. We have

$$\mathbf{y}' = \frac{\partial F}{\partial\boldsymbol{\pi}'} = \mathbf{y}, \quad \boldsymbol{\pi} = \frac{\partial F}{\partial\mathbf{y}} = \boldsymbol{\pi}' + \frac{e}{c}\frac{\partial\chi}{\partial\mathbf{y}},$$
$$p'_\lambda = p_\lambda + \frac{\partial F}{\partial z^\lambda} = p_\lambda + \frac{e}{c}\frac{\partial\chi}{\partial z^\lambda}. \tag{102}$$

Then $\mathbf{y}\cdot\boldsymbol{\pi}_E = \mathbf{y}\cdot\boldsymbol{\pi}'$ provided χ is any solution of

$$\mathbf{y}\cdot\frac{\partial\chi}{\partial\mathbf{y}} = -y_\kappa A^\kappa. \tag{103}$$

The general solution of (103) is

$$\chi(z^\lambda,\mathbf{y}) = -\int_\gamma^y \frac{y_\kappa}{y}A^\kappa\left(z^\lambda + \frac{ty^\lambda}{y}\right)dt, \tag{104}$$

where γ is an arbitrary function of z^λ and $\hat{\mathbf{y}}$. The theory presented in what follows does not depend on the choice of this function γ. [See (114) and (121).] The transformed evolution generator is then given by

$$p^0(\mathbf{y},\boldsymbol{\pi}') = \frac{e}{c}\left(A^0 + \frac{\partial\chi}{\partial z^0}\right) + \tfrac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi}')^{-1}[(\boldsymbol{\pi}'_E)^2 + m^2c^2]y,$$
$$\mathbf{p}(\mathbf{y},\boldsymbol{\pi}') = \frac{e}{c}\left(A - \frac{\partial\chi}{\partial z}\right) + \boldsymbol{\pi}'_E - \tfrac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi}')^{-1}[(\boldsymbol{\pi}'_E)^2 + m^2c^2]\mathbf{y}, \tag{105}$$
$$\boldsymbol{\pi}'_E = \boldsymbol{\pi}' - \frac{e}{c}\left(\mathbf{A} + \hat{\mathbf{y}}A^0 - \frac{\partial\chi}{\partial\mathbf{y}}\right).$$

The above does *not* represent an electromagnetic gauge transformation because $\chi(z^\lambda,\mathbf{y})$ does not depend on its arguments solely through the combination $z^\lambda + y^\lambda$.

A like treatment may be given for an antiparticle, with $e \to -e$ in all expressions. Thus we could treat particle and antiparticle as two disjoint systems with conjugate variables $\mathbf{y}_p,\boldsymbol{\pi}_p$ and $\mathbf{y}_a,\boldsymbol{\pi}_a$, respectively, with $\mathbf{y}_p\cdot\boldsymbol{\pi}_p \geqslant 0$ and $\mathbf{y}_a\cdot\boldsymbol{\pi}_a \geqslant 0$. The evolution generators $p_p^\lambda(\mathbf{y}_p,\boldsymbol{\pi}_p)$ and $p_a^\lambda(\mathbf{y}_a,\boldsymbol{\pi}_a)$ would then be given by (105) with $\mathbf{y} \to \mathbf{y}_p$, $\boldsymbol{\pi}' \to \boldsymbol{\pi}_p$, and $\mathbf{y} \to \mathbf{y}_a$, $\boldsymbol{\pi}' \to \boldsymbol{\pi}_a$, $e \to -e$, respectively. However, just as in the noninteracting case, particle and antiparticle may be treated more elegantly as a single system, with the variables $\mathbf{y},\boldsymbol{\pi}'$ now being in an enlarged phase space allowing both signs of $\mathbf{y}\cdot\boldsymbol{\pi}'$. Henceforth let us drop the prime and write $\boldsymbol{\pi}$ instead of $\boldsymbol{\pi}'$. The particle–antiparticle system then has the following classical description of its dynamics. The conjugate variables $\mathbf{y},\boldsymbol{\pi}$

each vary over the whole of $\mathbb{R}^3$. As we change the here-now z^λ from which the past light cone is drawn, the variables $\mathbf{y}$ and $\boldsymbol{\pi}$ evolve according to

$$\frac{\partial \mathbf{y}}{\partial z^\lambda} = \{\mathbf{y}, P_\lambda\}, \quad \frac{\partial \boldsymbol{\pi}}{\partial z^\lambda} = \{\boldsymbol{\pi}, P_\lambda\}, \tag{106}$$

where P^λ is given by

$$P^0 = \frac{e}{c}\left(A^0 + \frac{\partial \chi}{\partial z^0}\right) + \frac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi})^{-1}(\boldsymbol{\pi}_E^2 + m^2c^2)\, y,$$

$$\mathbf{P} = \frac{e}{c}\left(\mathbf{A} + \frac{\partial \chi}{\partial \mathbf{z}}\right) + \boldsymbol{\pi}_E - \frac{1}{2}(\mathbf{y}\cdot\boldsymbol{\pi})^{-1}(\boldsymbol{\pi}_E^2 + m^2c^2)\, \mathbf{y},$$

$$\boldsymbol{\pi}_E = \boldsymbol{\pi} - \frac{e}{c}\left(\mathbf{A} + \hat{\mathbf{y}}A^0 - \frac{\partial \chi}{\partial \mathbf{y}}\right). \tag{107}$$

If $\mathbf{y}\cdot\boldsymbol{\pi} > 0$, we interpret $\mathbf{y}_p = \mathbf{y}$, $\boldsymbol{\pi}_p = \boldsymbol{\pi}$ as the conjugate variables of a particle with $p_p^\lambda = P^\lambda$, and if $\mathbf{y}\cdot\boldsymbol{\pi} < 0$, $\mathbf{y}_a = \mathbf{y}$, $\boldsymbol{\pi}_a = -\boldsymbol{\pi}$ as those of an antiparticle with $p_a^\lambda = -P^\lambda$. In each case $p^\lambda = \mathrm{sgn}(\mathbf{y}\cdot\boldsymbol{\pi})P^\lambda$.

Equation (107) gives P^λ as a quadratic function of e/c. We may write (107) in the form

$$P^\lambda = P^\lambda_{(0)} + (e/c)P^\lambda_{(1)} + (e/c)^2 P^\lambda_{(2)}, \tag{108}$$

where $P^\lambda_{(0)}$ is given by (7), $P^\lambda_{(1)}$ is linear in A_κ and χ and $P^\lambda_{(2)}$ is quadratic in these fields. Explicitly,

$$P^\lambda_{(1)} = \frac{\partial \chi}{\partial z_\lambda} + \{\chi, P^\lambda_{(0)}\} - (\mathbf{y}\cdot\boldsymbol{\pi})^{-1}P^\kappa_{(0)}A_\kappa y^\lambda,$$

$$P^\lambda_{(2)} = (2\mathbf{y}\cdot\boldsymbol{\pi})^{-1}\left[(A^0)^2 - \left(\mathbf{A} - \frac{\partial \chi}{\partial \mathbf{y}}\right)^2\right]y^\lambda. \tag{109}$$

[The transition from (107) to (109) is facilitated by use of the identities $\{P^\lambda_{(0)}, y^\mu\} = g^{\lambda\mu} + (\mathbf{y}\cdot\boldsymbol{\pi})^{-1}y^\lambda P^\mu_{(0)}$ and $y_\lambda P^\lambda_{(0)} = -\mathbf{y}\cdot\boldsymbol{\pi}$.]

B. Invariances of the classical momentum function

The system (106) and (107) has a number of symmetries which one would expect to be preserved upon quantization. These are summarized in (110)–(114) below, and may be verified by direct computation. The evolution generator is given by (107) as a function of $\mathbf{y},\boldsymbol{\pi},e,A_\kappa$, and χ, which dependence will be indicated when necessary by the notation $P^\lambda(\mathbf{y},\boldsymbol{\pi},e,A_\kappa,\chi)$. Recall that A_κ is a function of $z^\lambda + y^\lambda$, and χ any function of z^λ and $\mathbf{y}$ which satisfies (103).

Invariance of the sign $\mathbf{y}\cdot\boldsymbol{\pi}$ means

$$\{\Theta(\mathbf{y}\cdot\boldsymbol{\pi}), P^\lambda\} = 0. \tag{110}$$

Self-consistency of the evolution equation (106) means

$$\frac{\partial P_\lambda}{\partial z^\mu} - \frac{\partial P_\mu}{\partial z^\lambda} + \{P_\lambda, P_\mu\} = 0. \tag{111}$$

Charge conjugation invariance means

$$P^\lambda(\mathbf{y},\boldsymbol{\pi},e,A_\kappa,\chi) = -P^\lambda(\mathbf{y},-\boldsymbol{\pi},-e,A_\kappa,\chi). \tag{112}$$

Electromagnetic gauge invariance means, for any $\Lambda(z^\lambda + y^\lambda)$,

$$P^\lambda(\mathbf{y},\boldsymbol{\pi},e,A_\kappa,\chi) = P^\lambda\left(\mathbf{y},\boldsymbol{\pi},e,A_\kappa + \frac{\partial \Lambda}{\partial z^\kappa}, \chi - \Lambda\right). \tag{113}$$

Invariance under changing χ by an arbitrary function $M(z^\lambda,\hat{\mathbf{y}})$ means

$$P_\lambda(\mathbf{y},\boldsymbol{\pi},e,A_\kappa,\chi) = P_\lambda\left(\mathbf{y},\boldsymbol{\pi} + \frac{e}{c}\frac{\partial M}{\partial \mathbf{y}}, e, A_\kappa, \chi - M\right) + \frac{e}{c}\frac{\partial M}{\partial z^\lambda}. \tag{114}$$

C. Quantization

Our aim is to find an operator P^λ_{OP} acting in $\mathscr{H}_y$ which is a quantum analog of the classical P^λ given by (107), and which governs the evolution of $\psi(\mathbf{y}) \in \mathscr{H}_y$ according to

$$i\hbar\frac{\partial \psi}{\partial z^\lambda} = P_{\mathrm{OP}\lambda}\psi. \tag{115}$$

We shall assume that the same relation holds between the Hilbert space $\mathscr{H}_y$ and $\mathscr{H}_{\mathrm{phys}}$ as applied without interactions [see (57)–(65)]. Thus particle and antiparticle are again associated, respectively, with the positive and negative eigenspaces $\mathscr{H}_y^+, \mathscr{H}_y^-$ of the operator $\Sigma = -i(\mathbf{y}\cdot\partial/\partial\mathbf{y} + 1)$. This hypothesis makes sense when the classical variables $\mathbf{y}$ and $\boldsymbol{\pi}$ [designated $\boldsymbol{\pi}'$ in (102)] are chosen as in Sec. IV A, because then the sign of $\mathbf{y}\cdot\boldsymbol{\pi}$ determines the particle's charge, just as was the case without electromagnetic interactions.

Analogy with (108) suggests that P^λ_{OP} should be quadratic in e/c:

$$P^\lambda_{\mathrm{OP}} = P^\lambda_{(0)} + (e/c)P^\lambda_{(1)} + (e/c)^2 P^\lambda_{(2)}. \tag{116}$$

Here $P^\lambda_{(0)}$ is the operator defined by (87) and (88), and the Hermitian operators $P^\lambda_{(1)}$ and $P^\lambda_{(2)}$ are functionals of A_κ and χ which are, respectively, linear and homogeneous of degree 2. Let us henceforth omit the suffix OP from P^λ_{OP}, it being understood that P^λ now represents an operator on $\mathscr{H}_y$. This operator is a function of z^μ and e, and a functional of the fields A_κ and χ, which dependence will be indicated when needed by the notation $P^\lambda(z^\mu,e;A_\kappa;\chi)$.

In addition to requiring that P^λ should correspond in the classical limit to the classical evolution generator specified by (107), we shall demand that the quantum counterparts of the classical symmetries (110)–(114) shall hold, namely

$$[\Theta(\Sigma), P^\lambda] = 0, \tag{117}$$

$$i\hbar\left(\frac{\partial P_\lambda}{\partial z^\mu} - \frac{\partial P_\mu}{\partial z^\lambda}\right) + [P_\lambda, P_\mu] = 0, \tag{118}$$

$$[P^\lambda(z^\mu,e;A_\kappa;\chi)\psi]^* = -P^\lambda(z^\mu,-e;A_\kappa;\chi)\psi^*, \tag{119}$$

$$P^\lambda(z^\mu,e;A_\kappa;\chi) = P^\lambda\left(z^\mu,e;A_\kappa + \frac{\partial \Lambda}{\partial z^\kappa};\chi - \Lambda\right), \tag{120}$$

$$P_\lambda(z^\mu,e;A_\kappa;\chi) = \exp\left(-\frac{ie}{\hbar c}M\right)P_\lambda(z^\mu,e;A_\kappa;\chi - M) \times \exp\left(\frac{ie}{\hbar c}M\right) + \frac{e}{c}\frac{\partial M}{\partial z^\lambda}. \tag{121}$$

A consequence of (117) is that $\mathscr{H}_y^+$ and $\mathscr{H}_y^-$ [see (57)–(61)] are invariant subspaces of P^λ. This implies that if $\psi \in \mathscr{H}_y$ evolves according to (115), then there exists an operator p^λ on $\mathscr{H}_{\mathrm{phys}}$ such that Ψ, the image in $\mathscr{H}_{\mathrm{phys}}$ of ψ under the mapping (65), satisfies

$$i\hbar\frac{\partial\Psi}{\partial z^\lambda}=p_\lambda\Psi. \tag{122}$$

Explicitly,

$$p_\lambda\begin{bmatrix}\psi_p\\ \psi_a\end{bmatrix}=\begin{bmatrix}P_\lambda(z^\mu,e;A_\kappa;\chi)\psi_p\\ P_\lambda(z^\mu,-e;A_\kappa;\chi)\psi_a\end{bmatrix}, \tag{123}$$

where use has been made of (119).

The consistency condition (118) implies the existence of a unitary operator $U(z)\equiv U(z,e;A_\kappa;\chi)$, a function of z^λ and e and a functional of A_κ and χ, such that

$$P_\lambda=i\hbar\frac{\partial U(z)}{\partial z^\lambda}U^\dagger(z). \tag{124}$$

The wave function ψ at the current here-now z^λ is related to its value ψ_0 at an initial here-now z_0^λ by

$$\psi=U(z)U^\dagger(z_0)\psi_0. \tag{125}$$

Note that if V is any unitary operator which does not depend on z^λ, then $U(z)$ and $U(z)V$ give the same results in (124) and (125), which equivalence will be denoted

$$U(z)V\cong U(z).$$

The symmetries (119) to (121) now become

$$\begin{aligned}&[U(z,e;A_\kappa;\chi)\psi]^*=U(z,-e;A_\kappa;\chi)\psi^*,\\ &U(z,e;A_\kappa;\chi)\cong U\left(z,e;A_\kappa+\frac{\partial\Lambda}{\partial z^\kappa};\chi-\Lambda\right),\\ &U(z,e;A_\kappa;\chi)\cong\exp\left(\frac{e}{i\hbar c}M\right)U(z,e;A_\kappa;\chi-M).\end{aligned} \tag{126}$$

The author has been unable to obtain any solution of (117) to (121), or equivalently of (126), for which P^λ takes the quadratic form (116). However a solution to first order in e/c may be obtained by writing

$$\begin{aligned}U(z)=&\exp\left[\frac{e}{i\hbar c}\left(H_1+\frac{e}{c}H_2+\cdots\right)\right]\\ &\times\exp\left[\frac{z_\lambda P^\lambda(0)}{i\hbar}\right],\end{aligned} \tag{127}$$

and expanding the first factor as a power series in e/c, keeping only the leading terms. Equating terms in e/c in (124) yields

$$\frac{\partial H_1}{\partial z^\lambda}+\frac{1}{i\hbar}[H_1,P_{(0)\lambda}]=P_{(1)\lambda}. \tag{128}$$

The classical counterpart of (128) is

$$\begin{aligned}&\frac{\partial H_{\rm Cl}}{\partial z^\lambda}+\{H_{\rm Cl},P_{(0)\lambda}\}=P_{(1)\lambda},\\ &\qquad=\frac{\partial\chi}{\partial z^\lambda}+\{\chi,P_{(0)\lambda}\}-\frac{P^\kappa_{(0)}A_\kappa y_\lambda}{\mathbf{y}\cdot\boldsymbol{\pi}},\end{aligned} \tag{129}$$

upon using (109) [$P_{(0)\lambda}$ and $P_{(1)\lambda}$ are now functions of $\mathbf{y},\boldsymbol{\pi}$]. A solution of (129) is

$$\begin{aligned}H_{\rm Cl}=&\chi+\frac{1}{2}\int_0^\infty P^\kappa_{(0)}[A_\kappa(z^\lambda+y^\lambda-tP^\lambda_{(0)})\\ &-A_\kappa(z^\lambda+y^\lambda+tP^\lambda_{(0)})]dt,\end{aligned} \tag{130}$$

which may be verified by direct substitution. We now seek a Hermitian operator H_1 on $\mathscr{H}_y$ which is an analog of (130), and further, is consistent to first order in (e/c) with the symmetries (117) to (121). Let us specify H_1 by its matrix elements with respect to ψ_k and ψ_k^*, the complete orthonormal set of eigenfunctions of $P^\lambda_{(0)}$ defined by (88). An operator H_1 which satisfies all the required constraints is given by

$$\begin{aligned}(\psi_k,H_1\psi_{k'})_y=&\int\frac{d^3\mathbf{y}}{y}\psi_k^*(\mathbf{y})\psi_{k'}(\mathbf{y})\\ &\times\left\{\chi+\frac{1}{2}\int_0^\infty K^\kappa[A_\kappa(z^\lambda-y^\lambda-tK^\lambda)\right.\\ &\left.-A_\kappa(z^\lambda-y^\lambda+tK^\lambda)]dt\right\},\end{aligned} \tag{131}$$

$$K_\kappa=k_\kappa+k'_\kappa,$$

$$(\psi_k^*,H_1\psi_{k'}^*)_y=(\psi_k,H_1\psi_{k'})^*,\quad(\psi_k^*,H_1\psi_{k'})=0.$$

This form was found by trial and error, guided by the classical result (130). Finally, the matrix elements of $P_{(1)\lambda}$ may be calculated from (128):

$$\begin{aligned}&(\psi_k,P_{(1)\lambda}\psi_{k'})_y=\left[\frac{\partial}{\partial z^\lambda}+i(k_\lambda-k'_\lambda)\right](\psi_k,H_1\psi_{k'})_y,\\ &(\psi_k^*,P_{(1)\lambda}\psi_{k'}^*)_y=(\psi_k,P_{(1)\lambda}\psi_{k'})_y^*,\\ &(\psi_k^*,P_{(1)\lambda}\psi_{k'})_y=0.\end{aligned} \tag{132}$$

No satisfactory way of extending these results to second order in e/c has been found.

V. DISCUSSION

A. Summary

In the previous sections we have arrived at the following past light cone quantum picture of a charged boson of spin-zero and nonzero rest mass. An observer at here-now z^λ ascribes to the particle an SO(1,3) scalar wave function $\psi(\mathbf{y},z^\lambda)$, which belongs to the Hilbert space $\mathscr{H}_y$ defined by (10). As z^λ varies the wave function evolves according to

$$i\hbar\frac{\partial\psi}{\partial z^\lambda}=P_\lambda\psi, \tag{133}$$

where P_λ is a Hermitian operator in $\mathscr{H}_y$. The wave function describes the quantum state on the past light cone with vertex at z^λ, i.e., on the three-surface $x^\lambda=z^\lambda+y^\lambda$, with $|\psi(\mathbf{y},z^\lambda)|d^3\mathbf{y}/y$ representing the probability that a position measurement will locate the particle on the light cone between $\mathbf{y}$ and $\mathbf{y}+d\mathbf{y}$ (ψ being normalized to unity).

An important role is played by the operator $\Sigma=-i(\mathbf{y}\cdot\partial/\partial\mathbf{y}+1)$ and the associated projection operator $\Theta(\Sigma)$. Particle and antiparticle states are eigenfunctions of $\Theta(\Sigma)$ with eigenvalues 1 and 0, respectively, so that we can resolve any wave function into a particle component and an antiparticle component according to the prescription (57). However, the Hilbert space of physical states $\mathscr{H}_{\rm phys}$ is *not* identified with $\mathscr{H}_y$, but is rather constructed by combining the particle amplitude with the complex conjugate of the antiparticle amplitude. [See (62)–(64).]

The evolution operator P_λ must satisfy the conditions (118) in order that (133) be self-consistent. For a free particle (88) gives a complete orthonormal set of eigenfunctions

of P_λ, which correspond to energy-momentum eigenstates in $\mathscr{H}_{\text{phys}}$. When an external electromagnetic field is present the theory is less complete. In this case (131) and (132) specify P_λ up to first order in the particle charge.

B. Alternative hypotheses

There are a number of places in the development of the theory where a hypothesis different from the one adopted might have been made. Particular examples follow.

(1) The probability density for finding the particle might be taken as $(|\psi_p|^2 + \psi_a|^2)/y$ rather than as $|\psi|^2/y$. [See (65).] Note that these quantities are not the same, though their integrals over all $\mathbf{y}$ are equal.

(2) Perhaps neither of these expressions for the probability density is valid and one should seek instead to find a set of position eigenfunctions $\Psi_Y(\mathbf{y})$ which span $\mathscr{H}_{\text{phys}}$. These should be labeled by a past light cone vector Y^λ with $Y^0 = -|\mathbf{Y}|$, and satisfy the orthogonality and completeness relations

$$(\Psi_Y,\Psi_{Y'})_{\text{phys}} = Y\delta(\mathbf{Y}-\mathbf{Y}'),$$
$$\int \frac{d^3\mathbf{Y}}{Y}\,\Psi_Y(\mathbf{y})\Psi_Y^\dagger(\mathbf{y}') = \langle \mathbf{y}|\Theta(\Sigma)|\mathbf{y}'\rangle \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{134}$$

We can then define a position operator in $\mathscr{H}_{\text{phys}}$ by

$$Y^\lambda_{\text{phys}}\Psi(\mathbf{y}) = \int \frac{d^3\mathbf{Y}}{Y}\, Y^\lambda \Psi_Y(\mathbf{y})(\Psi_Y,\Psi)_{\text{phys}},$$

and interpret $|(\Psi_Y,\Psi)_{\text{phys}}|^2\, d^3\mathbf{Y}/Y$ as the probability of finding the light cone position three-vector $\mathbf{y}$ in the range $\mathbf{Y}$ to $\mathbf{Y}+d\mathbf{Y}$. Note that the operator $\mathbf{y}_{\text{OP}}$ of (11) acts in $\mathscr{H}_y$, not in $\mathscr{H}_{\text{phys}}$. A possible candidate for $\Psi_Y(\mathbf{y})$, which satisfies (134), is

$$\Psi_Y(\mathbf{y}) = \int dS_k\, \psi_k(\mathbf{y}) \begin{bmatrix} \psi_k^*(\mathbf{Y}) \\ \psi_k(\mathbf{Y}) \end{bmatrix},$$

$$\equiv \Theta(\Sigma) \begin{bmatrix} y\delta(\mathbf{y}-\mathbf{Y}) \\ \dfrac{\kappa^2}{4\pi} J_0([2\kappa^2 y_\lambda Y^\lambda]^{1/2}) \end{bmatrix}.$$

[See (58), (60), (70), (91), (92).] With this choice, the effective coordinate space wave function is

$$\psi_{\text{phys}}(\mathbf{Y}) = (\Psi_Y,\psi)_{\text{phys}} = \psi_p(\mathbf{Y}) + W\psi_a(\mathbf{Y}), \tag{135}$$

where W is the unitary operator defined by (93).

(3) One might identify $\mathscr{H}_y$ directly with the space of physical states $\mathscr{H}_{\text{phys}}$. In this case the momentum operator would be $\Theta(\Sigma)P_\lambda$, the charge operator would be $\text{sgn}(\Sigma)$, and $\mathbf{y}_{\text{OP}}$ given by (11) would be acceptable as the light cone position operator. Such a theory would in fact be equivalent with that of (2) above, with the correspondence $\psi \to \psi_{\text{phys}}$ [see (135)]. This theory would be more akin to that of the nonrelativistic Schrödinger equation rather than that of the Dirac and Klein–Gordon equations.

(4) We could make a choice of phase function $g(\sigma)$ other than that of (88). It would be seen from the arguments of Secs. III E and III F that any $g(\sigma)$ which behaves asymptotically like $-\sigma\log\sigma + \sigma + O(1)$ for large σ will lead to the correct nonrelativistic and to the correct classical limits. An example of such a $g(\sigma)$ is given by (B5) in Appendix B. The price to pay if (88) is not adopted is that the operator W effecting the symmetry (91) no longer takes a simple form.

C. Some unsolved problems

This paper leaves many questions unanswered, particularly those relating to the measurement of observables. Suppose the quantum state of an observer at the here-now z^λ is $|\Psi\rangle$, and he measures the observable represented by a Hermitian operator $\mathscr{L}$ which has a complete orthonormal set of eigenstates $|\Phi_n\rangle$ with corresponding eigenvalues λ_n. One would like to postulate that $|\langle\Phi_n|\Psi\rangle|^2$ is the probability that the measurement will yield λ_n, and that if this eigenvalue is obtained, then immediately after the measurement the state will be $|\Phi_n\rangle$. However, how are we to interpret the phrase "a measurement at here-now z^λ?" If the measurement occupies negligible extension in space and time so that it can be carried out by a single observer at z^λ then there is no difficulty. Otherwise the observer at z^λ must collaborate with auxiliary observers in other parts of space time. Various possibilities suggest themselves: (1) a prearranged experiment, carried out by a number of observers on the past light cone with vertex at z^λ; (2) an unpremediated experiment, where the observer at z^λ radios instructions to his collaborators to measure some quantity in their locality, in which case the experiment involves a number of observers on the future light cone of z^λ; and (3) independent measurements made by different observers on a spacelike three-surface through z^λ, the results being communicated to the central observer.

Type (1) experiments comply most closely with the spirit of the present formalism, and have the advantage that all the data from the auxiliary observers reaches the central observer simultaneously at z^λ. However, the experiment can only involve a finite region of the past light cone and the central observer must plan the experiment sufficiently in advance to enable him to instruct his collaborators.

With experiments of type (2) and (3) the experimental results from the auxiliary observers will reach the central observer at different times, all in the future of z^λ. As each new piece of data arrives the central observer should presumably update his quantum state, but we lack a prescription for doing this. It seems likely that we shall need a description in terms of density matrices rather than of pure states to resolve this problem. That a solution to the problem of measurement in relativistic quantum theory may require the introduction of density matrices has been suggested by Houtappel, Van Dam, and Wigner.[10]

APPENDIX A: ORTHOGONALITY AND COMPLETENESS RELATIONS

Proof of (37)–(39): Since k^λ and $k^{\lambda'}$ both lie on the mass shell (14), $k^\lambda + k^{\lambda'}$ is future pointing timelike and there exists a "center of momentum" inertial frame in which this sum vector has no spatial component, i.e., $k^\lambda = (\epsilon_k, \mathbf{k})$ and $k^{\lambda'} = (\epsilon_k, -\mathbf{k})$. In this frame (37) becomes

$$(u_k,\Sigma u_{k'})_y = \tfrac{1}{2}(u_k,\Sigma u_{k'})_y + \tfrac{1}{2}(\Sigma u_k,u_{k'})_y$$
$$= (2\pi)^{-3}\int \epsilon_k e^{-2i\mathbf{k}\cdot\mathbf{y}}\, d^3\mathbf{y}$$
$$= \epsilon_k\delta(2\mathbf{k}) = \epsilon_k\delta(\mathbf{k}-\mathbf{k}'). \tag{A1}$$

Since each side of (A1) is an SO(1,3) scalar, the truth of this equation in any particular inertial frame implies its truth in all inertial frames, and hence (37) has been proved. Similar proofs may be given for (38) and (39).

Proof of (40): The left-hand side of (40) takes the form

$$\Sigma(2\pi)^{-3/2}\int dS_k\left[e^{-ik^\lambda(y_\lambda - y'_\lambda)} - e^{ik^\lambda(y_\lambda - y'_\lambda)}\right]$$
$$= 2i\Sigma\Delta(y_\lambda - y'_\lambda). \tag{A2}$$

The function Δ occurs in quantum field theory in connection with spin-zero field commutators, and may be expressed in terms of Bessel functions.[11] Putting in the explicit form for Δ and applying the differential operator Σ then yields (40).

Proof of (50) and (51): The Beltrami operator belonging to the mass shell (14) is proportional to the operator

$$\mathscr{K} = \tfrac{1}{2}\hbar^{-2}\eta_{\lambda\nu}\eta_{\mu\rho}j_k^{\lambda\mu}j_k^{\rho\nu},$$
$$= \left(\mathbf{k}\times\frac{\partial}{\partial\mathbf{k}}\right)^2 - \left(\epsilon_k\frac{\partial}{\partial\mathbf{k}}\right)^2. \tag{A3}$$

[See (22) and (23).] Direct computation shows that $(-k^\lambda y_\lambda)^{i\sigma-1}$ is an eigenfunction of $\mathscr{K}$ with eigenvalues $1+\sigma^2$, if y^λ is any vector lying on the past light cone. Now $\mathscr{K}$ is Hermitian with respect to the measure $dS_k = d^3\mathbf{k}/\epsilon_k$, which implies orthogonality for eigenfunctions belonging to different values of σ^2. Hence

$$\int dS_k(-k^\lambda y_\lambda)^{i\sigma-1}(-k^\mu y'_\mu)^{-i\sigma'-1}$$
$$= C_1\delta(\sigma-\sigma') + C_2\delta(\sigma+\sigma'), \tag{A4}$$

where C_1 and C_2 are functions of σ, y_λ, and y'_λ which are SO(1,3) scalars. A scalar function must yield zero when operated upon by any component of the appropriate total angular momentum tensor. Thus

$$\left(\mathbf{y}\times\frac{\partial}{\partial\mathbf{y}} + \mathbf{y}'\times\frac{\partial}{\partial\mathbf{y}'}\right)C_s = 0,$$
$$\left(y\frac{\partial}{\partial\mathbf{y}} + y'\frac{\partial}{\partial\mathbf{y}'}\right)C_s = 0, \tag{A5}$$

with $s = 1,2$. [See (76).] The general solution of (A5) is readily found to be

$$C_s = \alpha_s(\sigma, y^\lambda y'_\lambda) + (yy')^{-1}\beta_s(\sigma, y/y')\delta(\hat{\mathbf{y}},\hat{\mathbf{y}}'), \tag{A6}$$

where α_s and β_s are arbitrary functions of their arguments, and $\delta(\hat{\mathbf{y}},\hat{\mathbf{y}}')$ is the Dirac delta function for the surface of a unit sphere. Multiplying (A4) by $y^{-i\sigma+1}(y')^{i\sigma'+1}$ yields

$$\int dS_k(\epsilon_k + \mathbf{k}\cdot\hat{\mathbf{y}})^{i\sigma-1}(\epsilon_k + \mathbf{k}\cdot\hat{\mathbf{y}}')^{-i\sigma'-1}$$
$$= yy'(y/y')^{-i\sigma}C_1\delta(\sigma-\sigma')$$
$$+ (yy')^{-i\sigma+1}C_2\delta(\sigma+\sigma'). \tag{A7}$$

The left-hand side of (A7) is independent of y and y', and hence so must be the right-hand side. The only way in which this independence can be consistent with (A6) is for

$$\alpha_1 = 0, \quad \beta_1 = \beta(y/y')^{i\sigma},$$
$$\alpha_2 = \alpha(y^\lambda y'_\lambda)^{i\sigma-1}, \quad \beta_2 = 0, \tag{A8}$$

where α and β are functions of σ only. To find these functions, apply the integration $\int d^2\hat{\mathbf{y}}'$ over the unit sphere, to both sides of (A7). The left-hand side yields

$$8\pi^3(\sigma\sigma')^{-1}\kappa^{i\sigma}\kappa^{-i\sigma'}[\delta(\sigma-\sigma) - \delta(\sigma+\sigma')],$$

and the right-hand side

$$\beta\delta(\sigma-\sigma') - 2\pi i\alpha\sigma^{-1}2^{i\sigma}\delta(\sigma+\sigma').$$

Identification of the coefficients of the two delta functions then gives $\beta = 8\pi^3\sigma^{-2}$, $\alpha = 4\pi^2 i\sigma^{-1}(\kappa^2/2)^{i\sigma}$. Thus

$$\int dS_k(\epsilon_k + \mathbf{k}\cdot\hat{\mathbf{y}})^{i\sigma-1}(\epsilon_k + \mathbf{k}\cdot\hat{\mathbf{y}}')^{-i\sigma'-1}$$
$$= 8\pi^3\sigma^{-2}\delta(\hat{\mathbf{y}},\hat{\mathbf{y}}')\delta(\sigma-\sigma') \tag{A9}$$
$$+ 4\pi^2 i\sigma^{-1}(\kappa^2/2)^{i\sigma}(1-\hat{\mathbf{y}}\cdot\hat{\mathbf{y}}')^{i\sigma-1}\delta(\sigma+\sigma').$$

Multiplying (A9) by $\sigma\sigma'/(16\pi^3)$ and making the identification $\mathbf{w} = \hat{\mathbf{y}}$, $\mathbf{w}' = \hat{\mathbf{y}}'$ then yields (50) and (51).

APPENDIX B: COMPLETE ORTHONORMAL SETS IN $\mathscr{H}_y$

In Sec. III C we considered function sets of the form

$$\psi_k(\mathbf{y}) = (2\pi)^{-2}i\int_0^\infty d\sigma\,\sigma e^{ig(\sigma)}\zeta^{i\sigma-1}, \tag{B1}$$

where $\zeta = -k_\lambda y^\lambda$. These functions belong to the eigenvalue $+1$ of the operator $\Theta(\Sigma)$ defined by (58). It was shown that for any choice of the real function $g(\sigma)$ the functions $\psi_k(\mathbf{y})$ and $\psi_k^*(\mathbf{y})$ form a complete orthonormal set in $\mathscr{H}_y$:

$$(\psi_k,\psi_{k'})_y = (\psi_k^*,\psi_{k'}^*)_y = \epsilon_k\delta(\mathbf{k}-\mathbf{k}'),$$
$$(\psi_k^*,\psi_{k'})_y = 0, \tag{B2}$$
$$\int dS_k[\psi_k(\mathbf{y})\psi_k^*(\mathbf{y}') + \psi_k^*(\mathbf{y})\psi_k(\mathbf{y}')] = y\delta(\mathbf{y}-\mathbf{y}').$$

We exhibit here the form taken by $\psi_k(\mathbf{y})$ for three particular choices for $g(\sigma)$ (the case $g(\sigma) = \arg(\Gamma(-i\sigma)]$ was discussed in Sec. III D):

$$e^{ig(\sigma)} = 1,$$
$$\psi_k(\mathbf{y}) = [4\pi^2 i\zeta(\log\zeta + i\epsilon)^2]^{-1}, \quad \epsilon\to 0, \tag{B3}$$
$$= \Theta(\Sigma)(2\pi)^{-1}\frac{d}{d\zeta}\delta(\zeta-1);$$
$$e^{ig(\sigma)} = -\Gamma(-i\sigma)/\Gamma(i\sigma),$$
$$\psi_k(\mathbf{y}) = \Theta(\Sigma)(2\pi)^{-1}J_0(2\zeta^{1/2}); \tag{B4}$$
$$e^{ig(\sigma)} = \pi^{-1/2}\Gamma(\tfrac{1}{2} - i\sigma)[\cosh(\tfrac{1}{2}\pi\sigma) - i\sinh(\tfrac{1}{2}\pi\sigma)],$$
$$\psi_k(\mathbf{y}) = \Theta(\Sigma)(2\pi^3)^{-1/2}\frac{d}{d\zeta}(\zeta^{1/2}\sin\zeta). \tag{B5}$$

If we drop the requirement that our functions be eigenfunctions of $\Theta(\Sigma)$, then we can find further orthonormal, complete sets from a generalization of the theorem given by (82). The more general result is that any set of functions $v_{kq}(\mathbf{y})$, $q = 1$ or 2, of the form

$$v_{k1}(\mathbf{y}) = (2\pi)^{-2} i \int_{-\infty}^{\infty} d\sigma\, v(\sigma) \zeta^{i\sigma - 1},$$

$$v_{k2}(\mathbf{y}) = \left[(2\pi)^{-2} i \int_{-\infty}^{\infty} d\sigma\, \mathrm{sgn}(\sigma) v(\sigma) \zeta^{i\sigma - 1} \right]^*, \quad \text{(B6)}$$

where $v(\sigma)$ satisfies

$$|v(\sigma)|^2 + |v(-\sigma)|^2 = \sigma^2,$$

are orthonormal and complete

$$(v_{kq}, v_{k'q'})_y = \epsilon_k \delta(\mathbf{k} - \mathbf{k}') \delta_{qq'},$$

$$\sum_q \int dS_k\, v_{kq}(\mathbf{y}) v^*_{kq}(\mathbf{y}') = y \delta(\mathbf{y} - \mathbf{y}').$$

This theorem is proved by techniques similar to those used to establish (82). Two examples of sets of type (B6) are furnished by (B7) and (B8) below:

$$v(\sigma) = (2\pi)^{-1/2} \sigma \Gamma(\tfrac{1}{2} - i\sigma) e^{1/2\pi(\sigma + (1/2)i)},$$
$$v_{k1}(\mathbf{y}) = (2\pi)^{-3/2} \frac{d}{d\zeta} (\zeta^{1/2} e^{i\zeta}); \quad \text{(B7)}$$

$$v(\sigma) = (2\pi)^{-1/2} \sigma \frac{\Gamma(-i\sigma)}{\Gamma(i\sigma)} \Gamma(\tfrac{1}{2} + i\sigma) e^{1/2\pi(\sigma - i)},$$
$$v_{k1}(\mathbf{y}) = 2^{-5/2} \pi^{-1} M(\tfrac{3}{2}, 1, i\zeta), \quad \text{(B8)}$$
$$= 2^{-5/2} \pi^{-1} [(1 + i\zeta) J_0(\tfrac{1}{2}\zeta) - \zeta J_1(\tfrac{1}{2}\zeta)] e^{(1/2)i\zeta}.$$

In (B8) M denotes Kummer's confluent hypergeometric function. In neither of the above examples is the function $v_{k2}(\mathbf{y})$ expressible in elementary form.

APPENDIX C: PROOF OF (94)

Because (94) has manifest Lorentz invariance, it is sufficient to prove it in the inertial frame in which $k^\lambda = (\kappa,0,0,0)$. In this frame

$$W\psi_k(\mathbf{y}) = \int Y\, dY\, d^2\hat{\mathbf{Y}}\, w(y_\lambda Y^\lambda)(2\pi)^{-2} i \times \int_0^\infty d\sigma\, \sigma e^{ig(\sigma)} (\kappa Y)^{i\sigma - 1}. \quad \text{(C1)}$$

The function $w(y_\lambda Y^\lambda)$ may be expressed as the integral transform

$$w(y_\lambda Y^\lambda) = \frac{i\kappa^2}{8\pi^2} \int_{-\infty}^{\infty} d\sigma' \times \left(\frac{1}{2} \kappa^2 y_\lambda Y^\lambda \right)^{-i\sigma' - 1} \sigma' \frac{\Gamma(i\sigma')}{\Gamma(-i\sigma')}. \quad \text{(C2)}$$

Upon substituting (C2) into (C1) one obtains the product of the two integrals

$$\int dY\, Y^{i(\sigma - \sigma') - 1} = 2\pi \delta(\sigma - \sigma'),$$

$$\int d^2\hat{\mathbf{Y}} \left[\frac{1}{2} (1 - \hat{\mathbf{y}} \cdot \hat{\mathbf{Y}}) \right]^{-i\sigma - 1} = 4\pi i \sigma^{-1}.$$

This leads to

$$W\psi_k(\mathbf{y}) = -(2\pi)^{-2} i \int_0^\infty d\sigma\, \sigma e^{ig(\sigma)} \times \frac{\Gamma(i\sigma)}{\Gamma(-i\sigma)} (\kappa y)^{-i\sigma - 1},$$

which is just (94) in this special inertial frame. Invoking Lorentz invariance, (94) holds generally.

[1] G. H. Derrick, J. Math. Phys. **28**, 64 (1987).
[2] P. A. M. Dirac, Rev. Mod. Phys. **21**, 392 (1949).
[3] *Alphabet conventions:* Greek lowercase letters $\iota,\kappa,\lambda,\ldots = 0,1,2,3$, with summation over repeated indices. *Metric tensor:* $\eta_{\lambda\mu} = \mathrm{diag}(1,-1,-1,-1)$. *Conjugation operations:* A superscript $*,T,\dagger$ applied to a quantity denotes, respectively, the complex conjugate, transpose, Hermitian conjugate. *Step function:* $\Theta(x) = 1$ if $x > 0$ and 0 if $x < 0$. *Sign function:* $\mathrm{sgn}(x) = \Theta(x) - \Theta(-x)$.
[4] C. Itzykson and J-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980), p. 114.
[5] L. Cohen, J. Math. Phys. **11**, 3296 (1970); E. Kerner and W. Sutcliffe, J. Math. Phys. **11**, 391 (1970); F. J. Testa, J. Math. Phys. **12**, 1471 (1971); I. W. Mayes and J. S. Dowker, J. Math. Phys. **14**, 434 (1973); M. M. Mizrahi, J. Math. Phys. **16**, 2201 (1975); J. S. Dowker, J. Math. Phys. **17**, 1873 (1976); A. C. Hirschfield, Phys. Lett. A **67**, 5 (1978).
[6] E. P. Wigner, *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra* (Academic, New York, 1959), Chap. 11.
[7] Note that if the mass m is zero, Lemma 1 and the derived theorem are false, because in this case there are an infinite number of unsymmetric SO(1,3) scalars derived from k^λ and $k^{\lambda'}$ of the form $(kk')^{-1}\beta(k/k')\delta(\hat{\mathbf{k}},\hat{\mathbf{k}}')$, with β an arbitrary function, cf. (A6).
[8] E. P. Wigner, Ann. Math. **40**, 149 (1939); V. Bargmann and E. P. Wigner, Proc. Natl. Acad. Sci. USA **34**, 211 (1948); L. L. Foldy, Phys. Rev. **102**, 568 (1956).
[9] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics* (Pergamon, Oxford, 1965), p. 20.
[10] R. M. F. Houtappel, H. Van Dam, and E. P. Wigner, Rev. Mod. Phys. **37**, 595 (1965).
[11] S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1961), p. 177.

Sufficient conditions for zero not to be an eigenvalue of the Schrödinger operator

A. G. Ramm
Mathematics Department, Kansas State University, Manhattan, Kansas 66506

(Received 9 September 1986; accepted for publication 4 February 1987)

It is proved that if $H = -\nabla^2 + q(x) \geqslant 0$, $\operatorname{Im} q = 0$, $|q(x)| \leqslant c(1+|x|)^{-a}$, $c = \text{const} > 0$, $a > 2$, then zero is not an eigenvalue of H. An example is given of $H \geqslant 0$, with zero a resonance (half-bound state) and $q = q(|x|)$ compactly supported and integrable. An example of a potential $q = O(r^{-2})$ is known, for which $H \geqslant 0$ and zero is an eigenvalue. This shows that $a > 2$ is the optimal condition for zero not to be an eigenvalue of $H \geqslant 0$. If the condition $H \geqslant 0$ does not hold and H is an operator in $L^2(\mathbb{R}^3)$, then zero can be an eigenvalue even if $q \in C_0^\infty$. If H is an operator in $L^2(\mathbb{R}^1)$ or in $L^2(\mathbb{R}^1_+)$, $\mathbb{R}^1_+ = [0,\infty)$, then zero cannot be an eigenvalue of H provided that $a > 2$; here conditions $H \geqslant 0$ and $\operatorname{Im} q = 0$ can be dropped. Global estimates of the Green's function of H from below and above are given.

I. INTRODUCTION

In Ref. 1 Newton asked the following question. Let $H = -\nabla^2 + q(x)$, $\operatorname{Im} q(x) = 0$, $x \in \mathbb{R}^3$, $\int (1+|x|) \times |q(x)|\,dx < \infty$, $\int = \int_{R_3}$. Assume that

$$H \geqslant 0. \tag{1}$$

Can 0 be an eigenvalue of H?

The assumption about the decay of $q(x)$ can be relaxed: It is sufficient to assume that for all sufficiently large x the estimate

$$|q(x)| \leqslant c(1+|x|)^{-a}, \quad a > 2, \quad c = \text{const} > 0 \tag{2}$$

holds and that $q \in L^2_{\text{loc}}$. By c we denote below various constants. The assumption $(*)$ $\int (1+|x|)|q(x)|\,dx < \infty$ means, roughly speaking, that $a > 4$ in (2). On the other hand, $(*)$ allows local singularities in a neighborhood of infinity, which are excluded by (2). Since assumption (2) covers most, if not all, of the potentials of interest which decay at infinity faster than $|x|^{-2}$, we will use this assumption.

The question raised by Newton was discussed in Ref. 2, where it was proved that if $a > 3$ then the answer is no. The assumptions on q in Ref. 2 were given in terms of weighted L^p spaces. Our argument is different from the one in Ref. 2.

The purpose of this paper is to give the exact value of a for which the answer is no. We prove that if $a > 2$, then the answer is no and if $a \leqslant 2$, then zero can be an eigenvalue of $H \geqslant 0$.

The second part of this statement is known: It is shown in Ref. 3 (p. 375) that for central potentials $q = q(r)$, $r = |x|$, if $H\psi = 0$, $\psi = r^{-1}u(r)Y_l(x^0)$, where Y_l is the spherical harmonic $x^0 = xr^{-1}$, then $u = O(r^{-l})$ as $r \to \infty$. Therefore $\psi \in L^2(\mathbb{R}^3)$ provided that $l > \frac{1}{2}$. This conclusion does not use assumption (1). The potential $q(r)$ can be chosen in C_0^∞.

If $l = 0$, $q = q(r)$, $\operatorname{Im} q = 0$, is compactly supported and integrable in a neighborhood of the origin and (1) holds, then zero can be a half-bound state (a resonance) although it cannot be a bound state.

In the one-dimensional case, if condition (2) holds and

$$u'' - q(r)u = 0, \quad r \geqslant 0, \quad u \in L^2[0,\infty), \tag{3}$$

then $u = 0$. This conclusion holds without assumption (1) and without assumption $\operatorname{Im} q = 0$.

We prove that the Green's function $G(x,y)$ of H under the assumptions $q(x) \geqslant 0$ and (2) satisfies the global estimates

$$c|x-y|^{-1} \leqslant G(x,y) \leqslant (4\pi|x-y|)^{-1}, \quad c = \text{const} > 0. \tag{4}$$

The low energy scattering has been studied in Refs. 4 and 5.

II. THE RESULTS

Theorem 1: If $H \geqslant 0$ and $a > 2$, then zero is not an eigenvalue of H.

Proof: Let $H = H_0 + q$, $H_0 = -\nabla^2$, $H\psi = 0$, $\psi \in L^2(\mathbb{R}^3)$. Then $\psi > 0$, being the ground state of H. Let us first assume that $q(x) = -p(x)$ and $p(x) \geqslant 0$. If $q(x) = q_+(x) - q_-(x)$, where $q_+ = \max(0, q(x))$, $q_-(x) = \max(0, -q(x))$, then we can use a similar argument taking $p(x) = q_-(x)$ and $H_0 = -\nabla^2 + q_+(x)$. The key estimate for the Green's function of the operator $-\nabla^2 + q_+$ that is needed for the proof is the estimate (4). This estimate is proved in Lemma 2.

Step 1: If $q(x) = -p(x)$, $p(x) \geqslant 0$, then $H\psi = 0$ can be written as

$$\psi = H_0^{-1} p\psi = \int g p \psi\, dy, \quad g = (4\pi|x-y|)^{-1}. \tag{5}$$

Since $\psi > 0$ and $p \geqslant 0$, we have $p\psi \geqslant 0$ and

$$\psi = (4\pi|x|)^{-1} \int p\psi\, dy + o(|x|^{-1}), \quad |x| \to \infty. \tag{6}$$

Estimate (6) is proved in Lemma 1.

If $\psi \in L^2$ then (6) implies that $\int p\psi\, dy = 0$. Since $p\psi \geqslant 0$, this means that $p\psi = 0$. Therefore $\nabla^2\psi = 0$ and $\psi \in L^2(\mathbb{R}^3)$. Thus $\int |\nabla\psi|^2\, dx = 0$ and $\psi = 0$. The same conclusion follows from the equation $p\psi = 0$ if $p \neq 0$ on an open set. On this set $\psi = 0$ and, by the unique continuation property for elliptic equations, $\psi \equiv 0$.

Step 2: If $q = q_+ - p(x)$, $H_0 = -\nabla^2 + q_+$, and $H_0 G$

 0022-2488/87/061341-03$02.50

$=\delta(x-y)$, then G satisfies estimate (4) and (5) holds with G in place of g. Therefore

$$c|x|^{-1}\int p\psi\,dy+o(|x|^{-1})$$
$$\leqslant\psi\leqslant(4\pi|x|)^{-1}\int p\psi\,dy+o(|x|^{-1}),\quad |x|\to\infty. \tag{7}$$

If $\psi\in L^2$, then $\int p\psi\,dy=0$, and the rest of the argument is the same. Theorem 1 is proved.

Lemma 1: If $a>2$ and $\psi\in L^2(\mathbb{R}^3)$, then (5) implies (6).

Proof: One has

$$|\psi|\leqslant 4\pi\left(\int|x-y|^{-2}|p|^2\,dy\right)^{1/2}\left(\int|\psi|^2\,dy\right)^{1/2}$$
$$\leqslant c(1+|x|)^{-1}. \tag{8}$$

Here and below c denotes various positive constants, the first integral in (8) was estimated with the help of inequality (2), and the inequality $p>0$ was not used. From (5) and (8) one obtains

$$\psi=(4\pi|x|)^{-1}\int_{|y|<\epsilon|x|}dy\,p\psi(1+O(\epsilon))$$
$$+\int_{|y|>\epsilon|x|}(4\pi|x-y|)^{-1}p\psi\,dy:=J_1+J_2. \tag{9}$$

Here $0<\epsilon=\epsilon(r)\to 0$, $\epsilon(r)r\to\infty$ as $r=|x|\to\infty$. Since $a>2$ and (8) holds we have $\int p\psi\,dy<\infty$. Therefore $J_1=(4\pi|x|)^{-1}\int dy\,p\psi+o(|x|)^{-1}$. Let us show that $J_2=o(|x|^{-1})$ as $|x|\to\infty$. One has

$$J_2\leqslant c\int_{\epsilon r}^{\infty}dt\,t^2(1+t)^{-(a+1)}\int_{-1}^{1}(r^2+t^2-2rts)^{-1/2}\,ds$$
$$\leqslant c\int_{\epsilon r}^{\infty}dt(1+t)^{-a+1}(2rt)^{-1}[r+t-|r-t|]$$
$$\leqslant cr^{-1}\left[\int_{\epsilon r}^{r}dt(1+t)^{-a}2t+\int_{r}^{\infty}dt(1+t)^{-a}2r\right]$$
$$\leqslant cr^{-a+1}+o(r^{-1}). \tag{10}$$

Here $s=\cos\theta$ and we used the spherical coordinates with the y_3 axis along the vector x, $|y|+t$. Lemma 1 is proved.

Lemma 2: If $q\geqslant 0$ satisfies estimates (2), with $a>2$ for $|x|\geqslant R$, where $R>0$ is an arbitrary large number, and $q\in L^2_{\mathrm{loc}}$, then the Green's function of $H=-\nabla^2+q(x)$ satisfies estimate (4).

Proof: The right inequality in (4) follows from the maximum principle. Indeed, first note that $G>0$: If $G(x_0,y_0)\leqslant 0$, then, since $G(x,y_0)\to+\infty$ as $x\to y_0$ and $G(x,y_0)\to 0$ as $|x|\to\infty$, the function $G(x,y_0)$ attains a nonpositive minimum at a certain point $\xi\neq y_0$. Since $G\not\equiv$const, $q\geqslant 0$, and $\nabla^2 G=qG$ in a neighborhood of ξ, we have a contradiction which proves that $G(x,y)>0$. Since $G=g-\int gqG\,dz\leqslant g$, we conclude that $g\geqslant G$.

Let us prove the left inequality (4).

We have

$$G(x,y)=g(x,y)-\int g(x,z)q(z)G(z,y)dz. \tag{11}$$

Suppose that the left inequality (4) does not hold. Then there exist sequences x_n and y_n such that

$$|x_n-y_n|G(x_n,y_n)\leqslant n^{-1},\quad n\to\infty. \tag{12}$$

We will show this is impossible, so that (4) holds with some $c>0$. There are three cases to consider.

Case 1: There exist numbers m and δ which do not depend on n, such that $|x_n|\leqslant m$, $|y_n|\leqslant m$, $|x_n-y_n|\geqslant\delta>0$. In this case choose $x_n\to x_0$, $y_n\to y_0$ and pass to the limit in (12) to get $G(x_0,y_0)=0$, which is a contradiction since $G(x_0,y_0)>0$.

Case 2: There exists a number m such that $|x_n|\leqslant m$, $|y_n|\leqslant m$, and $|x_n-y_n|\to 0$. Then choose $x_n\to x_0, y_n\to y_0$, multiply (11) by $|x-y|$, set $x=x_n, y=y_n$, and pass to the limit $n\to\infty$ to get $0=(4\pi)^{-1}$, which is a contradiciton. Here we used $\lim|x_n-y_n|\int g(x_n,z)qG(z,y_n)dz=0$, which holds since the integral is bounded.

Case 3: Either $|x_n|\to\infty$ or $|y_n|\to\infty$, or $|x_n|\to\infty$ and $|y_n|\to\infty$. From (11) and (12) one obtains

$$0=(4\pi)^{-1}-\lim_{n\to\infty}|x_n-y_n|\int g(x_n z)q(z)G(z,y_n)dz. \tag{13}$$

If $|x_n-y_n|\leqslant m$, where m does not depend on n, then the limit in (13) is zero since $g(x_n,z)=(4\pi|x_n-z|)^{-1}\to 0$ as $|x_n|\to\infty$, and $G(z,y_n)\leqslant(4\pi|z-y_n|)^{-1}\to 0$ as $|y_n|\to\infty$. In this case Eq. (13) becomes $0=(4\pi)^{-1}$, which is a contradiction.

If $|x_n-y_n|\to\infty$, then

$$\lim|x_n-y_n|\int g(x_n,z)q(z)G(z,y_n)dz$$
$$\leqslant\lim c|x_n-y_n|\left[\int_{|z|<R}\frac{dz}{|x_n-z||z-y_n|(1+|z|)^a}\right.$$
$$\left.+\int_{|z|>R}\frac{dz}{|x_n-z||y_n-z|(1+|z|)^a}\right]$$
$$\leqslant c\lim|x_n-y_n|\cdot O(|x_n-y_n|^{1-a})=0.$$

So again Eq. (13) leads to a contradiction. Therefore (12) cannot hold and the left inequality (4) is proved (cf. Ref. 6, p. 314).

Remark: In the one-dimensional case if $H\psi=-\psi''+q(r)\psi=0$, $r\geqslant 0$, $a>2$, and $\psi\in L^2[R,\infty)$, then $\psi\equiv 0$. Here $R>0$ is an arbitrary (large) fixed number. This conclusion holds without assumption $H\geqslant 0$ and without assumption $\operatorname{Im} q=0$. Indeed, the differential equation implies

$$\psi=A+Br+\int_r^{\infty}(t-r)q\psi\,dt, \tag{14}$$

where A and B are constants. If $\psi\in L^2[R,\infty)$ and $a>2$, then

$$\left|\int_r^{\infty}(t-r)q\psi\,dt\right|$$
$$\leqslant\left(\int_r^{\infty}t^2(1+t)^{-2a}\,dt\right)^{1/2}\cdot\left(\int_r^{\infty}|\psi|^2\,dt\right)^{1/2}$$
$$=o(r^{-a+1.5}),\quad \text{as } r\to+\infty.$$

Therefore (14) and $\psi\in L^2[R,\infty)$ imply that $A=B=0$. If $A=B=0$, then (14) becomes a homogeneous Volterra equation for ψ and therefore $\psi=0$.

Remark 2: (a) Although, as we stated in Introduction, it is shown in Ref. 3 that there are central potentials such that

$\nabla^2\psi - q\psi = 0$, $\psi \in L^2(\mathbb{R}^3)$, no specific examples are given in Ref. 3. We give such examples and show that $q(r)$ can be chosen in C_0^∞. The corresponding $\psi = h(r)Y_1(x^0) \in L^2(R^3)$, $x^0 = xr^{-1}$. The construction is simple. Take $h(r) = r^{-2}$ for $r \geqslant 1$, $h(r) > 0$ for $0 \leqslant r \leqslant 1$, $h \in C^\infty$. Define $q = \psi^{-1}\nabla^2\psi$. Then $q \in C^\infty$ and $q = 0$ for $r \geqslant 1$ since $\nabla^2 r^{-2}Y_1 = 0$ for $r \geqslant 1$. Clearly $\psi \in L^2(\mathbb{R}^3)$. A similar example is in Ref. 7.

(b) An explicit example of an integrable near $r = 0$ compactly supported $q(r)$, such that $H \geqslant 0$ and zero is a half-bound state can also be constructed. One defines $\psi = r^{-1}$ for $r \geqslant 1$, $\psi = r^{-1}u(r)$ for $r \leqslant 1$, and chooses $u(r)$ so that $q(r)$: $= \psi^{-1}\nabla^2\psi$ is integrable and $H \geqslant 0$. Note that $q(r) = 0$ for $r \geqslant 1$. The desired u one can choose, for example, in the form $u = r^\gamma(1 + \gamma - \gamma r)$, with any $0 < \gamma < (\sqrt{2} - 1)/2$.

III. CONCLUSIONS

In this paper we prove that $H \geqslant 0$ does not have zero eigenvalue if (2) with $a > 2$ holds and may have zero if q falls off as $O(r^{-2})$, $r \to \infty$. An example of compactly supported integrable potential is given for which $H \geqslant 0$ and zero is a resonance.

ACKNOWLEDGMENT

The author thanks Professor R. G. Newton and Professor O. L. Weaver for discussions.

This work was supported by ONR. It was written while the author was a research professor at the Mathematical Institute of Academia Sinica. The author is grateful to the Academia for hospitality.

[1] R. C. Newton, J. Math. Phys. **18**, 1353 (1977).
[2] M. Klaus and B. Simon, Ann. Phys. (NY) **130**, 251 (1980).
[3] R. C. Newton, *Scattering Theory of Waves and Particles* (Springer, New York, 1982).
[4] R. C. Newton, J. Math. Phys. **27**, 2720 (1986).
[5] A. G. Ramm, J. Math. Phys. **21**, 308 (1980); *Theory and Applications of Some New Classes of Integral Equations* (Springer, New York, 1980), Appendix 4.
[6] A. G. Ramm, *Scattering by Obstacles* (Reidel, Dordrecht, 1986).
[7] C. L. Dolph, B. McLeod, and D. Thoe, J. Math. Anal. Appl. **16**, 311 (1966).

The Korteweg–de Vries hierarchy of isospectral transformations: Towards a general explicit expression

Avia Rosenhouse[a] and Jacob Katriel
Department of Chemistry, Technion—Israel Institute of Technology, Haifa 32000, Israel

(Received 15 October 1985; accepted for publication 4 February 1987)

The structure of the Korteweg–de Vries hierarchy of evolution equations, generating isospectral transformations, is elucidated by means of a study of its recurrence relations. For the mth member of the KdV hierarchy, which can be written in the form $V_t = -2A_{m+1,x}$, where the A_i satisfy the recurrence relation $A_{m+1,x} = VA_{m,x} + \frac{1}{2}A_m V_x - \frac{1}{4}A_{m,xxx}$, it is shown that A_m is a homogeneous polynomial in $\partial^i V/\partial x^i$. A general combinatorial formula for the coefficients of all the monomials entering A_m, up to a set of constants determined by means of a recurrence relation, is derived.

I. INTRODUCTION

The Korteweg–de Vries (KdV) hierarchy of isospectral transformations was introduced by Lax[1] and by Gardner *et al.*[2] It is very intimately related to the question of uniqueness of spectral inversion. The connection with the inverse scattering method has been particularly clearly studied with respect to the original KdV equation.[3]

The classical limit of the KdV hierarchy was recently discussed,[4] and it was shown that in this limit the hierarchy reduces to the first-order equation

$$V_t = f(V)\cdot V_x \tag{1}$$

in which $f(V)$ is an arbitrary function of V. This equation generates an isoperiodic transformation of $V_0(x) = V(x,0)$ into $V(x,t)$. Thus, to each $f(V)$ corresponds some isoperiodic transformation. It was also shown in Ref. 4 that if the two isoperiodic potentials $V(x,0)$ and $V(x,1)$ are given, the form of $f(V)$ generating the transformation between them via Eq. (1) can easily be written down.

The complete KdV hierarchy has so far not been studied extensively at all. In particular, the general form of an arbitrary isospectral transformation has not been explicitly derived.

In the present article we present an attempt to derive the explicit form of the higher-order members of the hierarchy. General explicit forms are obtained, up to some numerical coefficients for which a set of recurrence relations is derived. The form of the results enables the association of an isospectral transformation with any $F(V)$, although the inversion problem, concerning the determination of the form of $F(V)$ generating a particular isospectral transformation, has not been solved.

A related problem, which has been worked out in considerable detail, concerns the determination of the infinite sequence of polynomial conservation laws of the KdV equation.[5,6] In view of the relation between these conservation laws and the members of the KdV hierarchy, one could consider the results, in particular in Ref. 6, as almost providing the explicit form of the KdV hierarchy. However, the elucidation of the structure of the KdV hierarchy achieved in the present article is a prerequisite for the consideration of the general isospectral transformation as presented in the concluding section, i.e., in terms of an arbitrary $F(V)$.

[a] Based on a part of a thesis to be submitted by AR to the senate of the Technion—Israel Institute of Technology, in partial fulfillment of the requirements for the M.Sc. degree.

II. PRELIMINARY CONSIDERATIONS

The evolution equation

$$V_t = -2A_{m+1,x}\,, \tag{2}$$

where

$$A_{m+1,x} = V\cdot A_{m,x} + \tfrac{1}{2}A_m\cdot V_x - (\hbar^2/4)A_{m,xxx} \tag{3}$$

specifies an isospectral transformation of the one-dimensional Schrödinger equation $H\psi = E\psi$, where $H = -\hbar^2(d^2/dx^2) + V$. Here x is the dynamical coordinate, t is a parameter such that $V = V(x,t)$, $V_t = \partial V/\partial t$, $V_x = \partial V/\partial x$, and $A_{m,x} = \partial A_m/\partial x$. Here m is a running index, $m = 0,1,2,\ldots$.

The fact that the transformation is isospectral means that for all eigenvalues we have $\partial E/\partial t = 0$, or equivalently $\int_{-\infty}^{\infty} \psi^* V_t \psi\, dx = 0$, where ψ is any one of the eigenfunctions. That $\int_{-\infty}^{\infty} \psi^* A_{m+1,x}\psi\, dx = 0$ was shown in Refs. 1, 2, and 4. It follows immediately that an arbitrary linear combination $\Sigma_m \alpha_m A_{m+1,x}$, where $\{\alpha_m\}$ is a set of constant coefficients, will also specify an isospectral transformation.

A_m was implicitly presented in Refs. 1 and 2 for $m = 0,1,2,3$, i.e.,

$$A_0 = -1, \quad A_1 = -V/2\,,$$

$$A_2 = -\tfrac{3}{8}V^2 + (\hbar^2/8)\,V_{xx}\,,$$

$$A_3 = -\tfrac{5}{16}V^3 + \tfrac{5}{16}\hbar^2\left[V\cdot V_{xx} + \tfrac{1}{2}V_x^2\right] - (\hbar^4/32)V_{(4)}\,,$$

resulting in the evolution equations

$$V_t = V_x, \quad V_t = \tfrac{3}{2}VV_x - (\hbar^2/4)V_{(3)}\,,$$

$$V_t = \tfrac{15}{8}V^2V_x - \tfrac{5}{8}\hbar^2\left[V\cdot V_{(3)} + 2V_xV_{xx}\right] - (\hbar^4/32)V_{(5)}\,,$$

where $V_{(i)} = \partial^i V/\partial x^i$.

To elucidate the explicit form of the general isospectral transformation we first consider the successive terms that can be generated using Eq. (3).

Writing

 0022-2488/87/061344-07$02.50

$$A_m = \sum_{i=0}^{m-1} (-1)^{i+1} \hbar^{2i} A_{i,m} , \tag{4}$$

we obtain

$$A_{i+1,m+1,x} = V \cdot A_{i+1,m,x} + \tfrac{1}{2} A_{i+1,m} V_x + \tfrac{1}{4} A_{i,m,xxx} . \tag{5}$$

The nature of this two-dimensional recurrence relation is illustrated in Fig. 1. It determines $A_{i+1,m+1,x}$, $0 \leqslant i < m-1$, in terms of the two elements $A_{i+1,m}$ and $A_{i,m}$ or, if $i = m-1$, in terms of $A_{i,m}$ only. It is therefore necessary to specify both $A_{0,0}$ and the constants of integration entering upon evaluation of $A_{i+1,m+1}$ from $A_{i+1,m+1,x}$. We shall set $A_{0,0} = -1$, and all the integration constants will be chosen to be zero. Although some further flexibility in the final expression for the general isospectral transformation could be incorporated by allowing the constants of integration to be arbitrary functions of t, this further flexibility, whose classical analog was discussed in Ref. 4, will not be explicitly retained.

The recurrence relation becomes particularly simple for $A_{0,m}$, obtaining the form

$$A_{0,m+1,x} = V \cdot A_{0,m,x} + \tfrac{1}{2} A_{0,m} \cdot V_x ,$$

which is just the classical limit of the original recurrence relation, Eq. (3). This recurrence relation, which involves stepping along the horizontal sequence $i = 0$ in Fig. 1, was solved in Ref. 4, where it was shown that

$$A_{0,m} = [(2m-1)!!/2^m \cdot m!] V^m . \tag{6}$$

Having obtained all the terms corresponding to $i = 0$, it is now a simple matter to proceed along the horizontal line corresponding to $i = 1$ in Fig. 1. The terms along this line constitute the lowest-order quantal terms. Thus substituting Eq. (6) in Eq. (5) we obtain

$$\begin{aligned} A_{1,m+1,x} = {} & V \cdot A_{1,m,x} + \tfrac{1}{2} A_{1,m} \cdot V_x \\ & + [(2m-1)!!/2^m \cdot (m-1)!] \\ & \times [(m-1)(m-2) V^{m-3} V_{(1)}^3 \\ & + 3(m-1) V^{m-2} V_{(2)} V_{(1)} + V^{m-1} \cdot V_{(3)}] . \end{aligned}$$

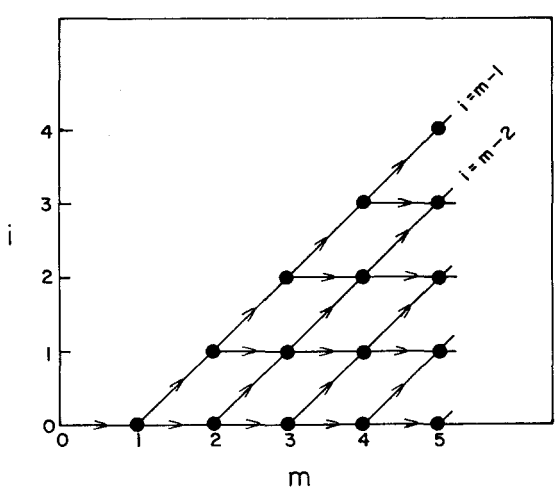

FIG. 1. The two-dimensional recurrence relation for $A_{m,i}$.

It follows straightforwardly that

$$A_{1,2} = \tfrac{1}{8} V_{(2)} ,$$

$$A_{1,3} = \tfrac{5}{16} (V \cdot V_{(2)} + \tfrac{1}{2} V_{(1)}^2) ,$$

$$A_{1,4} = \tfrac{35}{64} (V^2 \cdot V_{(2)} + V \cdot V_{(1)}^2) ,$$

$$\vdots$$

and it can be shown by induction that

$$A_{1,m} = \frac{(2m-1)!!}{2^m \cdot 6} \left[\frac{V^{m-2}}{(m-2)!} \cdot V_{(2)} + \frac{V^{m-3} \cdot V_{(1)}^2}{(m-3)! \cdot 2!} \right] . \tag{7}$$

This result can be substituted in Eq. (5) to obtain

$$A_{2,3} = \tfrac{1}{32} V_{(4)} ,$$

$$A_{2,4} = \tfrac{7}{64} V \cdot V_{(4)} + \tfrac{7}{32} V_{(3)} V_{(1)} + \tfrac{21}{128} V_{(2)}^2 ,$$

$$A_{2,5} = \tfrac{63}{256} V^2 V_{(4)} + \tfrac{63}{64} V V_{(3)} V_{(1)} + \tfrac{189}{256} V \cdot V_{(2)}^2 + \tfrac{231}{256} V_{(2)} V_{(1)}^2 ,$$

$$A_{2,6} = \tfrac{231}{512} V^3 V_{(4)} + \tfrac{693}{256} V^2 V_{(3)} V_{(1)} + \tfrac{2079}{1024} V^2 \cdot V_{(2)}^2 + \tfrac{2541}{512} V V_{(2)} V_{(1)}^2 + \tfrac{1155}{2048} V_{(1)}^4 ,$$

$$\vdots .$$

On the basis of these results one can already make certain observations concerning the form of $A_{i,m}$ in general. First, $A_{i,m}$ is homogeneous, of degree $m-i$, with respect to V (including its derivatives). Furthermore, $A_{i,m}$ is a sum of terms, each one of which contains a product of derivatives of V, the sum of whose orders is $2i$. This sum is called the derivative index in Ref. 6.

Thus the orders of the derivatives in each term of $A_{i,m}$ constitute a partition of $2i$. These two statements can be established by induction, using Eq. (5). The induction has to be carried out over the two indices i and m, which can be done by imbedding an induction over m within an induction over i, as follows: checking Eq. (6) to establish that the two statements are true for all $A_{0,m}$ (i.e., $i = 0$ and all m) we shall assume that they are true for i and all m and show that they are, consequently, true for $i+1$ and all m.

The last statement will be proved by induction over m, with fixed i: From Eq. (5) and the induction hypothesis it follows that the theorem holds for $A_{i+1,i+2}$. We assume that it holds for $A_{i+1,m}$ and show that it holds for $A_{i+1,m+1}$ by noting that each one of the three terms in the rhs of the recurrence relation, Eq. (5), is of order $2i+3$ and degree $m-i$ from which it follows that $A_{i+1,m+1}$ is of order $2(i+1)$ and degree $(m+1)-(i+1) = m-i$. This completes the proof.

Denoting by ξ the number of summands in a given partition of $2i$, we observe that the ratios between the coefficients

of terms corresponding to partitions of equal ξ are independent of m. Thus, for $i=2$, $\xi=2$ the possible partitions of $2i=4$ are $3+1$ and $2+2$. The ratios of the corresponding terms are 4/3 in $A_{2,4}$, $A_{2,5}$ as well as in $A_{2,6}$. The general validity of this observation will be established later on. It suggests that the coefficient of each term in $A_{i,m}$ can be written as a product of two factors, one of which depends on the partition but not on m, the other one depending on m, i, and ξ but not on the specific partition.

The observations just made suggest that the general form of $A_{2,m}$ is

$$A_{2,m}=\frac{(2m-1)!!}{2^m\cdot 180}\left\{3\cdot\frac{V^{m-3}\cdot V_{(4)}}{(m-3)!}+\frac{V^{m-4}}{(m-4)!}\left[6V_{(3)}V_{(1)}+9\cdot\frac{V_{(2)}^2}{2!}\right]+11\cdot\frac{V^{m-5}\cdot V_{(2)}V_{(1)}^2}{(m-5)!\cdot 2!}+15\cdot\frac{V^{m-6}\cdot V_{(1)}^4}{(m-6)!\cdot 4!}\right\}. \tag{8}$$

This relation can be established by induction. One can proceed and obtain

$$A_{3,m}=\frac{(2m-1)!!}{2^m\cdot 2520}\left\{3\cdot\frac{V^{m-4}}{(m-4)!}\cdot V_{(6)}+\frac{V^{m-5}}{(m-5)!}\left[9V_{(5)}V_{(1)}+19\cdot V_{(4)}V_{(2)}+\frac{23}{2!}V_{(3)}^2\right]+\frac{V^{m-6}}{(m-6)!}\left[\frac{25}{2!}V_{(4)}V_{(1)}^2+43V_{(3)}V_{(2)}V_{(1)}+\frac{61}{3!}V_{(2)}^3\right]+\frac{V^{m-7}}{(m-7)!}\left[\frac{60}{3!}V_{(3)}V_{(1)}^3+\frac{83}{2!\cdot 2!}V_{(2)}^2V_1^{(2)}\right]+\frac{V^{m-8}}{(m-8)!}\cdot\frac{119}{4!}V_{(2)}V_{(1)}^4+\frac{V^{m-9}}{(m-9)!}\cdot\frac{175}{6!}V_{(1)}^6\right\}. \tag{9}$$

Further progress along these lines becomes rather cumbersome. However, before we attempt to undertake some more general considerations, let us obtain one further simple special case, along the line characterized by $i=m-1$, in Fig. 1.

Along that line the recurrence relation obtains the form

$$A_{m,m+1,x}=\tfrac{1}{4}A_{m-1,m,xxx}$$

or

$$A_{m,m+1}=\tfrac{1}{4}A_{m-1,m,xx}\,.$$

It follows immediately that

$$A_{m-1,m}=(1/2^{2m-1})\cdot V_{(2m-2)}\,. \tag{10}$$

A comparison of this result with Eqs. (6)–(9) suggests that

$$A_{i,m}=\frac{(2m-1)!!}{3^i\cdot(2i+1)!!\cdot 2^{m+i}}\cdot\sum_{i_0,i_1,\cdots i_{2i}}C\left[1^{i_1}2^{i_2}\cdots(2i)^{i_{2i}}\right]\cdot\prod_{j=0}^{2i}\frac{V_{(j)}^{i_j}}{i_j!}\left(\sum_{j=0}^{2i}i_j=m-i,\quad\sum_{j=1}^{2i}j\cdot i_j=2i\right), \tag{11}$$

where $C[1^{i_1}2^{i_2}\cdots(2i)^{i_{2i}}]$ are numerical coefficients which depend on the partition of $2i$ but will be shown to be independent of m. These coefficients have already been determined, via Eqs. (6)–(9), for $i=0,1,2,3$. They are presented, along with further coefficients, in Table I.

For a given i and a large enough m, the number of terms in $A_{i,m}$ is equal to the number of partitions of $2i$ into sums of positive integers. Since i_j is the number of times that j appears in the partition, it follows that $\xi=\Sigma_{j=1}^{2i}i_j$ is the number of summands in the partition considered. The maximum number of summands, obtained by writing $2i=1+1+\cdots+1$ (i.e., $i_1=2i$; $i_j=0, j>1$) is $\xi_m=2i$. The power of V in a term corresponding to a partition into ξ summands is

$$i_0=m-i-\sum_{j=1}^{2i}i_j=m-i-\xi\,.$$

Therefore, the minimum value of m for which all possible partitions of $2i$ appear in $A_{i,m}$ is $m=i+\xi_m=3i$.

For $i=0$ the minimal value of m is zero. However, for $i>0$ the minimal value of ξ is 1, corresponding to $i_j=0$, $i\leqslant j<2i$; $i_{2i}=1$. Therefore, the minimal value of m, corresponding to $i_0=0$, is $i+1$. For $i+1\leqslant m<3i$ only partitions into at most $m-i$ summands are allowed.

III. THE GENERAL RECURRENCE RELATION FOR THE COEFFICIENTS

Having written $A_{i,m}$ in Eq. (11), we shall now derive general recurrence relations for the coefficients $C[1^{i_1}2^{i_2}\cdots(2i)^{i_{2i}}]$ which will enable us to show that these coefficients are indeed independent of m. The recurrence relations for the coefficients are obtained by substituting Eq. (11) in Eq. (5) and comparing coefficients. A typical term appearing in Eq. (5) after substitution of (11) is

$$g=\prod_{j=0}^{n}V_{(j)}^{i_j}\,,$$

where

$$\sum_{j=0}^{n}i_j=m-1,\quad\sum_{j=1}^{n}j\cdot i_j=2i+3,\quad n\leqslant 2i+3\,.$$

To obtain the coefficient of g in Eq. (5) we note that it appears in the following ways.

A. From $A_{i+1,m+1,x}$

Here g can only arise upon differentiation of terms of the form $h=gV_{(k)}/V_{(k+1)}$ which, being of degree $m-i$ and order $2(i+1)$, are contained in $A_{i+1,m+1}$. Such a term is only present if g contains $V_{(k+1)}$ (i.e., $i_{k+1}>0$). Upon differentiation it will generate many other terms in addition to g, but all these other terms are not relevant. The coefficient of the corresponding contribution to g will be

$$\alpha_{i+1,m+1}(h)\cdot(i_k+1)\{k+1\}\,.$$

TABLE I. The coefficients for $i \leqslant 5$.

$C[0]$	= 1	$C[8]$	= 81	$C[10]$	= 243	$C[2^2 3^2]$	= 2183 517/5
		$C[1,7]$	= 324	$C[1,9]$	= 1 215	$C[1^4,6]$	= 101 331
$C[2]$	= 3	$C[2,6]$	= 891	$C[2,8]$	= 4 131	$C[1^3,2,5]$	= 1223 262/5
$C[1^2]$	= 3	$C[3,5]$	= 1 539	$C[3,7]$	= 9 234	$C[1^3,3,4]$	= 370 332
		$C[4^2]$	= 1 863	$C[4,6]$	= 14 823	$C[1^2,2^2,4]$	= 2539 593/5
$C[4]$	= 9	$C[1^2,6]$	= 1 215	$C[5^2]$	= 17 253	$C[1^2,2,3^2]$	= 3192 777/5
$C[1,3]$	= 18	$C[1,2,5]$	= 2 916	$C[1^2,8]$	= 5 751	$C[1,2^3,3]$	= 4386 879/5
$C[2^2]$	= 27	$C[1,3,4]$	= 4 374	$C[1,2,7]$	= 17 658	$C[2^5]$	= 1209 411
$C[1^2,2]$	= 33	$C[2^2,4]$	= 6 075	$C[1,3,6]$	= 176 904/5	$C[1^5,5]$	= 363 285
$C[1^4]$	= 45	$C[2,3^2]$	= 7 533	$C[1,4,5]$	= 248 427/5	$C[1^4,2,4]$	= 754 515
		$C[1^3,5]$	= 4 131	$C[2^2,6]$	= 243 081/5	$C[1^4,3^2]$	= 949 887
$C[6]$	= 27	$C[1^2,2,4]$	= 42 687/5	$C[2,3,5]$	= 85 293	$C[1^3,2^2,3]$	= 6518 718/5
$C[1,5]$	= 81	$C[1^2,3^2]$	= 53 379/5	$C[2,4^2]$	= 514 593/5	$C[1^2,2^4]$	= 8960 139/5
$C[2,4]$	= 171	$C[1,2^2,3]$	= 73 548/5	$C[3^2,4]$	= 643 059/5	$C[1^6,4]$	= 1136 025
$C[3^2]$	= 207	$C[2^4]$	= 102 141/5	$C[1^3,7]$	= 25 272	$C[1^5,2,3]$	= 1964 655
$C[1^2,4]$	= 225	$C[1^4,4]$	= 12 393	$C[1^2,2,6]$	= 346 923/5	$C[1^4,2^3]$	= 13495 977/5
$C[1,2,3]$	= 387	$C[1^3,2,3]$	= 106 677/5	$C[1^2,3,5]$	= 610 983/5	$C[1^7,3]$	= 2993 760
$C[2^3]$	= 549	$C[1^2,2^3]$	= 146 853/5	$C[1^2,4^2]$	= 147 177	$C[1^6,2^2]$	= 4113 747
$C[1^3,3]$	= 540	$C[1^5,3]$	= 31 590	$C[1,2^2,5]$	= 837 621/5	$C[1^8,2]$	= 6330 555
$C[1^2,2^2]$	= 747	$C[1^4,2^2]$	= 216 999	$C[1,2,3,4]$	= 1265 949/5	$C[1^{10}]$	= 9823 275
$C[1^4,2]$	= 1071	$C[1^6,2]$	= 65 205	$C[1,3^3]$	= 1588 734/5		
$C[1^6]$	= 1575	$C[1^8]$	= 99 225	$C[2^3,4]$	= 1742 553/5		

The factor $i_k + 1$ is due to the differentiation of $V_{(k)}^{i_k+1}$ and the factor $\{k+1\} \equiv 1 - \delta i_{k+1,0}$ takes care of the requirement $i_{k+1} > 0$. Here $\alpha_{i+1,m+1}(h)$ is the coefficient of h in $A_{i+1,m+1}$, which, according to Eq. (11), can be written in the form

$$\alpha_{i+1,m+1}(h) = \frac{(2m+1)!!}{3^{i+1}\cdot(2i+3)!!\cdot 2^{m+i+2}} C[h] \left(\prod_j (i_j(h)!)\right)^{-1}.$$

B. From $A_{l+1,m,x}\cdot V$

Here g/V is obtained (among other terms) upon differentiation of each one of the functions $h' = (g/V)\cdot(V_{(k)}/V_{(k+1)})$, $k = 0,1,\ldots,n-1$. Since all these functions are contained in $A_{i+1,m}$, each one of them contributes the quantity

$$\alpha_{i+1,m}(h')\cdot(i_k + 1 - \delta_{k,0})\{k+1\}$$

to the coefficient of g in Eq. (5).

C. From $\frac{1}{2} A_{l+1,m}\cdot V_{(1)}$

The term in $A_{i+1,m}$ contributing to g will be $h'' = g/V_{(1)}$, with the coefficient

$$\tfrac{1}{2}\cdot\alpha_{i+1,m}\cdot\{1\}.$$

As indicated by the last factor, there is a contribution only if g contains $V_{(1)}$, i.e., $i_1 > 0$.

D. From $\frac{1}{4} A_{l,m,xxx}$

The general form of terms in $A_{i,m}$ which contribute to g upon triple differentiation is

$$h''' = g\cdot(V_{(j)}V_{(k)}V_{(l)}/V_{(j+1)}V_{(k+1)}V_{(l+1)}),$$
$$0 \leqslant j \leqslant k \leqslant l \leqslant n-1.$$

The coefficient of g in the third derivative of h''' is specified for each one of the following cases.

1. $j+1<k<l-1$

$$6\cdot(i_j+1)(i_k+1)(i_l+1)\{j+1\}\times\{k+1\}\{l+1\}\cdot\alpha_{i,m,}(h''').$$

2. $j+1=k<l-1$

In this case h''' reduces to $g\cdot V_{(j)}V_{(l)}/(V_{(j+2)}\times V_{(l+1)})$ and the coefficient of g in the third derivative is $3(i_j+1)(2i_{j+1}+1)(i_l+1)\{j+2\}\{l+1\}\alpha_{i,m}(h''')$.

3. $j=k<l-1$

Since $h''' = g\cdot V_{(j)}^2 V_{(l)}/(V_{(j+1)}^2 V_{(l+1)})$ we obtain $3\cdot(i_j+2)(i_j+1)(i_l+1)\{\{j+1\}\}\{l+1\}\alpha_{i,m}(h''')$, where

$$\{\{\alpha\}\} \equiv 1 - \delta_{i_a,0} - \delta_{i_a,1} = \begin{cases} 0, & i_a = 0,1, \\ 1, & i_a \geqslant 2. \end{cases}$$

4. $j+1<k=l-1$

$$3(i_j+1)(i_k+1)(2i_{k+1}+1)\{j+1\}\{k+2\}\alpha_{i,m}(h''').$$

5. $j+1<k=l$

$$3(i_j+1)(i_l+2)(i_l+1)\{j+1\}\{\{l+1\}\}\alpha_{i,m}(h''').$$

6. $j+1=k=l-1$

Here $h''' = g\cdot V_{(j)}/V_{(j+3)}$ and the coefficient of g is $(i_j+1)[6i_{j+1}\cdot i_{j+2} + 3(i_{j+1}+i_{j+2}) + 1]\times\{j+3\}\alpha_{i,m}(h''')$.

7. j+1=k=l

$$3(i_j+1)(i_{j+1}+1)^2\cdot\{\{j+2\}\}\cdot\alpha_{i,m}(h''')\,.$$

8. j=k=l−1

$$3(i_j+1)(i_j+2)i_{j+1}\{j+1\}\{j+2\}\alpha_{i,m}(h''')\,.$$

9. j=k=l

$$(i_j+3)(i_j+2)(i_j+1)\{\{\{j+1\}\}\}\alpha_{i,m}(h''')\,.$$

Using these results it is a straightforward, though tedious, task to equate the coefficients of any term g in the recurrence relation, Eq. (5).

The resulting equation is

$$\begin{aligned}
&\sum_{K=0}^{n-1}\frac{i_{K+1}}{6(2i+3)}\left[\left(m+\frac{1}{2}\right)\cdot C\left[\frac{g(V)V_{(K)}}{V_{(K+1)}}\right]+(i+\xi-m)\cdot C\left[\frac{g(V)\cdot V_{(K)}}{V\cdot V_{(K+1)}}\right]\right]-\frac{1}{2}\frac{i_1}{6(2i+3)}\cdot C\left[\frac{g(v)}{V_{(1)}}\right]\\
&=\frac{1}{4}\Bigg\{\sum_{j=0}^{n-5}\sum_{K=j+2}^{n-3}\sum_{l=K+2}^{n-1}6i_{j+1}i_{K+1}i_{l+1}\cdot C\left[g(V)\cdot\frac{V_{(j)}V_{(K)}V_{(l)}}{V_{(j+1)}V_{(K+1)}V_{(l+1)}}\right]\\
&\quad+\sum_{j=0}^{n-3}\sum_{l=j+2}^{n-1}3i_{j+1}(i_{j+1}-1)i_{l+1}\cdot C\left[g(V)\cdot\frac{V_{(j)}^2V_{(l)}}{V_{(j+1)}^2V_{(l+1)}}\right]\\
&\quad+\sum_{j=0}^{n-3}\sum_{l=j+2}^{n-1}3i_{j+1}i_{l+1}(i_{l+1}-1)\cdot C\left[g(V)\cdot\frac{V_{(j)}V_{(l)}^2}{V_{(j+1)}V_{(l+1)}^2}\right]\\
&\quad+\sum_{j=0}^{n-1}i_{j+1}(i_{j+1}-1)(i_{j+1}-2)\cdot C\left[g(V)\cdot\frac{V_{(j)}^3}{V_{(j+1)}^3}\right]\\
&\quad+\sum_{j=0}^{n-4}\sum_{l=j+3}^{n-1}3i_{j+2}(2i_{j+1}+1)i_{l+1}\cdot C\left[g(V)\frac{V_{(j)}V_{(l)}}{V_{(j+2)}V_{(l+1)}}\right]\\
&\quad+\sum_{j=0}^{n-4}\sum_{K=j+2}^{n-2}3i_{j+1}i_{K+2}(2i_{K+1}+1)\cdot C\left[g(V)\frac{V_{(j)}V_{(K)}}{V_{(j+1)}V_{(K+2)}}\right]\\
&\quad+\sum_{j=0}^{n-2}3i_{j+2}(i_{j+2}-1)(i_{j+1}+1)\cdot C\left[g(V)\cdot\frac{V_{(j)}V_{(j+1)}}{V_{(j+2)}^2}\right]\\
&\quad+\sum_{j=0}^{n-2}3i_{j+1}^2i_{j+2}C\left[g(V)\frac{V_{(j)}^2}{V_{(j+1)}V_{(j+2)}}\right]\\
&\quad+\sum_{j=0}^{n-3}i_{j+3}(6i_{j+1}i_{j+2}+3(i_{j+1}+i_{j+2})+1)\cdot C\left[g(V)\frac{V_{(j)}}{V_{(j+3)}}\right]\Bigg\}\,.
\end{aligned}\tag{12}$$

In Appendix B we use Eq. (12) to show the m independence of the coefficients $C[\cdots]$. Using this result the left-hand side of Eq. (12) obtains the form

$$\sum_{k=0}^{n-1}\frac{i_{k+1}}{6(2i+3)}\left(i+\xi+\frac{1}{2}\right)C\left[\frac{gV_{(k)}}{V_{(k+1)}}\right]-\frac{1}{2}\cdot\frac{i_1}{6(2i+3)}C\left[\frac{g}{V_{(1)}}\right],\tag{12'}$$

the right-hand side remaining the same as in Eq. (12).

The number of distinct terms appearing in the recurrence relation is larger than the number of terms in $A_{i+1,m+1}$, because upon differentiation of a typical term of the latter several contributions to the former appear. More precisely, the number of terms in $A_{i+1,m+1}$ is the number of partitions of $2(i+1)$, whereas the number of terms in the recurrence relation is the number of partitions of $2i+3$. Thus the set of linear equations obtained upon equating coefficients in the recurrence relation is redundant. It is shown in Appendix A that one way of obtaining a nonredundant set of equations consists of considering only those terms in which the highest derivative appears linearly, i.e.,

$$g=V^{i_0}V_{(1)}^{i_1}\cdots V_{(n-1)}^{i_{n-1}}V_{(n)}\,.$$

IV. RESULTS

The general expression for $A_{i,m}$ which we now write in the form

$$\begin{aligned}
A_{i,m}=&\frac{(2m-1)!!}{3^i(2i+1)!!\cdot 2^{m+i}}\sum_{\xi=\{i\}}^{2i}\frac{V^{m-i-\xi}}{(m-i-\xi)!}\\
&\cdot\sum_{i_1,i_2,\ldots,i_{2i}}C\left[1^{i_1}2^{i_2}\cdots(2i)^{i_{2i}}\right]\prod_{j=1}^{2i}\left(\frac{V_{(j)}^{i_j}}{i_j!}\right)\\
&\left(\sum_{j=1}^{2i}i_j=\xi,\quad\sum_{j=1}^{2i}j\cdot i_j=2i\right),
\end{aligned}$$

where $\{i\}=1-\delta_{i,0}$ involves a set of coefficients which have been shown in Appendix B to be independent of m. These coefficients satisfy the recurrence relation, Eq. (12'), which was used to obtain the coefficients appearing in Table I. The recurrence relation was further used to derive closed form expressions for certain types of coefficients, which are pre-

TABLE II. The coefficients for certain sets of terms.

$\xi=1$	$[2i]=3^i$
$\xi=2$	$[1,2i-1]=3^i\cdot i$ $[2,2i-2]=3^{i-1}(2i^2+1)$ $[3,2i-3]=3^{i-1}(i^3-i^2/2+i/2-1)$ $[4,2i-4]=(3^{i-1}/5)(2i^4-3i^3+2i^2-3i+5)$
$\xi=3$	$[1^2,2i-2]=3^{i-1}(3i^2-i+1)$ $[1,2,2i-3]=3^{i-1}(2i^3-3i^2/2+3i/2-2)$ $[1,3,2i-4]=(3^{i-1}/10)(10i^4-17i^3+13i^2-22i+28)$ $[2^2,2i-4]=(3^{i-2}/5)(20i^4-32i^3+28i^2-52i+63)$
$\xi=4$	$[1^3,2i-3]=3i(i^3-i^2+i-1)$ $[1^2,2,2i-4]=(3^{i-2}/5)(30i^4-55i^3+54i^2-95i+99)$
$\xi=5$	$[1^4,2i-4]=3^{i-1}(3i^4-6i^3+7i^2-12i+11)$
$\xi=2i$	$[1^{2i}]=(2i+1)[(2i-1)!!]^2$

sented in Table II. These expressions provide some clues to the form of the general expression for an arbitrary coefficient, but the actual derivation of a closed form general expression has not been achieved.

V. THE GENERAL FORM OF AN ISOSPECTRAL TRANSFORMATION

It was pointed out in Sec. II that the general form of an isospectral transformation is

$$V_t=\sum_m \alpha_m A_{m+1,x}\,, \tag{13}$$

where $\{\alpha_m, m=0,1,...\}$ is an arbitrary set of constants. As a matter of fact, if the α_m were arbitrary functions of t the above expression would still be a valid isospectral transformation.

Let

$$F(V)=\sum_m \frac{(2m-1)!!}{2^m}\alpha_m\cdot\frac{V^m}{m!}$$

and note that

$$F_{(k)}(V)\equiv\frac{\partial^k F}{\partial V^k}=\sum_m \frac{(2m-1)!!}{2^m}\alpha_m\cdot\frac{V^{m-k}}{(m-k)!}\,. \tag{14}$$

To obtain the general expression for an isospectral transformation characterized by an arbitrary $F(V)$ we substitute Eqs. (4) and (11) in Eq. (13) and use Eq. (14) to express the sums over m. The resulting equation is

$$V_t=-2\Bigg[\sum_{i=0}\frac{(-1)^{i+1}\cdot\hbar^{2i}}{2^i\cdot 3^i\cdot(2i+1)!!}\cdot\sum_{\xi=\{i\}}^{2i}\sum_{i_1,i_2,\ldots,i_{2i}} C\left[1^{i_1}\cdots(2i)^{i_{2i}}\right]\times\prod_{j=1}^{2i}\left(\frac{V_{(j)}^{i_j}}{i_j!}\right)\cdot F_{(i+\xi)}(V)\Bigg]_x$$

$$\left(\sum_{j=1}^{2i} i_j=\xi,\quad \sum_{j=1}^{2i} j\cdot i_j=2i\right).$$

Comparison of the classical limit of this expression, $V_t=2F_x=2F_{(1)}V_x$ with Eq. (1), indicates that $f=2F_{(1)}$.

In a publication which appeared after the present article was submitted for publication, Torriani[7] presented a conjecture enabling the combinatorial enumeration of the terms appearing in the KdV densities as well as the partial determination of their numerical coefficients. The relation suggested by these conjectures between the results derived in the present article and the representation theory of the symmetric group seem to deserve further attention.

ACKNOWLEDGMENTS

Helpful discussions with Professor R. Pauncz are gratefully acknowledged. We are grateful to the referee for his suggestions.

This research was supported by the Fund for the Promotion of Research at Technion and the Technion V. P. R. Fund—Lawrence Deutsch Research Fund.

APPENDIX A: NONREDUNDANT SET OF EQUATIONS FOR THE COEFFICIENTS

We shall now present a nonredundant set of equations for the coefficients appearing in $A_{i+1,m}$. The number of coefficients is equal to $p_{2(i+1)}$, the number of partitions of $2(i+1)$, and this will also be the number of equations to be presented. These equations will be ordered in such a way that each one of them contains one new coefficient, in addition to coefficients appearing in preceding equations.

The set of equations described will be obtained by equating the coefficient of a particular set of terms in Eq. (5), and in a particular order which we now specify.

Since each term g appearing in Eq. (5) is specified by a partition of $2i+3$, let us arrange these partitions in "increasing order" as in the following example, corresponding to $2i+3=5$:

$\xi=1$ (5)

$\xi=2$ (1) (4)

(2) (3)

$\xi=3$ (1) (1) (3)

(1) (2) (2)

$\xi=4$ (1) (1) (1) (2)

$\xi=5$ (1) (1) (1) (1) (1) .

This ordering can either be specified as increasing in ξ and arranged dictionarywise for each ξ, or, if the partition is read as a number (in the basis $2i+4$), the ordering is according to increasing numerical value ($5<14<23<113<\cdots$ etc.). The same ordering was used in Ref. 6.

Let us first consider the relation obtained by equating the coefficients of

$$g=V^{m-i-\xi}V_{(1)}^{i_1}\cdots V_{(n-1)}^{i_{n-1}}V_{(n)} \tag{A1}$$

in Eq. (5). One of the terms in $A_{i+1,m+1}$ which contribute to g upon differentiation is

$$f_1(n-1)=V^{m-i-\xi}V_{(1)}^{i_1}\cdots V_{(n-1)}^{i_{n-1}+1}=g\cdot V_{(n-1)}/V_{(n)}\,.$$

The other terms obtained by differentiation of $f_1(n-1)$ are of the form

$$\tilde{g}=V^{m-i-\xi}\cdots V_{(j)}^{i_j-1}V_{(j-1)}^{i_j+1}\cdots V_{(n-1)}^{i_{n-1}+1}\,.$$

Since g precedes all possible $\tilde{g}$ in the ordering specified above, it is obvious that $f_1(n-1)$ could not have contribut-

ed to any one of the linear equations corresponding to the coefficients of terms preceding g in the above ordering. Other terms in $A_{i+1,m+1}$ contributing to g upon differentiation, such as

$$f_1(k) = V^{m-i-\xi} V_{(1)}^{i_1} \cdots V_{(k)}^{i_k+1} V_{(k+1)}^{i_{k+1}-1} \cdots V_{(n)}$$

also generate additional terms which precede g, which means that the coefficient of $f_1(k)$ has already appeared in a preceding equation. Since Eq. (5) is a two-dimensional recurrence relation we have to assume that when we attempt to determine $A_{i+1,m+1}$ all the terms in the rhs, i.e., $A_{i+1,m}$ and $A_{i,m}$, are already available. Hence, the only new coefficient appearing is $C[g\cdot V_{(n-1)}/V_{(n)}]$. For a given $1 \leqslant n \leqslant 2i+3$ the number of different g's of the form (A1) is the number of partitions of $2i+3-n$ into integers not larger than $n-1$, $P_{n-1}(2i+3-n)$. In view of the identity

$$\sum_{n=2}^{2i+3} P_{n-1}(2i+3-n) = \sum_{n^*=1}^{2i+2} P_{n^*}(2i+2-n^*) = p(2i+2)$$

it is obvious that the number of equations generated by all g of the form (A1) is equal to the number of coefficients in $A_{i+1,m+1}$. Since each one of these equations, if ordered as specified above, contains one new coefficient, together they determine all $p_{(2i+2)}$ coefficients in $A_{i+1,m+1}$.

APPENDIX B: m INDEPENDENCE OF THE COEFFICIENTS

To demonstrate the fact that the coefficients are not m dependent, we start from the recurrence relation, Eq. (12), which we write in the form

$$\sum_{k=0}^{n-1} \frac{i_{k+1}\cdot m}{6(2i+3)} \left[C\left[\frac{gV_{(k)}}{V_{(k+1)}}\right] - C\left[\frac{g\cdot V_k}{(V\cdot V_{(k+1)})}\right]\right] + \sum_{k=0}^{n-1} \frac{i_{k+1}}{6(2i+3)} \left[\frac{1}{2} C\left[\frac{g\cdot V_{(k)}}{V_{(k+1)}}\right] + (i+\xi) C\left[\frac{g\cdot V_{(k)}}{(V\cdot V_{(k+1)})}\right]\right] - \frac{1}{2}\frac{i_1}{6(2i+3)} C\left[\frac{g}{V_{(1)}}\right] = \cdots . \quad \text{(B1)}$$

All the terms not explicitly written down are neither explicitly m dependent nor containing coefficients corresponding to partitions of $2(i+1)$.

It was shown in Appendix A that each equation introduces one new coefficient. Since for $i = 0,1$ the coefficients were explicitly shown to be m independent, we can invoke the following inductive argument to establish the m independence of all coefficients. Assuming that all coefficients preceding the last one are m independent we have in particular

$$C[gV_{(k)}/V_{(k+1)}] = C[g\cdot V_{(k)}/(V\cdot V_{(k+1)})], \quad k < n-1 .$$

Thus, the only remaining m dependence is in

$$[i_n\cdot m/6(2i+3)]\,[C[g\cdot V_{(n-1)}/V_{(n)}] - C[g\cdot V_{(n-1)}/(V\cdot V_{(n)})]] + [i_n/6(2i+3)]\,[\tfrac{1}{2} C[gV_{(n-1)}/V_{(n)}] + (i+\xi) C[g\cdot V_{(n-1)}/(V\cdot V_{(n)})]] = \cdots \equiv \mu . \quad \text{(B2)}$$

The lowest value of m for which $C[gV_{(n-1)}/V_{(n)}]$ can appear is $m = i+\xi$. For this value of m $C[gV_{(n-1)}/(V\cdot V_{(n)})]$ cannot appear so that Eq. (B2) results in

$$C[g\cdot V_{(n-1)}/V_{(n)}] = \mu\cdot 6(2i+3)/i_n\cdot(i+\xi+\tfrac{1}{2}) .$$

Assuming that up to some m^* $C[V^{m^*}\cdot g\cdot V_{(n-1)}/V_{(n)}] = C[gV_{(n-1)}/V_{(n)}]$ we show that the same applies to $C[V^{m^*+1}\cdot g\cdot V_{(n-1)}/V_{(n)}]$. Note that the latter corresponds to $m = i+\xi+m^*+1$, so that from (B2)

$$C[V^{m^*+1}\cdot g\cdot V_{(n-1)}/V_{(n)}] = (\mu\cdot[6(2i+3)/i_n] + (m^*+1)\cdot C[V^{m^*}\cdot g\cdot V_{(n-1)}/V_{(n)}])/(m+\tfrac{1}{2}) = \mu\cdot 6(2i+3)/i_n(i+\xi+\tfrac{1}{2}) = C[gV_{(n-1)}/V_{(n)}] .$$

This concludes the proof of the m independence of the coefficients.

[1] P. D. Law, Commun. Pure Appl. Math. **21**, 467 (1968).
[2] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, Commun. Pure Appl. Math. **27**, 97 (1974).
[3] G. L. Lamb, Jr., *Elements of Soliton Theory* (Wiley, New York, 1980).
[4] J. Katriel and A. Rosenhouse, Phys. Rev. D **32**, 884 (1985).
[5] R. M. Miura, C. S. Gardner, and M. D. Kruskal, J. Math. Phys. **9**, 1204 (1968).
[6] M. D. Kruskal, R. M. Miura, C. S. Gardner, and N. J. Zabusky, J. Math. Phys. **11**, 952 (1970).
[7] H. H. Torriani, Phys. Lett. A **113**, 345 (1986).

Schrödinger-like equation for relativistic particles

I. B. Goldberg
Racah Institute of Physics, Hebrew University, Jerusalem, 91904 Israel

R. H. Pratt
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 17 November 1986; accepted for publication 11 February 1987)

The Dirac equation in a spherically symmetric screened Coulomb potential is transformed to a modified Schrödinger equation of the form $d^2u/dr^2 + k^2(r)u = 0$. This transformation is induced by expressing the Dirac function as a linear combination of the function u and its derivative du/dr. Various properties of the transformation and of the resulting equations are studied. The close similarity between the modified Schrödinger equation and the Schrödinger equation suggests that methods applied to the Schrödinger equation to derive nonrelativistic relations can be applied to the modified Schrödinger equation to derive the analogous relativistic relations. As an example, this approach is applied to the single channel quantum defect theory to give a new derivation of its relativistic form.

I. INTRODUCTION

Our objective in this paper is to develop a Schrödinger-equation-like formalism for the Dirac equation in a screened Coulomb potential permitting relativistic calculations utilizing procedures previously developed for the nonrelativistic case. In various atomic processes, such as photoionization, Compton scattering, or bremsstrahlung in an ionic field, the potential seen by the free electron at large distances from the ion is a point Coulomb potential, corresponding to the ionic charge Z_{ion}. Closer to the ion, when the free electron is penetrating the charge distribution of the bound electrons, bound electron screening of the nuclear charge is no longer complete and the potential is no longer of the point Coulomb type; in addition exchange and correlation effects also affect the free electron. Following the terminology of the R-matrix method[1] we thus divide the space around the ion into two regions: the external region, where the potential is point Coulomb, and the internal, non-Coulomb region. (Of course at the center of the internal region is another Coulombic region, but now characterized by the nuclear charge Z rather than the ionic charge Z_{ion}.) Several theoretical models have been applied in order to calculate a local atomic potential in the internal region. The most commonly used are the Hartree–Fock–Slater[2] potential in the nonrelativistic case and the Dirac–Fock–Slater[3] potential in the relativistic case. Other methods (local-density approximation,[4] random phase approximation,[5] and Fock approximation[6]), which may go beyond a local potential description for the interior, are used as well.

Except for very special cases, there is no analytic solution to the Dirac equation for a screened central potential in the internal region, and numerical methods must be applied. On the other hand, in the external region (referred to hereafter also as the tail region) the solutions are well known.[7] The complete solution to the Dirac equation, describing both internal and external regions, may be obtained by matching numerically integrated functions in the internal region to an appropriate linear combination of the regular and irregular solutions in the point Coulomb potential. With this matching, carried out at some point in the tail region, one is able to normalize the continuum wave function, and to calculate the phase shift caused by the presence of the potential in the interior region.

The analytic solution of the Dirac equation in the point Coulomb potential may be expressed in terms of confluent hypergeometric functions. These functions appear also in the Coulomb functions which provide the solution of the Schrödinger equation in the point Coulomb potential, and it turns out that the Dirac functions in the point Coulomb potential are given as linear combinations of such functions and their derivatives, albeit for modified arguments. Utilizing this linear combination of solutions, the Dirac equation is transformed into a modified Schrödinger equation[8] in the Coulomb potential, which we may call a modified Coulomb equation.

The question now arises as to whether there exists a similar transformation for the Dirac equation in a screened Coulomb potential, which would result in a modified Schrödinger equation of the form $d^2u/dr^2 + k^2(r)u = 0$, where the effective potential $k^2(r)$ is a function to be determined. It turns out that there is such a transformation. The Dirac functions can be expressed as a linear combination of the function u and its derivative, where u is a solution of the modified Schrödinger equation.

The close similarity between the modified Schrödinger equation and the Schrödinger equation suggests that various relativistic quantities can be obtained with the same methods used to obtain the corresponding nonrelativistic quantities. This approach is applied here to the quantum defect theory, and application of the formalism to a relativistic WKB approximation is in progress.[9] We expect that other features and consequences of the Dirac equation can be obtained by applying nonrelativistic methods to the modified Schrödinger equation.

In Sec. II we will begin by reviewing the Dirac equation for the point Coulomb potential and how it may be written in terms of a modified Schrödinger equation (modified Coulomb equation). In Sec. III we demonstrate that the Dirac equation in a screened Coulomb potential may also be written in terms of a modified Schrödinger equation. In Sec. IV,

 0022-2488/87/061351-09$02.50

as an application of this formalism, we rederive the relativistic quantum defect theory relations.

II. THE DIRAC EQUATION FOR THE POINT COULOMB POTENTIAL

A. The Dirac equation

The Dirac equation for an electron moving in a spherically symmetric potential $V(r)$ is

$$H\psi = (c\alpha\cdot p + \beta mc^2 + V)\psi = E\psi. \tag{2.1}$$

We choose the usual representation in which

$$\alpha = \begin{pmatrix} 0 & \sigma \\ \sigma & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}. \tag{2.2}$$

I is the 2×2 unit matrix and σ are the usual 2×2 Pauli matrices. Here $E = mc^2 + T$ is the total energy of the electron, T being the energy relative to the rest mass.

The eigenfunction ψ is a four-rank spinor which can be written as

$$\psi_{\kappa m} = \frac{1}{r}\begin{pmatrix} g_\kappa\cdot\Omega_{\kappa m} \\ if_\kappa\cdot\Omega_{-\kappa m} \end{pmatrix}, \tag{2.3}$$

where κ is a quantum number which combines angular momentum j and parity l,

$$\kappa = \mp(j+\tfrac{1}{2}) \quad \text{as } j = l \pm \tfrac{1}{2}. \tag{2.4}$$

The spherical spin orbit spinor of rank 2 is given by

$$\Omega_{\kappa m} = \sum_{s=\pm 1/2} \left(l\,\frac{1}{2}\,m-s\,s|jm\right) Y_{l,m-s}X^s; \tag{2.5}$$

$(l\,\frac{1}{2}\,m-s\,s|jm)$ are the Clebsch–Gordan coefficients, $Y_{l,m-s}$ are the spherical harmonics, and the spinors X^s are defined by

$$X^{1/2} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad X^{-1/2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{2.6}$$

With this form for the eigenfunction, the Dirac equation reduces to

$$\frac{dg}{dr} = -\frac{\kappa}{r}g + \left(\frac{1+\epsilon}{\lambda_c} - U\right)f, \tag{2.7a}$$

$$\frac{df}{dr} = \frac{\kappa}{r}f + \left(\frac{1-\epsilon}{\lambda_c} + U\right)g, \tag{2.7b}$$

where $\lambda_c = \hbar/mc$ is the Compton wavelength, $\epsilon = E/mc^2 = 1 + T/mc^2$, and $U = V/\hbar c$.

B. The modified Coulomb equation

When the potential is given by $V = -Ze^2/r$, the Dirac equation (2.7) takes the form

$$\frac{dg}{dr} = \frac{\kappa}{r}g + \left(\frac{1+\epsilon}{\lambda_c} + \frac{a}{r}\right)f, \tag{2.8a}$$

$$\frac{df}{dr} = \frac{\kappa}{r}f + \left(\frac{1-\epsilon}{\lambda_c} - \frac{a}{r}\right)g, \tag{2.8b}$$

where $a = [(Ze^2)/(\hbar c)] = Z\alpha$, α being the fine structure constant.

The solutions of (2.8) are well known[7]; they can be written in terms of solutions of the Schrödinger equation in the same potential, but with modified parameters. Here we will deduce a form for this connection by only considering boundary conditions, to prepare for a generalization of the connection to other potentials. When the energy T is positive ($\epsilon > 1$), the regular solution of (2.8) has the following properties.

(i) As $r\to 0$ we have

$$g = Nr^\gamma, \quad f = [(\kappa+\gamma)/a]g. \tag{2.9}$$

Here $\gamma = (\kappa^2 - a^2)^{1/2}$ and N is a constant which depends on the normalization of the continuum solution of the Dirac equation at large distances.

(ii) As $r\to\infty$ we have

$$\begin{aligned} g &\sim A\sin(pr - \tfrac{1}{2}l\pi + \delta_c + n\ln|2pr|), \\ f &\sim AQ\cos(pr - \tfrac{1}{2}l\pi + \delta_c + n\ln|2pr|), \end{aligned} \tag{2.10}$$

where

$$\delta_c = -\arg\Gamma(\gamma + i\eta) + (l+1-\gamma)(\pi/2) + \xi, \tag{2.11a}$$

and

$$\begin{aligned} &p = (1/\lambda_c)(\epsilon^2-1)^{1/2}, \quad \eta = a\epsilon/(\epsilon^2-1)^{1/2}, \\ &Q = ((\epsilon-1)/(\epsilon+1))^{1/2}, \quad \tan\xi = aQ/(\kappa-\gamma). \end{aligned} \tag{2.11b}$$

The constant A in (2.10), like N in (2.9), is determined by the normalization of the continuum wave functions. When the normalization is on the energy scale, i.e.,

$$\int_0^\infty dr[g(r,E)g(r,E') + f(r,E)f(r,E')] = \delta(E-E'), \tag{2.12}$$

we have

$$A = [(mc^2\cdot\pi\lambda_c\cdot Q)^{1/2}]^{-1}, \tag{2.13a}$$

$$N = A\cdot[|\Gamma(\gamma+i\eta)|e^{\pi\eta/2}/\Gamma(2\gamma+1)] \times (2p\lambda_c)^\gamma(\gamma\cos\xi - \eta\sin\xi). \tag{2.13b}$$

The boundary conditions (2.9) and (2.10) for the large component function g are similar to the boundary conditions of the regular Coulomb function u_c, which provide the continuum solution to the Schrödinger equation in the point Coulomb potential, but corresponding to a noninteger angular momentum parameter $\gamma - 1$. Namely,

$$\begin{aligned} &u_c \underset{r\to 0}{\propto} r^\gamma, \\ &u_c \underset{r\to\infty}{\propto} \sin(pr - \tfrac{1}{2}l\pi + \sigma_c + \eta\ln(|2pr|), \end{aligned} \tag{2.14}$$

where

$$\sigma_c = -\arg\Gamma(\gamma+i\eta) + (l+1-\gamma)(\pi/2) = \delta_c - \xi. \tag{2.15}$$

The combination $v_c = (1/p)(u_c' - (\gamma/r)u_c)$ behaves like r^γ near the origin and like

$$\cos(pr - \tfrac{1}{2}l\pi + \sigma_c + \eta\ln|2pr|)$$

as $r\to\infty$. Thus the boundary conditions suggest the following form for the functions g and f:

$$g = \cos\xi\cdot u + \sin\xi\cdot(1/p)\cdot(u' - (\gamma/r)u), \tag{2.16a}$$

$$f = Q[-\sin\xi\cdot u + \cos\xi\cdot(1/p)\cdot(u' - (\gamma/r)u)]. \tag{2.16b}$$

Indeed, on substituting (2.16) in the Dirac equations we find that both equations, (2.8a) and (2.8b), reduce to the same modified Schrödinger equation in the point Coulomb potential,

$$\frac{d^2u}{dr^2}+\left(p^2+\frac{2a\epsilon}{\lambda_c r}-\frac{\gamma(\gamma-1)}{r^2}\right)u=0, \tag{2.17}$$

which we shall call the *modified Coulomb equation*. The coefficients of the first derivative du/dr in (2.17) vanish,

$$\begin{aligned}&-(\kappa-\gamma)\tan\xi+aQ=0,\\&-(\kappa+\gamma)/(\tan\xi)+a/Q=0,\end{aligned} \tag{2.18}$$

since $\gamma=(\kappa^2-a^2)^{1/2}$ and $\tan\xi=aQ/(\kappa-\gamma)$, according to our previous definitions.

However, if we tried to solve (2.18) to determine γ and ξ, we would find a second solution which also leads to the modified Coulomb equation (2.17), namely $\gamma=-(\kappa^2-a^2)^{1/2}$. The regular solution of (2.17) in this case behaves near the origin like $r^{1-\gamma}$ and the phase σ_c would be obtained by replacing γ with $1-\gamma$ in Eq. (2.15). Either solution leads to the same solution of the Dirac equation.

It follows then that γ and ξ, which we have related initially to the properties of the Dirac functions, may alternately be considered as free parameters, to be determined by the requirements that (a) g and f given by (2.16) solve the Dirac equation (2.8) and (b) u is a solution of the modified Coulomb equation (2.17).

The arguments which led us to the form (2.16) for the Dirac wave functions were based on the boundary condition of the regular continuum functions. However, since we have shown that the Dirac equation (2.8) is equivalent to the modified Coulomb equation (2.17) when g and f have the form (2.16), it follows that this relation is not restricted to the regular continuum case. Thus, when u is an irregular solution of the modified Coulomb equation, we get from (2.16) an irregular solution of the Dirac equation. The relation also holds for the negative energy ($\epsilon<1$) case. In this situation the momentum p and the functions η, Q, and $\sin\xi$ become purely imaginary. Equations (2.16) remain, however, real, and so does the modified Coulomb equation (2.17).

C. The nonrelativistic limit

The choice of the appropriate sign of γ in Eq. (2.16) will now be made by considering the behavior of the modified Coulomb equation (2.17) in the nonrelativistic limit. This limit is obtained by letting the speed of light tend to infinity. Thus we get

$$P\to\sqrt{2mT}/\hbar,\quad a/\lambda_c\to Z_{\rm nuc}e^2/\hbar^2,\quad |\gamma|\to|\kappa|. \tag{2.19}$$

The modified Coulomb equation reduces to the Schrödinger equation for the point Coulomb potential,

$$\frac{d^2u}{dr^2}+\left(\frac{2mT}{\hbar^2}+\frac{2m}{\hbar^2}\cdot\frac{Z_{\rm nuc}e^2}{r}-\frac{L(L+1)}{r^2}\right)u=0, \tag{2.20}$$

where

$$L=\begin{cases}|\kappa|-1, & \gamma>0,\\ |\kappa|, & \gamma<0.\end{cases} \tag{2.21}$$

We require that in the nonrelativistic limit, the solution of (2.20) coincides (up to a sign) with the large component function g. Therefore $L(L+1)=\kappa(\kappa+1)$ and $\mathrm{sgn}(\gamma)=-\mathrm{sgn}(\kappa)$.[10]

With this choice of the sign γ we get in the nonrelativistic limit

$$g=\mathrm{sgn}(\kappa)\cdot u,\quad f=\frac{1}{c}\frac{\hbar}{2m}r^{-\kappa}\frac{d}{dr}(r^{\kappa}\cdot g). \tag{2.22}$$

(Note that other conventions are often used; for example in Rose[7] the sgn κ factor is omitted.)

D. Free particle case

The Dirac equation for a free particle in spherical coordinates is given by Eq. (2.8) with $a=0$. The expressions we have derived are valid for this case, for which we have

$$\gamma=-\kappa,\quad \eta=0,\quad \tan\xi=0, \tag{2.23}$$

and the modified Coulomb equation reduces to the Bessel equation

$$\frac{d^2u}{dr^2}+\left(p^2-\frac{\kappa(\kappa+1)}{r^2}\right)u=0. \tag{2.24}$$

The solution of (2.24), regular at the origin, is

$$u=\tilde{A}rj_l(pr), \tag{2.25}$$

where $\tilde{A}$ is a constant of normalization. The corresponding Dirac functions are obtained from Eq. (2.16),

$$\begin{aligned}&g=u=\tilde{A}rj_l(pr),\\&f=(Q/p)(u'+(\kappa/r)u)=(\kappa/|\kappa|)\tilde{A}Qrj_{\bar{l}}(pr),\end{aligned} \tag{2.26}$$

where $\bar{l}=l-\kappa/|\kappa|$. The solution of the Dirac equation which is irregular at the origin is obtained by replacing in Eq. (2.26) the Bessel functions j_l and $j_{\bar{l}}$ with the Neumann functions n_l and $n_{\bar{l}}$, respectively.

III. SCREENED COULOMB POTENTIAL AND THE MODIFIED SCHRÖDINGER EQUATION

We consider now the case of an electron in a screened Coulomb potential. We assume that the potential is of the form

$$U(r)=-(a_0/r)s(r), \tag{3.1}$$

where $a_0=\alpha Z_{\rm nuc}$ and $s(r)$ is a smooth and monotomic function such that

$$s(r)=\begin{cases}1, & r=0,\\ s_t, & r\geqslant r_t.\end{cases} \tag{3.2}$$

The tail radius r_t is the position at which screening has its full effect. Beyond this point the potential is point Coulomb ($s_t>0$) or it vanishes ($s_t=0$),

$$U(r)=-a_t/r,\quad r\geqslant r_t, \tag{3.3}$$

where $a_t=a_0\cdot s_t$.

A. The modified Schrödinger equation

We look for a solution of the Dirac equation of the form (2.16) in the screened potential, with u again the solution of

a Schrödinger-like equation. Since the effective charge seen by the electron is now position dependent, we now must allow ξ and γ to be also position dependent. Thus we now have three functions u, γ, ξ; we shall use our freedom in the choice of γ and ξ to obtain a convenient result for u. (Other choices might also be useful.)

When we substitute Eqs. (2.16) in the Dirac equation, we find that the function u must satisfy two (second-order) differential equations, resulting from the two equations (2.7a) and (2.7b). It turns out that these two equations for u are the same if we require that the coefficients of the first derivative du/dr vanish. Thus we get two equations connecting the "rotation angle" ξ and the "angular momentum parameter" γ with the screening function $s(r)$

$$r\frac{d\xi}{dr} = -(\kappa-\gamma(r))\tan\xi(r) + a_0 s(r) Q, \tag{3.4a}$$

$$r\frac{d\xi}{dr} = -\frac{\kappa+\gamma(r)}{\tan\xi(r)} + \frac{a_0 s(r)}{Q}. \tag{3.4b}$$

When these equations are satisfied, u is a solution of the equation

$$\frac{d^2u}{dr^2} + \left(p^2 + \frac{2a_0\epsilon}{\lambda_c r} w(r) - \frac{\gamma(r)[\gamma(r)-1]}{r^2}\right)u = 0, \tag{3.5}$$

where

$$w(r) = \frac{\lambda_c}{2a_0\epsilon}\left[-\frac{d\gamma}{dr} + p\left(\frac{\kappa+\gamma(r)}{\tan\xi(r)} + (\kappa-\gamma(r))\tan\xi(r)\right)\right]. \tag{3.6}$$

Equation (3.5) will be referred to as the modified Schrödinger equation. It should be emphasized that Eqs. (3.4) are energy dependent through the parameter $Q = \sqrt{(\epsilon-1)/(\epsilon+1)}$. Therefore, both $\tan\xi(r)$ and $\gamma(r)$ are energy dependent and so is the effective screening function $w(r)$. However, as we will show below, we can obtain both ξ and γ by solving for one energy-*independent* function.

Before solving the modified Schrödinger equation (3.5) we have to solve Eqs. (3.4), which are the generalization of Eqs. (2.18) for the case of a screened Coulomb potential. We eliminate $\gamma(r)$ between (3.4a) and (3.4b) and get

$$r\frac{d\tan\xi}{dr} = \frac{a_0 s}{Q}\tan^2\xi - 2\kappa\tan\xi + a_0 s Q, \tag{3.7a}$$

$$\gamma = \{\kappa(\tan^2\xi - 1) + a_0 s(1-Q^2)(\tan\xi/Q)\} \times\{\tan^2\xi + 1\}^{-1}. \tag{3.7b}$$

There are two "natural" initial values for the solution of Eqs. (3.7), which correspond to the values of ξ and γ in a point Coulomb potential of an appropriate nuclear charge, which depends on the point r_{init} at which the integration of Eq. (3.7a) is started. Namely,

$$\gamma_0 = -\operatorname{sgn}(\kappa)\sqrt{\kappa^2 - a_0^2}, \qquad r_{\text{init}} = 0, \tag{3.8a}$$

$$\tan\xi_0 = a_0 Q/(\kappa-\gamma_0),$$

and

$$\gamma_t = -\operatorname{sgn}(\kappa)\sqrt{\kappa^2 - a_t^2}, \qquad r_{\text{init}} = r_t. \tag{3.8b}$$

$$\tan\xi_t = a_t Q/(\kappa-\gamma_t),$$

We will choose that initial value for which the solution of Eqs. (3.7) is smooth and well behaved, as will be discussed in the next section.

B. The subsidiary equation

We transform now Eq. (3.7a) to an energy-independent form. We define the subsidiary function

$$\theta(r) = [\tan\xi(r)]/Q, \tag{3.9}$$

and we get

$$r\frac{d\theta}{dr} = a_0 s\theta^2 - 2\kappa\theta + a_0 s. \tag{3.10}$$

Equation (3.10), which will be referred to as the subsidiary equation, is energy independent, and so are the natural Coulombic initial values which correspond to Eqs. (3.8)

$$\theta_0 = a_0/(\kappa-\gamma_0), \quad r_{\text{init}} = 0, \tag{3.11a}$$

$$\theta_t = a_t/(\kappa-\gamma_t), \quad r_{\text{init}} = r_t. \tag{3.11b}$$

It follows, then, that when one of these initial values is chosen, $\theta(r)$ is energy independent everywhere. This property of $\theta(r)$ makes the numerical solution of the modified Schrödinger equation (3.5) more tractable than would have appeared from Eqs. (3.4).

The question now arises as to whether there is a solution $\theta(r)$ which varies smoothly between these two initial values as r moves from the origin to the tail region or vice versa. Such a solution, if it exists, has a Coulombic feature in the two regions where the potential is point Coulombic. It is usually impossible for a solution of a first-order differential equation to satisfy two boundary conditions. However, the solution $\theta(r)$ of the subsidiary equation (3.10) has the property that its limiting values as $r\to 0$ and $r\to\infty$ are independent of the initial values (see Secs. 1 and 2 in the Appendix). In particular we have

$$\lim_{r\to 0}\theta(r) = \theta_0, \quad \text{for } \kappa<0, \quad r_{\text{init}}>0,$$

$$\lim_{r\to\infty}\theta(r) = \theta_t, \quad \text{for } \kappa>0, \quad r_{\text{init}}<\infty.$$

Therefore, when $\kappa<0$ we start the integration at $r_{\text{init}} = r_t$ with the initial value θ_t and integrate inward. The solution $\theta(r)$ is a monotonically increasing function (see Appendix C) satisfying the two boundary values (3.11). When $\kappa>0$ we start the integration at $r_{\text{init}} = 0$ with the initial value θ_0 and integrate outward. The solution $\theta(r)$ in this case is a monotonically decreasing function which approaches θ_t as r tends to infinity (see Appendix D). At $r = r_t$, however, we have (for $\kappa>0$) $\theta(r_t) > \theta_t$ and the boundary condition at this point is not satisfied. As far as the solution of the Dirac equation is concerned, this poses no special difficulty, as we shall see later.

C. Some properties of the modified Schrödinger equation and its solution

With the function $\theta(r)$ given, we calculate $\tan\xi(r)$ and $\gamma(r)$ by (3.9) and (3.7b),

$$\tan\xi(r) = Q\cdot\theta(r),$$
$$\gamma(r) = \{\kappa[Q^2\theta^2(r) - 1] + a_0 s(r)(1-Q^2)\theta(r)\} \times\{Q^2\theta^2(r)+1\}^{-1}. \tag{3.12}$$

Then we integrate the modified Schrödinger equation (3.5) to obtain u, and using Eq. (2.16) we obtain the solution g and f to the Dirac equation. Let us look at some properties of Eq. (3.5) and of its solution u.

(i) *Near the origin* ($r\to 0$). Since $\theta(0)=\theta_0$, we get from (3.12) and (3.10) that $\gamma(0)=\gamma_0$ and $\lim_{r\to 0} r(d\gamma/dr) = 0$. Therefore the dominant term of the modified Schrödinger equation near the origin is the centrifugal potential $\gamma_0(\gamma_0-1)/r^2$. This assures $r^{\gamma_0}(r^{1-\gamma_0})$ behavior of u near the origin for the regular (irregular) solution for $\kappa<0$, or for the irregular (regular) solution when $\kappa>0$. This, in turn, assures $r^{|\gamma_0|}(r^{-|\gamma_0|})$ behavior for the regular (irregular) solution (2.14) of the Dirac equation.

(ii) *The tail region* ($r\geqslant r_t$). When $\kappa<0$ we have $\theta(r\geqslant r_t)=\theta_t$ and by (3.6) and (3.12) we get in this region

$$\gamma(r)=\gamma_t,\quad \tan\xi(r)=\tan\xi_t,\quad w(r)=s_t\quad (r\geqslant r_t). \tag{3.13}$$

Thus the modified Schrödinger equation (3.5) reduces to the modified Coulomb equation (2.17) for charge $Z = Z_{\rm nuc}\cdot s_t$ (ionic potential, $s_t>0$) or to the Bessel equation (2.24) (neutral atom potential, $s_t=0$).

When $\kappa>0$ the boundary value (3.8b) is not satisfied and the modified Schrödinger equation (3.5) does not reduce to either Eq. (2.17) or to Eq. (2.24). However, since $\lim_{r\to\infty}\theta(r)=\theta_t$, $\xi(r)$ and $\gamma(r)$ approach the Coulombic values ξ_t and γ_t as $r\to\infty$. The asymptotic behavior of the solution $u(r)$ of Eq. (3.5) is the same as the asymptotic behavior of a particular solution u_t of the modified Coulomb equation (2.17) with $\gamma=\gamma_t$ and $a=a_t$, as we will now show.

The solution (g,f) of the Dirac equation in the tail region can be obtained in two ways: (i) by solving the modified Schrödinger equation (3.5) and substituting its solution in Eqs. (2.16); and (ii) by solving the modified Coulomb equation (2.17) with $\gamma=\gamma_t$ and $\xi=\xi_t$, and substituting its solution in Eqs. (2.16). We denote by u_t, that particular solution of (2.17) for which the two solutions of the Dirac equation coincide at $r=r_t$ (and therefore for $r\geqslant r_t$),

$$g_t(r) = \cos\xi_t\cdot u_t(r) + \sin\xi_t\cdot v_t(r),$$
$$f_t(r) = Q[-\sin\xi t\cdot ut(r) + \cos\xi t\cdot vt(r)], \tag{3.14}$$

where $v_t = (1/p)[u_t'(r) - (\gamma_t/r)u_t(r)]$.

By equating the two solutions we get

$$u(r) = \cos(\xi_t-\xi)\cdot u_t(r) + \sin(\xi_t-\xi)\cdot v_t(r),$$
$$v(r) = -\sin(\xi_t-\xi)u_t(r) + \cos(\xi_t-\xi)\cdot v_t(r), \tag{3.15}$$

where $v(r) = (1/p)[u'(r) - (\gamma(r)/r)u(r)]$ and $\xi = \arctan[Q\cdot\theta(r)]$. Since $\xi(r)\to_{r\to\infty}\xi_t$ we see that the asymptotic behavior of $u(r)$ and its derivative $u'(r)$ is the same as the asymptotic behavior of the Coulomb function $u_t(r)$ and its derivative $u_t'(r)$, respectively.

(iii) *The nonrelativistic limit.* The boundary values θ_0 and θ_t [Eq. (3.11)] are given in this limit by

$$\theta_0\approx(1/c)\cdot(e^2Z_{\rm nuc}/2h\kappa),\quad \theta_t\approx\theta_0\cdot s_t.$$

Therefore from Eq. (3.10) we find that to the lowest order in $1/c$, $\theta(r)$ is given by $\theta(r)=(1/c)\hat\Theta(r)$, where $\hat\Theta(r)$ does not depend on c. On expanding Eqs. (3.6) and (3.12) in 1/c we get

$$\gamma(r)\approx -\kappa + (1/c^2)(e^2Z_{\rm nuc}/\hbar)\cdot s(r)\cdot\hat\Theta(r),$$
$$\tan\xi(r)\approx(1/c^2)\sqrt{(T/2m)}\hat\Theta(r), \tag{3.16}$$
$$w(r)\approx s(r),$$

and the modified Schrödinger equation reduces to the Schrödinger equation for a screened Coulomb potential,

$$\frac{d^2u}{dr^2} + \left(\frac{2mT}{\hbar^2} + \frac{2me^2Z_{\rm nuc}}{\hbar^2}\cdot\frac{s(r)}{r} - \frac{l(l+1)}{r^2}\right)u = 0, \tag{3.17}$$

and again, g and f are given by (2.22).

IV. APPLICATION: DERIVATION OF RELATIVISTIC QUANTUM DEFECT THEORY

Here we illustrate the application of these ideas, using as our example a rederivation of single channel relativistic quantum defect theory[11,12] with the methods of the nonrelativistic derivation. Applying a nonrelativistic derivation based on the modified Schrödinger equation (3.5), we obtain the corresponding relativistic theory from Eqs. (2.16).

It was shown in Sec. III that the asymptotic behavior of the solution u of Eq. (3.5) is given by the function u_t, which is a solution of the modified Coulomb equation

$$\frac{d^2u}{dr^2} + \left(p^2 + \frac{2a\epsilon}{\lambda_c r} + \frac{\gamma(\gamma-1)}{r^2}\right)u = 0, \tag{4.1}$$

where

$$a = \alpha Z_{\rm nuc}s_t, \tag{4.2a}$$

and

$$\gamma = \begin{cases}\gamma_t, & \kappa<0,\\ 1-\gamma_t, & \kappa>0.\end{cases} \tag{4.2b}$$

We consider the case of an ionic potential ($a>0$) and we apply Seaton's quantum defect theory[13] to the modified Schrödinger equation. We follow the QDT method as closely as possible, bearing in mind that the angular momentum parameter is not an integer in our case.

A. Solutions of the modified Coulomb equation

We first change the variable r in (4.1) to $x = a\epsilon r/\lambda_c$ and we obtain the modified Coulomb equation in its dimensionless form

$$\frac{d^2u}{dx^2}+\left(-\frac{1}{\sigma^2}+\frac{2}{x}-\frac{\gamma(\gamma-1)}{x^2}\right)u=0, \tag{4.3}$$

where

$$\sigma=\begin{cases} i\eta = ia\epsilon/\sqrt{\epsilon^2-1}, & \epsilon>1, \\ \nu = a\epsilon/\sqrt{1-\epsilon^2}, & \epsilon<1. \end{cases} \tag{4.4}$$

We consider the following solutions to Eq. (4.3) (Ref. 14):

$$y_1(\sigma,\gamma;x)=[(2x)^\gamma/\Gamma(2\gamma)]\,e^{-x/\sigma}M(\gamma-\sigma,\ 2\gamma;2x/\sigma),$$
$$y_2(\sigma,\gamma;x)=y_1(\sigma,\ 1-\gamma;\ x), \tag{4.5}$$

where $M(\alpha,\beta;Z)$ is the confluent hypergeometric function as defined in Ref. 15. It was shown by Seaton[13] that these functions are analytic functions of $1/\sigma^2$ and therefore they are analytic functions of the energy for $\epsilon\geqslant\epsilon_0>0$. Moreover, since γ is not an integer or half-integer, these functions are algebraically independent.

B. Asymptotic behavior

From the asymptotic behavior of the confluent hypergeometric function[16] we get the following expression for y_1:

$$y_1 \underset{x\to\infty}{\sim} \sigma^\gamma\left\{\frac{e^{x/\sigma}}{\Gamma(\gamma-\sigma)}\cdot\left(\frac{2x}{\sigma}\right)^{-\sigma} + \frac{e^{-x/\sigma}}{\Gamma(\gamma+\sigma)}\cdot\left(\frac{2x}{\sigma}\right)^{\sigma}e^{-i\pi(\gamma-\sigma)}\right\}. \tag{4.6}$$

(i) Below threshold ($\epsilon<1$). In this case $\sigma=\nu$ and the function grows exponentially as $x\to\infty$. The function y_1 can represent a bound state only when the exponentially growing part of the function vanishes. This is the case when $\gamma-\sigma$ is a nonpositive integer. The conditions for the energy levels are therefore

$$\nu(\epsilon_{n,\kappa})=a\epsilon_{n,\kappa}/\sqrt{1-\epsilon^2_{n,\kappa}}=\gamma-|\kappa|+n$$
$$(n=|\kappa|,|\kappa|+1,|\kappa|+2,\ldots). \tag{4.7}$$

This relation is equivalent to the expression for the energy levels of hydrogenlike atoms in the Dirac theory.

(ii) Above threshold ($\epsilon>1$). In this case $\sigma=i\eta$, the function remains bounded as $x\to\infty$, and we have

$$y_1 \underset{x\to\infty}{\sim} [2n^\gamma e^{-\pi\eta/2}/|\Gamma(\gamma+i\eta)|]\sin\omega, \tag{4.8}$$

where

$$\omega=x/\eta+\eta\ln(2x/\eta)+(\pi/2)(1-\gamma)-\arg\Gamma(\gamma+i\eta). \tag{4.9}$$

The asymptotic behavior of $y_2(\sigma,\gamma;x)$ is obtained from Eq. (4.6) by replacing γ with $1-\gamma$,

$$y_2 \underset{x\to\infty}{\sim} \sigma^{1-\gamma}\left\{\frac{e^{x/\sigma}}{\Gamma(1-\gamma-\sigma)}\left(\frac{2x}{\sigma}\right)^{-\sigma} + \frac{e^{-x/\sigma}}{\Gamma(1-\gamma+\sigma)}\left(\frac{2x}{\sigma}\right)^{\sigma}e^{-i\pi(1-\gamma-\sigma)}\right\}. \tag{4.10}$$

(i) Below threshold ($\epsilon<1$). Here this function also grows exponentially as $x\to\infty$. Since the function is irregular at the origin, bound states of the point Coulomb potential are not given by the function y_2 for hydrogenlike atoms. However, in the case of a screened potential with a Coulomb tail, bound states are represented in the exterior region by the linear combination of y_1 and y_2 which decays exponentially as $x\to\infty$. Therefore we have to look for the linear combination of y_1 and y_2 in which the exponentially growing part vanishes.

Using the relation $\Gamma(Z)\Gamma(1-Z)=\pi/\sin\pi Z$, we find that the exponentially growing part of y_1 is given by $K\sin\pi(\gamma-\nu)$, where

$$K=(1/\pi)\nu^\gamma\Gamma(1-\gamma+\nu)(2x/\nu)^{-\nu}e^{x/\nu}. \tag{4.11}$$

To describe bound states (for the screened case) in a convenient form, we look for a linear combination of y_1 and y_2 in which the exponentially growing part is given by $-K\cos\pi(\gamma-\nu)$. The coefficients of the linear combination are given by

$$K_1=-\cot 2\pi\gamma,$$
$$K_2=-\nu^{2\gamma-1}[\Gamma(1-\gamma+\nu)/\Gamma(\gamma+\nu)]\cdot(\sin 2\pi\gamma)^{-1}, \tag{4.12}$$

and we get

$$y_0=K_1y_1+K_2y_2=-K\cos\pi(\gamma-\nu) - \nu^\gamma\cdot[e^{-x/\nu}/\Gamma(\gamma+\nu)](2x/\nu)^\nu[\cot 2\pi\gamma e^{-i\pi(\gamma-\nu)} + (\sin 2\pi\gamma)^{-1}e^{-i\pi(1-\gamma-\nu)}]. \tag{4.13}$$

The coefficient K_2 in Eq. (4.12) can be analytically continued to energies above threshold[17]

$$K_2(\epsilon>1)=-\sigma^{2\gamma-1}\frac{\Gamma(1-\gamma+\sigma)}{\Gamma(\gamma+\sigma)}\cdot\frac{1}{\sin 2\pi\gamma} = -\eta^{2\gamma-1}\left|\frac{\Gamma(1-\gamma+i\eta)}{\Gamma(\gamma+i\eta)}\right|\frac{1}{\sin 2\pi\gamma}. \tag{4.14}$$

Therefore the function y_0 defined in Eq. (4.13) is an analytic function of the energy near the threshold.

(ii) Above threshold ($\epsilon>1$). Here we have

$$y_2 \underset{x\to\infty}{\sim} [2\eta^{1-\gamma}e^{-\pi\eta/2}/|\Gamma(1-\gamma+i\eta)|]\sin(\tilde{\omega}), \tag{4.15}$$

where

$$\tilde{\omega}=x/\eta+\eta\ln(2x/\eta)+\pi\gamma/2-\arg\Gamma(1-\gamma+i\eta). \tag{4.16}$$

In the case of a screened potential, we characterize continuum wave functions by phase shifts. It is thus more convenient to have two independent solutions to the modified Coulomb equation (4.3); the regular solution behaves asymptotically like $\sin(\omega)$ and the irregular solution like $\cos(\omega)$. We define therefore

$$L_1=-\cot(\tilde{\omega}-\omega),$$
$$L_2=\eta^{2\gamma-1}\left|\frac{\Gamma(1-\gamma+i\eta)}{\Gamma(\gamma+i\eta)}\right|\frac{1}{\sin(\tilde{\omega}-\omega)} = -K_2\frac{\sin 2\pi\gamma}{\sin(\tilde{\omega}-\omega)}, \tag{4.17}$$

and we get

$$\tilde{y}_0=L_1y_1+L_2y_2 \underset{x\to\infty}{\sim} [2\eta^\gamma e^{-\pi\eta/2}/|\Gamma(\gamma+i\eta)|]\cos(\omega). \tag{4.18}$$

At threshold we have

$$K_1(\epsilon=1)=L_1(\epsilon=1)=-\cot 2\pi\gamma,$$
$$K_2(\epsilon=1)=L_2(\epsilon=1)=-(\sin 2\pi\gamma)^{-1}, \tag{4.19}$$

C. Quantum defect and phase shift of the solution of the modified Schrödinger equation

Let $u(r)$ be a solution of Eq. (3.5), regular at the origin. Let $u_t(r)$ be a solution of Eq. (4.1) such that in the tail region ($r\geqslant r_t$); for $\kappa<0$ we have $u_t(r)=u(r)$, and for $\kappa>0$ Eqs. (3.15) are satisfied. We will use the basis functions defined above and we change therefore the variable r to $x=a\epsilon r/\lambda_c$,

$$y(x)=u(r),\quad y_t(x)=u_t(r). \tag{4.20}$$

(i) Below threshold ($\epsilon<1$). We write $y_t(x)$ as

$$y_t(x)=c(\cos\pi\mu\cdot y_1(x)+\sin\pi\mu\cdot y_0(x)), \tag{4.21}$$

where c is a constant and μ is an analytic function of the energy.[18]

For the function $y(x)$ to remain bounded, so that corresponding Dirac functions g and f represent a bound state, the exponentially growing part of $y_t(x)$ must vanish. This condition is satisfied when $\sin\pi(\gamma-\nu-\mu)=0$, or

$$\gamma-\nu(\epsilon)-\mu(\epsilon)=|\kappa|-n=\text{integer}, \tag{4.22}$$

from which the eigenvalue $\epsilon_{n,k}$ is obtained. We write Eq. (4.22) as $\nu=\gamma-|k|+(n-\mu)$ and compare it with (4.7). It follows that $\mu(\epsilon_{n,k})$ is the quantum defect caused by the non-Coulomb part of the modified Schrödinger equation (3.5).

(ii) Above threshold ($\epsilon>1$). We write $y_t(x)$ as

$$y_t(x)=c(\cos\delta\cdot y_1(x)+\sin\delta\cdot\tilde{y}_0(x))$$
$$\underset{x\to\infty}{\sim} c[2\eta^{\gamma}e^{-\pi\eta/2}/|\Gamma(\gamma+i\eta)|]\sin(\omega+\delta). \tag{4.23}$$

Thus, $\delta(\epsilon)$ is the phase shift caused by the non-Coulomb part of Eq. (3.5). We may also express $y_t(x)$ in the form of Eq. (4.21) for energies above threshold. Comparing coefficients of y_1 and y_2 we get

$$\cot\delta(\epsilon)=\cot(\tilde{\omega}-\omega)-[(\sin 2\pi\gamma)/\sin(\tilde{\omega}-\omega)]$$
$$\times(\cot\pi\mu(\epsilon)-\cot 2\pi\gamma). \tag{4.24}$$

Near threshold, for $\eta\gg 1$, we have

$$\cot\delta(\epsilon)=(1-\cos 2\pi\gamma e^{-2\pi\eta})\cot\pi\mu(\epsilon)$$
$$-\sin 2\pi\gamma e^{-2\pi\eta}, \tag{4.25}$$

and at threshold

$$\cot\delta(\epsilon=1)=\cot\pi\mu(\epsilon=1). \tag{4.26}$$

Equations (4.25) and (4.26) are the quantum defect relations of the modified Schrödinger equation. As we will see below, $\mu(\epsilon)$ and $\delta(\epsilon)$ are also the quantum defect and phase shift of the Dirac functions, and therefore Eqs. (4.24)–(4.26) are the relativistic quantum defect relations. They are identical with the relations obtained by Johnson and Cheng.[11]

D. Quantum defect and phase shift of the Dirac functions

We substitute now an unnormalized regular solution $u(r)$ of the modified Schrödinger equation (3.5) in the expressions (2.16) and we get an unnormalized regular solution of the Dirac equation,

$$\begin{pmatrix}g\\f\end{pmatrix}=\begin{pmatrix}\cos\xi\\-Q\sin\xi\end{pmatrix}u+\frac{1}{p}\begin{pmatrix}\sin\xi\\Q\cos\xi\end{pmatrix}\left(u'-\frac{\gamma}{r}u\right), \tag{4.27}$$

where ξ and γ are the solution of Eqs. (3.10) and (3.12).

In the tail region ($r\geqslant r_t$) the Dirac functions g and f are the same as the functions defined in Eq. (3.14) in terms of the Coulomb function $u_t(r)$. The function $u_t(r)$ can be expressed as a linear combination of the basis functions $u_1(r)$ and $u_0(r)$,

$$u_1(r)=y_1(x),\quad u_0(r)=y_0(x), \tag{4.28}$$

and therefore g and f are linear combinations, with the same coefficients, of the two solutions (g_1,f_1) and (g_0,f_0) obtained by substituting u_1 and u_0 in Eqs. (2.16), respectively.

(i) Below threshold ($\epsilon<1$). From Eq. (4.21) we get

$$\begin{pmatrix}g\\f\end{pmatrix}=c\cdot\left\{\cos\pi\mu\begin{pmatrix}g_1\\f_1\end{pmatrix}+\sin\pi\mu\begin{pmatrix}g_0\\f_0\end{pmatrix}\right\},\quad r\geqslant r_t. \tag{4.29}$$

When μ satisfies Eq. (4.22), g and f decay exponentially as $r\to\infty$, and thus represent a bound state. Therefore, the quantum defect function $\mu(\epsilon)$ defined for the solutions of the modified Schrödinger equation is identical with the quantum defect function of the Dirac functions.

(ii) Above threshold ($\epsilon>1$). Let $(\tilde{g}_0,\tilde{f}_0)$ be the Dirac functions obtained by substituting $\tilde{u}_0(r)=\tilde{y}_0(x)$ in Eqs. (2.16). Then, from Eq. (4.23) we get

$$\begin{pmatrix}g\\f\end{pmatrix}=c\left\{\cos\delta\begin{pmatrix}g_1\\f_1\end{pmatrix}+\sin\delta\begin{pmatrix}\tilde{g}_0\\\tilde{f}_0\end{pmatrix}\right\},\quad r\geqslant r_t. \tag{4.30}$$

The asymptotic behavior of g_1 and f_1 is given by (2.10). The asymptotic behavior of $\tilde{g}_0$ and $\tilde{f}_0$ is obtained from (4.18),

$$\tilde{g}_0\sim A\cos(pr-\tfrac{1}{2}l\pi+\delta_c+\eta\ln(2pr)),$$
$$\tilde{f}_0\sim -AQ\sin(pr-\tfrac{1}{2}l\pi+\delta_c+\eta\ln(2pr)), \tag{4.31}$$

where the relativistic Coulomb phase shift is given by Eq. (2.11a). It follows from (4.30) that $\delta(\epsilon)$ is indeed the phase shift, with respect to the regular solution of the Dirac equation in the point Coulomb potential, caused by the non-Coulomb part of the potential. As we already saw, $\delta(\epsilon)$ is also the phase shift, with respect to the Coulomb function $u_1(r)$, caused by the non-Coulomb part in the modified Schrödinger equation (3.5).

Thus, Eqs. (4.24)–(4.26), which connect the phase shift $\delta(\epsilon)$ and the quantum defect $\mu(\epsilon)$ of the modified Schrödinger equation, are also the desired single channel relativistic quantum defect relations.

V. CONCLUSION

We have shown that the Dirac equation in a screened Coulomb potential can be transformed to a modified Schrödinger equation [Eq. (3.5)] with an effective potential. The close similarity between the modified Schrödinger equation and the Schrödinger equation suggests that various proper-

ties of the Dirac equation can be derived from the corresponding properties of the Schrödinger equation. We have demonstrated the usefulness of this approach by obtaining the relativistic quantum defect relations from the corresponding relations of the modified Schrödinger equation. We expect this approach to be useful in obtaining other properties associated with the Dirac equation.

ACKNOWLEDGMENT

Helpful discussions with Mr. J. C. Parker, III and Dr. J. Stein are gratefully acknowledged.

APPENDIX: PROPERTIES OF THE SOLUTION OF THE SUBSIDIARY EQUATION (3.10)

1. Solution near the origin

We expand $\theta(r)$ and $s(r)$ in power series,

$$\theta(r) = \sum_{l=0}^{\infty} \tilde{\theta}_l r^l, \qquad s(r) = \sum_{l=0}^{\infty} s_l r^l = 1 + \sum_{l=1}^{\infty} s_l r^l. \tag{A1}$$

On substituting in Eq. (3.10) and equating coefficients of the same order in r we get the following set of equations:

$$a_0 \tilde{\theta}_0^2 - 2\kappa \tilde{\theta}_0 + a_0 = 0, \tag{A2a}$$

$$\tilde{\theta}_l = a_0[(c_l + s_l)/(2\kappa + l - 2a_0\theta_0)], \quad l = 1,2,..., \tag{A2b}$$

where

$$C_l = \sum_{\substack{m,n=0 \\ m+n<l}}^{l-1} S_{l-m-n} \cdot \tilde{\theta}_m \cdot \tilde{\theta}_n. \tag{A3}$$

The point $r = 0$ is a singular point of the subsidiary equation (3.10). The behavior of its solution near the origin depends on the sign of the quantity

$$S = a_0 \tilde{\theta}_0 - \kappa,$$

where $\tilde{\theta}_0$ is one of the two solutions of Eq. (A2a),[19] $\tilde{\theta}_{0,1} = \theta_0$; $\tilde{\theta}_{0,2} = 1/\theta_0$ [see (3.11a)]. When $S > 0$, the point $(r = 0, \theta = \tilde{\theta}_0)$ is a nodal point, and all the curves $\theta(r)$ reach the point $\tilde{\theta}_0$ as $r \to 0$. On the other hand, when $S < 0$, the point $(r = 0, \theta = \theta_0)$ is a saddle point, and apart from two lines (the principal lines), no other curve $\theta(r)$ reaches the point $\tilde{\theta}_0$. For the solution $\tilde{\theta}_0 = \theta_0$, which corresponds to the initial value (3.11a), we have

$$\mathrm{sgn}(S) = -\mathrm{sgn}(k),$$

and therefore $(0,\theta_0)$ is a nodal point for $\kappa < 0$ and a saddle point for $\kappa > 0$. On the other hand, the point $(0, 1/\theta_0)$ is a saddle point for $\kappa < 0$ and a nodal point for $\kappa > 0$. It follows that for $\kappa < 0$ we get $\lim_{r \to 0} \theta(r) = \theta_0$ even when the integration of (3.10) begins at $r_{\text{init}} > 0$, with arbitrary initial value, while for $\kappa > 0$ we get $\lim_{r \to 0} \theta(r) = \theta_0$ only if we start the integration at $r_{\text{init}} = 0$ with the initial value θ_0, and integrate outward (or integrate along the principal line, which is very unlikely when the integration is carried out numerically).

2. Solution in the tail region

When $r \geqslant r_t$ the screening is constant and Eq. (3.10) can be solved analytically.

(i) Ionic potential, $s_t > 0$. Let $\theta(r_t)$ be the initial value at $r = r_t$. Then

$$\theta(r) = \theta_t + (2\gamma_t/a_t) \cdot [cr^{2\gamma_t}/(1 - cr^{2\gamma_t})] \quad (r \geqslant r_t), \tag{A4}$$

where the constant c is given by

$$c = (1/r_t^{2\gamma_t}) \cdot ([\theta(r_t) - \theta_t]/[\theta(r_t) - 1/\theta_t]). \tag{A5}$$

(ii) Neutral atom potential, $s_t = 0$. Here we have $\theta_t = 0$ and

$$\theta(r) = \theta(r_t)(r_t/r)^{2k} \quad (r \geqslant r_t). \tag{A6}$$

For $k < 0$ we choose $\theta(r_t) = \theta_t$ and get

$$\theta(r) = \theta_t = 0 \quad \text{for } r \geqslant r_t. \tag{A7}$$

For $\kappa > 0$ $\theta(r_t)$ is determined by integrating from the origin outward, and we get

$$\lim_{r \to \infty} \theta(r) = \theta_t = 0. \tag{A8}$$

3. Solution for $\kappa < 0$, $r \leqslant r_t$

We will show that in this case $\theta(r)$ is a monotonically increasing function. We start the integration of Eq. (3.10) at $r_{\text{init}} = r_t$ with the initial value $\theta(r_t) = \theta_t$. Let n be the order of the first nonvanishing, left side, derivative of the screening function $s(r)$ at $r = r_t$. Then by induction the first nonvanishing, left side, derivative of $\theta(r)$ at $r = r_t$ is of the order $n + 1$, and we have

$$r_t \left(\frac{d^{n+1}\theta}{dr^{n+1}}\right)_{r_t} = a_0(\theta_t^2 + 1)\left(\frac{d^n s}{dr^n}\right)_{r_t}. \tag{A9}$$

Since $s(r)$ is assumed to be a monotonically decreasing function we get

$$\left(\frac{d^{n+1}\theta}{dr^{n+1}}\right)_{r_t} \begin{cases} > 0, & n \text{ even}, \\ < 0, & n \text{ odd}, \end{cases} \tag{A10}$$

and therefore, in the vicinity of r_t ($r \lesssim r_t$) $\theta(r)$ is an increasing function of r and $d\theta/dr > 0$ near r_t.

Suppose that while integrating Eq. (3.10) inward we encounter a point $\bar{r}$ for which $(d\theta/dr)_{\bar{r}} = 0$. Then, at this point we have [$s(r)$ is a monotonically decreasing function]

$$\left(\frac{d^2\theta}{dr^2}\right)_{r_t} = a_0(\theta^2(\bar{r}) + 1)\left(\frac{ds}{dr}\right)_{\bar{r}} < 0. \tag{A11}$$

Thus $\theta(r)$ has a maximum at $r = \bar{r}$, which is impossible since $d\theta/dr > 0$ for $\bar{r} < r < r_t$. Therefore $d\theta/dr > 0$ for the whole range $(0, r_t)$.

4. Solution for $\kappa > 0$

We will show that $\theta(r)$ is a monotonically decreasing function for $r > 0$. Let s_n $(n \geqslant 1)$ be the first nonvanishing coefficient in (A1). Then, the first nonvanishing coefficient (for $n \geqslant 1$) in the expansion of $\theta(r)$ is

$$\theta_n = 2\theta_0 \kappa S_n/(n - 2\gamma_0) < 0, \tag{A12}$$

and therefore $d\theta/dr < 0$ for $r \geqslant 0$.

Suppose that while integrating Eq. (3.10) outward we encounter a point $\bar{r}$ for which $(d\theta/dr)_{\bar{r}} = 0$. Then $\theta(r)$ has a maximum at this point, which is, however, impossible since $d\theta/dr < 0$ for $r < \bar{r}$. Therefore $d\theta/dr < 0$ for the whole range $r > 0$.

[1]P. G. Burke and W. D. Robb, Adv. At. Mol. Phys. **11**, 153 (1975).

[2]F. Herman and S. Skillman, *Atomic Structure Calculations* (Prentice–Hall, Englewood Cliffs, NJ, 1963).

[3]D. Liberman, J. T. Waber, and D. T. Cromer, Phys. Rev. **137**, A27 (1965).

[4]Nonrelativistic: A Zangwill and Paul Soven, Phys. Rev. A **21**, 1561 (1980); relativistic: F. A. Parpia and W. R. Johnson, J. Phys. B **17**, 531 (1984).

[5]Nonrelativistic: M. Ya Amusia and N. A. Cerepkov, Case Stud. At. Phys. **5**, 47 (1975); relativistic: W. R. Johnson and C. D. Lin, Phys. Rev. A **20**, 964 (1979).

[6]Nonrelativistic: A. Dalgarno and G. A. Victor, Proc. R. Soc. London Ser. A **291**, 291 (1966); relativistic: I. P. Grant, Adv. Phys. **19**, 747 (1970).

[7]See, e.g., H. A. Bethe and E. E. Salpeter, *Quantum Mechanic of One and Two Electron Atoms* (Academic, New York, 1957); M. E. Rose, *Relativistic Electron Theory* (Wiley, New York, 1961).

[8]For the point Coulomb potential this is essentially the transformation studied by L. C. Biedenharn, Phys. Rev. **126**, 845 (1962), which also has recently been discussed by J. Y. Su, Phys. Rev. A **32**, 3251 (1985). In this paper we are concerned with the possibility of generalizing the transformation to the screened potential case.

[9]I. B. Goldberg, J. Stein, Akiva Ron, and R. H. Pratt, Fourteenth International Conference on the Physics of Electronic and Atomic Collisions, Abstracts of Contributed Papers, Palo Alto, CA, 1985, p. 135.

[10]This choice of the sign of γ differs from the choice $\text{sign}(\gamma) = \text{sign}(\kappa)$ made in Ref. 8.

[11]W. R. Johnson and K. T. Cheng, J. Phys. B **12**, 863 (1979).

[12]V. A. Zilitis, Opt. Spectrosc. (USSR) **50**, 227 (1981).

[13]M. J. Seaton, Mon. Not. R. Astron. Soc. **118**, 504 (1958).

[14]Reference 13, Eqs. (8) and (9).

[15]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, Washington, 1964), Eq. (13.1.2).

[16]Bateman Manuscript Project, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. 1, p. 278.

[17]Reference 16, p. 48, Eq. (12), and p. 37, Eq. (12).

[18]F. S. Ham, *Solid State Physics*, edited by F. Seitz and D. Turnbull (Academic, New York, 1955), Vol. 1, p. 127.

[19]E. L. Ince, *Integration of Ordinary Differential Equations* (Interscience, New York, 1952), pp. 35–39.

A geometric approach to quantum scattering with group symmetry

Yihren Wu
Department of Mathematics, Hofstra University, Hempstead, New York 11550

(Received 16 October 1986; accepted for publication 11 February 1987)

A geometric theory for quantum scattering when the symmetry group is semisimple is presented. This theory is seen as a generalization of the partial wave analysis. As an application of this theory, the S-matrix elements for scattering in the Pöschl–Teller potential with symmetry group SO(1,2), Coulomb potential with SO(1,3), and a perturbed Coulomb potential with SO(2,3) are calculated. The last example may be considered as a model for heavy-ion scattering.

I. INTRODUCTION

In a series of papers,[1] Alhassid *et al.* have studied the quantum scattering problems having an SU(1,1) symmetry group. Some examples are scattering with the Morse potential and Pöschl–Teller potential. Their work involves a procedure called the "Euclidean connection" in which they show that the shift operators acting on the set of asymptotic states can be written in terms of operators from the Euclidean algebra, thus giving rise to recursive relations for the transmission and reflection coefficients in the adjacent states, and the S-matrix elements are determined. This procedure is completely group theoretical, in a sense that no physical aspect of the problem enters the calculation, so the result will hold [modulo the relation between the Casimir operator of SU(1,1) and the energy Hamiltonian] for all situations having such a group symmetry. This procedure is extended to study the SO(2,n) groups.

Recently, we have shown[2] that the scattering problem is closely related to the Radon transform if the symmetry group is semisimple. Based on the assumption that the S-matrix elements should look the same, we consider the "standard" scattering problem, namely the geodesic flow problem on the symmetric space associated with the group. We quantize (via the techiques of geometric quantization[3]) this classical system and show that the wave operators are the dual Radon transform on the symmetric space corresponding to different choices of the Weyl chambers. If the group is of rank 1, there are only two Weyl chambers, we can label the resulting wave operators as incoming and outgoing, and the scattering operator can be calculated accordingly. The calculation is the same in the higher ranking cases, their physical significance still remains to be explored.

The dual Radon transform (wave operator) operates from the set of functions on the dual space (free states) to functions on the symmetric space (interacting states). This seems to suggest that the free states are expressed more conveniently as functions on the dual space, rather than the Euclidean space. This observation leads us to consider an alternative to the Euclidean connection procedure. Our objective here is to show that the S-matrix elements can be found by comparing the asymptotic states on the symmetric space and the free states on the dual space, a method similar to the one employed by Frank and Wolf.[4] Moreover, the comparison has been performed for the spherically symmetric functions [the states $|k,0\rangle$ in the SU(1,1) case or, in the Coulomb case, the states that have only radial dependence]. The results are the celebrated Harish-Chandra c-functions, and have been computed for all semisimple Lie groups by Gindikin and Karpelevic.[5]

This paper is organized as follows. In Sec. II we present the formal setup for this comparison, and we show that this is a generalization of the partial wave analysis. If we adhere to this formalism, there is a constraint on the dimension of the phase space relative to the dimension of the various groups and subgroups in question. In the cases when the dimensional constraints are satisfied, the calculation is straightforward. Examples for the nice dimensions are the Pöschl–Teller potential and the Coulomb potential with symmetry group SO(1,3). These will be done in Sec. III. In the other cases, we need to know more about the geometry of the symmetry group so as to choose the state spaces with the right dimensions. In Sec. IV we give an example for the group SO(2,3) as the symmetry group for a perturbed Coulomb potential. The procedure is analogous, we hope to report on the formal principle in the future.

II. PARTIAL WAVE ANALYSIS AND RADON TRANSFORM

The relevant details on Lie group theory can be found in Helgason.[6]

Let G be a semisimple Lie group of noncompact type with finite center, K a maximal compact subgroup. Fix a Cartan decomposition $\mathbf{p}+\mathbf{k}$ of the Lie algebra $\mathbf{g}$, where the analytic subgroup of $\mathbf{k}$ is K. Let $\mathbf{a}$ be the maximal Abelian subalgebra of $\mathbf{p}$, $\mathbf{a}^+$ the Weyl chamber. Let $A=\exp\mathbf{A}$, $A^+=\exp\mathbf{a}^+$ be the analytic subgroups of $\mathbf{a}$ and $\mathbf{a}^+$ in G, respectively. Let M,M' denote the centralizer and normalizer of A in K, then $B=K/M$ can be viewed as the boundary of the symmetric space $X=G/K$. Let $G=KAN$ be the Iwasawa decomposition of G corresponding to our previous choices. Define a function H: $X\times B\to\mathbf{a}$, $H(x,b)=\alpha$ if $X=b\exp\alpha nK$ for some $n\in N$. Let σ be one-half the sum of positive roots, $\mu\in\mathbf{a}^*$, the dual vector space, then $\exp\langle i\mu-\sigma,H(x,b)\rangle$ are the plane waves. We get the plane wave decomposition of functions on X,

$$\tilde{F}(b,\mu)=\int_X f(x)\exp\langle i\mu-\sigma,H(x,b)\rangle dx. \tag{1}$$

 0022-2488/87/061360-04$02.50

Here $\tilde{F}(-,\mu)$, which is considered as a function on B, can be decomposed further once we have a representation of the compact group K. This allows us to label the states as $(\mu_1,\mu_2,\ldots, m_1,\ldots,m_2,\ldots)$, where m_i's are discrete corresponding to the state labeling of K, and μ_1's are continuous as components of $\mu\in\mathbf{a}^*$.

Closely related to the plane wave decomposition of functions on X, we have the Radon transform, first discussed by Newton[7] in the Euclidean case and was shown to have the same significance in all symmetric spaces of the form G/K (see Ref. 2). There the integration is over horocycles $C(b,\alpha)=\{x\in X\,|H(x,b)=\alpha\}$ or

$$F(b,\alpha)=\int_x f(x)\delta(H(x,b)-\alpha)dx. \tag{2}$$

The set of horocycles are parameterized by $\Sigma=G/MN$, where $C(b,\alpha)$ is given as $b\exp\alpha MN\in\Sigma$, where Σ is known as the dual space of X. We can rewrite (2) as

$$F(b\exp\alpha MN)=\int f(b\exp\alpha nK)dn. \tag{3}$$

This is the Radon transform.

Physically, b is the outgoing asymptotic "direction," α is the dual coordinate to the "energy" μ. So Σ is the space of asymptotic data, and (3) can be interpreted as the inverse of the scattering (outgoing) wave operator. In the incoming direction, the calculation is the same as above except we choose the opposite Weyl chamber $\mathbf{a}^-=-\mathbf{a}^+$.

Let $L^2(\Sigma)$ and $L^2(X)$ be carrier spaces of representations of G. Denote $|\pm\mu_i,m_i\rangle_\Sigma$ and $|\mu_i,m_i\rangle_X$ the states in the two spaces with the given label. We can consider them as the free and interacting states of given asymptotic data, since $|\pm\mu_i,m_i\rangle_\Sigma$ are the Radon transforms of $|\mu_i,m_i\rangle_X$ with respect to $\mathbf{a}^+$ and $\mathbf{a}^-$. The transmission and reflection coefficients are given by comparing these states. This comparison is well defined via the polar coordinates on X and Σ:

$$X=K/M\times A^+,\quad \Sigma=K/M\times A. \tag{4}$$

We now have asymptotically

$$|\mu,m\rangle_X\approx A\,|\mu,m\rangle_\Sigma+B\,|-\mu,m\rangle_\Sigma. \tag{5}$$

For the K invariant states, $A=c(\mu)$ is the Harish-Chandra c function. In this sense, the transmission and reflection coefficients are extensions to the c functions, corresponding to choosing $\mathbf{a}^+$ and $\mathbf{a}^-$, respectively. Furthermore, this comparison is just a generalization of the partial wave analysis. In the standard case, G is of rank 1 (rank $G=\dim\mathbf{a}$), μ is interpreted as the energy, $K\approx SO(3)$, $M\approx SO(2)$, so $B\approx S^2$. The explicit formulas calculated by Gindikin and Karpelevic will suggest a functional form of the matrix elements.

If the symmetry group is of rank 1, then there is an essentially unique Casimir operator. This operator plays the role of the quantum Hamiltonian for the dynamical system. In the case where the group is of higher rank, we have more than one Casimir operator, and the correspondence between the quantum observables and the Lie algebra representation is unclear. Different correspondence will result in completely different dynamical systems. For instance, we know that the group $SU(1,1)$ is the symmetry for a particle in the Pöschl–Teller potential, so if we have two noninteracting particles moving in this potential, the symmetry group for elastic scattering is $SU(1,1)\times SU(1,1)\approx SO(2,2)$. This is of course different from the case considered by Alhassid *et al.* in which one particle is moving in a two-dimensional space, with a modified Pöschl–Teller potential.

If the space G/K is appropriate, then the procedure as described will give correct S-matrix elements. Examples include the Pöschl–Teller potential with $SU(1,1)\approx SO(1,2)$, Coulomb potential with $SO(1,3)$, and the two noninteracting particles with $SO(2,2)$. We will also consider the group $SO(2,3)$ as the symmetry group of heavy-ion scattering, here the isotropy group is not the maximal compact subgroup. The procedure is analogous, it turns out that X is the coset space $SO(2,3)/SO(1,3)$. We note here that Radon transform has been generalized to these situations (known as double fibrations in Helgason[6]), although not as complete.

III. APPLICATIONS TO SCATTERING

A. Pöschl–Teller potential

Here we assume that the dynamical symmetry decomposes according to the chain $SO(1,2)\supset SO(2)$. The symmetric space $X=SO(1,2)/SO(2)$ can be realized as a hyperboloid. Let us take the upper sheet of the two-sheeted hyperboloid

$$X=\{x\in\mathbb{R}^3\,|\,[x,x]=1,\ x_0\geqslant 0\}, \tag{6}$$

where $x=(x_0,x_1,x_2)$, $[\ ,\ ]$ denote the pseudometric $(+\ -\ -)$. The dual space is the cone

$$\Sigma=\{y\in R^3\,|\,[y,y]=0, y_0\geqslant 0\}. \tag{7}$$

Since $SO(1,2)$ acts on X and Σ, we define interacting states and free states as eigenfuctions to the Casimir operator of $SO(1,2)$ on X and Σ, respectively. As indicated earlier, we will perform a comparison between the states on X and Σ corresponding to the same eigenvalues for the chain $SO(1,2)\supset SO(2)$. Denote $|\mu,m\rangle_X$ and $|\mu,m\rangle_\Sigma$ as the states on X and Σ, respectively, we want to determine the transmission and reflection coefficients A and B such that

$$|\mu,m\rangle_X\approx A\,|+\mu,m\rangle_\Sigma+B\,|-\mu,m\rangle_\Sigma. \tag{8}$$

Choosing coordinate $x_0=\cosh\alpha$, $x_1=\sinh\alpha\cos\theta$, $x_2=\sinh\alpha\sinh\theta$, the Casimir operator takes the form

$$C^2{}_x=-\frac{1}{\sinh\alpha}\frac{\partial}{\partial\alpha}\sinh\alpha\frac{\partial}{\partial\alpha}-\frac{1}{\sinh^2\alpha}\frac{\partial^2}{\partial\theta^2}, \tag{9}$$

and

$$\begin{aligned}&|\mu,m\rangle_X=u^\mu{}_m(\alpha)e^{im\theta},\\ &u^\mu{}_m(\alpha)=P^m{}_{i\mu-1/2}(\cosh\alpha),\\ &C^2{}_x|\mu,m\rangle_X=(-\mu^2-1/4)|\mu,m\rangle_X,\end{aligned} \tag{10}$$

where P denotes the conical functions.[8] Similarly, choose coordinate on Σ as $y_0=e^\alpha$, $y_1=e^\alpha\cos\theta$, $y_2=e^\alpha\sin\theta$, then the Casimir operator is

$$C^2{}_\Sigma\quad\frac{d^2}{d\alpha^2}+\frac{d}{d\alpha}. \tag{11}$$

The eigenfunction are of the form $e^\sigma f(\theta)$ with eigenvalue $\sigma^2+\sigma$. If $\sigma=\pm i\mu-\frac{1}{2}$, $\sigma^2+\sigma=\mu^2-\frac{1}{4}$, we set

$$|\pm\mu,m\rangle_\Sigma = e^{(\pm i\mu - 1/2)\alpha} e^{im\theta}. \tag{12}$$

The coordinates are chosen so that the comparison is performed under the limit as α approaches infinity. One need only to get the asymptotic expansion of $P^m{}_{i\mu-1/2}(\cosh\alpha)$. We have[8]

$$(2\pi\ \sinh\alpha)^{1/2} P^m{}_{i\mu-1/2}(\cosh\alpha) \approx e^{i\mu\alpha}\frac{\Gamma(i\mu)}{\Gamma(\frac{1}{2}+m+i\mu)} + e^{-i\mu\alpha}\frac{\Gamma(-i\mu)}{\Gamma(\frac{1}{2}+m-i\mu)}. \tag{13}$$

The S-matrix elements are

$$\frac{\Gamma(\frac{1}{2}+m-i\mu)}{\Gamma(\frac{1}{2}+m+i\mu)} \cdot \frac{\Gamma(i\mu)}{\Gamma(-i\mu)}. \tag{14}$$

B. Coulomb potential

Let $X = \{x\in R^4 | [x,x] = 1, x_0 \geqslant 1\}$ be the upper sheet of the unit hyperboloid on which the symmetry group of the Coulomb potential SO(1,3) acts naturally, [,] is the metric $+ - - -$, $X = \mathrm{SO}(1,3)/\mathrm{SO}(3)$. Choose coordinates

$$x_0 = \cosh\alpha, \quad (x_1,x_2,x_3) = \sinh\alpha\,\Omega, \quad \Omega\in S^2 \text{ in } R^3. \tag{15}$$

Via a similarity transform $1/\sinh\alpha$, the Casimir operator takes the form

$$-\frac{\partial^2}{\partial\alpha^2} + 1 - \frac{L}{\sinh^2\alpha} \tag{16}$$

(where L is the Laplace operator on S^2), the eigenvalues are μ^2+1. So the eigenfunctions satisfy the equation

$$\left(-\frac{\partial^2}{\partial\alpha^2} + \frac{l(l+1)}{\sinh^2\alpha}\right)\phi = \mu^2\cdot\phi, \tag{17}$$

where $\phi(\alpha,\Omega) = \sinh\alpha\cdot P^{-l-1/2}{}_{i\mu-1/2}(\cosh\alpha)\cdot Y_{lm}(\Omega)$, the P are the conical function and the Y are the spherical harmonics.[9] The work of Ref. 9 can now be reinterpreted as a partial wave analysis described above. Notice that in both these cases, via a similarity transform, the Casimir operator on Σ becomes simply $\partial^2/\partial\alpha^2$. We suspect that this is true, in general, thus we always compare the interacting states with $e^{i\mu\alpha}$.

IV. HEAVY-ION SCATTERING

It was reported[1] that heavy-ion scattering possesses an SO(2,3) symmetry. Here we will show that this problem can be recognized as a short range perturbation on the Coulomb interaction, and that the group SO(2,3) enters naturally. The comparison procedure described above is applied and the results are consistent with those obtained by Alhassid *et al.*

Let $X = \{x\in R^5\}|[x,x] = 1\}$, [,] is the metric $+ + - - -$, $X = \mathrm{SO}(2,3)/\mathrm{SO}(1,3)$. We parametrize X as

$$(x_1,x_2) = \cosh\alpha\cdot w, \quad (x_3,x_4,x_5) = \sinh\alpha\cdot\Omega, \tag{18}$$

with $w\in S^1$ in R^2, $\Omega\in S^2$ as before. Via the similarity transform $1/\sinh\alpha\cdot\cosh^{1/2}\alpha$, using the fact that the eigenvalue is $\mu^2+\frac{9}{4}$, the Laplace equation on X reads

$$C_{l,k\Phi} = \left(-\frac{\partial^2}{\partial\alpha^2} - \frac{l(l+1)}{\sinh^2\alpha} - \frac{k^2-\frac{1}{4}}{\cosh^2\alpha}\right)\Phi = \mu^2\cdot\Phi, \tag{19}$$

where $-k^2$ is the eigenvalue of the operator $\partial^2/\partial w^2$. One sees that this equation is a perturbation of (17), the last term corresponds to a potential of the form e^{-r}/r^2 (see the Appendix). Notice that if $k=\frac{1}{2}$, Eq. (19) becomes that of the Coulomb problem. Our procedure suggests that we should compare the eigenfunctions with eigenfunctions on the asymptotic cone $\Sigma = \{y\in R^5 | [y,y] = 0\}$.

From Ref. 1, the S-matrix elements are

$$S(\mu,l,k) = \frac{\Gamma(\frac{1}{2}(l+k+\frac{3}{2}+i\mu))\Gamma(\frac{1}{2}(l-k+\frac{3}{2}+i\mu))}{\Gamma(\frac{1}{2}(l+k+\frac{3}{2}-i\mu))\Gamma(\frac{1}{2}(l-k+\frac{3}{2}-i\mu))}. \tag{20}$$

Note than when $k=\frac{1}{2}$, this reduces to the Coulomb S matrix. So our comparison trick will give the same result in these cases. Equation (20) yields the following recursive relation:

$$S(\mu,l-1,k-1) = [(l+k-\tfrac{1}{2}+i\mu)/(l+k-\tfrac{1}{2}-i\mu)]S(\mu,l,k). \tag{21}$$

So we need to show that our comparison gives the same relations.

Let

$$J = \frac{\partial}{\partial\alpha} + l\cdot\coth\alpha + (k-\tfrac{1}{2})\cdot\tanh\alpha, \tag{22}$$

then we have, C as in (19),

$$C_{(l-1),(k-1)} J\,|\mu,l,k\rangle_X = JC_{l,k}\,|\mu,l,k\rangle_X. \tag{23}$$

Thus J plays the role of a shift operator, whether J belongs to the SO(2,3) algebra is immaterial. So asymptotically,

$$\begin{aligned}\text{const}\cdot|\mu,l-1,k-1\rangle &= J\,|\mu,l,k,\rangle \approx \left(\frac{\partial}{\partial\alpha} + l + k - \tfrac{1}{2}\right)\cdot(A\cdot e^{i\mu\alpha} + B\cdot e^{-i\mu\alpha}) \\ &= (i\mu + l + k - \tfrac{1}{2})A\cdot e^{i\mu\alpha} \\ &\quad + (-i\mu + l + k - \tfrac{1}{2})B\cdot e^{-i\mu\alpha},\end{aligned} \tag{24}$$

and we do get the same recursive relations.

In conclusion, we have embarked on a geometric theory, which complements the algebraic theory of Alhassid *et al.*, for scattering in the presence of a group symmetry. We see that this theory is an extension to the partial wave analysis, and that the geometric consideration simplifies the calculation.

APPENDIX: CANONICAL TRANSFORM FOR POSITIVE COULOMB SPACE AND THE HYPERBOLOID

Here we give explicitly the SO(1,3) equivariant canonical transform between the positive energy phase space of the Coulomb problem and the phase space of the geodesic flow problem on the hyperboloid. This allows us to estimate the size of the perturbation term in (19). The construction here is similar to the one given in Souriau and Onofri[10] for the negative energy case.

Let $T^*{}_+R^3$ denote the subspace of the cotangent bundle on R^3 on which the energy $E = p^2/2 - 1/q$ is positive. Denote T^*H the cotangent bundle on the hyperboloid

$H = \{y \in R^4 | [y,y] = 1, y_0 \geqslant 1\}$. For convenience, we can (via a reduction[11]) identify T^*H as the subspace $\{(y,\eta) \in R^4 \times R^4 | [y,y] = 1, \ [y,\eta] = 0\}$. The symplectic form is $d\eta_0 \wedge dy_0 - \Sigma d\eta_i \wedge dy_i$ so that the SO(1,3) actions are canonical. Then the canonical transform $T^*H \to T^*{}_+R^3$ is

$$y = A \cosh t + (B/\sqrt{-[B,B]}) \sinh t,$$
$$\eta = \sqrt{-[B,B]} \cdot A \cdot \sinh t + B \cosh t, \tag{A1}$$

where

$$t = \langle q,p \rangle \sqrt{2E},$$
$$A = \begin{bmatrix} p^2 q - 1 \\ \sqrt{2E} \cdot q \cdot p_i \end{bmatrix}, \quad B = \begin{bmatrix} \langle q,p \rangle \\ (\langle q,p \rangle p_i - q_i/q)/\sqrt{2E} \end{bmatrix}. \tag{A2}$$

The perturbation term $(k^2 - \frac{1}{4})/y_0{}^2$, both near and far away from the scattering center, assumes the form

$$(k^2 - \tfrac{1}{4}) \exp(-4Eq)/(2Eq)^2. \tag{A3}$$

Thus we see that this term is a short range potential which grows as $1/q^2$ at the origin.

[1] Y. Alhassid, F. Iachello, and J. Wu, Phys. Rev. Lett. **56**, 271 (1986); Y. Alhassid, F. Gürsey, and F. Iachello, *ibid.* **50**, 873 (1983); Ann. Phys. (NY) **148**, 346 (1983); Yale preprint No. 3074-837.

[2] Y. Wu, J. Phys. A **20**, 429 (1987).

[3] B. Kostant, *Quantization and Unitary Representation*, in *Lecture Notes in Mathematics*, Vol. 170 (Springer, New York, 1970); J. Sniatycki, *Geometric Quantization and Quantum Mechanics* (Springer, New York, 1980).

[4] A. Frank and K. B. Wolf, Phys. Rev. Lett. **52**, 1737 (1984).

[5] S. G. Gindikin and F. I. Karpelevic, Dok. Akad. Nauk SSSR **145**, 252 (1962).

[6] S. Helgason, *Groups and Geometric Analysis* (Academic, New York, 1984).

[7] R. G. Newton, *Scattering Theory in the Mixed Representation, Lecture Notes in Physics*, Vol. 130 (Springer, New York, 1980).

[8] W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics* (Springer, New York, 1966).

[9] Y. Wu, J. Phys. A **18**, L499 (1985).

[10] J. M. Souiau, *Sur la Variete de Kepler, Symposia Math* (Academic, London, 1974); E. Onofri, J. Math. Phys. **17**, 401 (1976); Y. Wu, Hadronic J. **8**, 101 (1985).

[11] V. W. Guillemin and S. Sternberg, Am. J. Math. **101**, 915 (1979).

Existence and observability of spinor structure

Anne M. R. Magnon[a)]
Département de Mathématiques, Université de Clermont-Ferrand, 63170, Aubière, France; Physics Department, University of Syracuse, Syracuse, New York 13244-1130; and Institute for Theoretical Physics, University of California, Santa Barbara, California 93106

(Received 3 March 1986; accepted for publication 7 January 1987)

A mechanism by which space-time topological modifications could have been controlled, in the early universe or at the Planck length, to enable onset of spinor structure is investigated. This mechanism (based on a reshuffling of topological charges and related modification of characteristic classes) could provide a gravitational analog of the Aharonov–Susskind *Gedankenexperiment* proposed to detect relative rotation in the universe, spinor behavior, or to keep track of the two homotopy classes of the Lorentz Lie group. The space-time topology [and in particular the trivial (nontrivial) bundle structure at conformal null infinity] provide a labeling of the asymptotic Lorentz homotopy classes which originates in the first Chern class (enclosed magnetic mass) or in the parametrization of the second homology group, and gives rise to a necessary (and sufficient) condition for the existence of spinor structure. This underlines the intertwined roles of topology and curvature. The mechanism could also be viewed as an "unwinding" of gravitational magnetic monopoles with one asymptotic region into electric mass (black-hole) solutions with two asymptotic regions. In such situations a discrete PT symmetry could emerge from a continous transformation. Possible implications on the CPT theorem are mentioned.

I. INTRODUCTION

Since the discovery, in 1956, of parity nonconservation in weak interactions, it is believed that the concepts of right and left physical systems can be unambiguously defined provided space is orientable. Indeed such a statement could not be made if a closed circuit within a disoriented laboratory could transform a particle into an antiparticle: in this case the combined transformations, charge conjugation (C) and space inversion (P), would not enable us to define the conservation of right–left symmetry in space. Equivalently, in a disoriented three-space, space inversion is not a discrete transformation but a continuous one[1] and CP invariance does not enable us to decide whether two remote particles are identical or are a particle and an antiparticle since the conclusion would depend on the path along which the two particles are brought to the same point. However, the absolute difference between matter and antimatter has been experimentally confirmed by the CP violation in the K^0-meson decay, and nonorientability of space seems to be excluded by the same token.[2,3] It is to be underlined here that such conclusions are drawn under the requirement that causality violation is excluded, i.e., that space-time does not contain closed timelike world lines.

Since one might question the availability of causality, e.g., in the early universe or on a cosmic scale (and we shall do so here), it is of interest to search for mechanisms that could explain the onset of orientability or even bring support to a further question: Why should there be orientability in the universe or PCT invariance?

It has often been underlined that this issue is related to that of observability of a relative 2π rotation of two systems (certainly a 2π rotation of the entire universe should not be observable), and to the existence of spin structure.[4–7]

Underlining a key point—a spinor changes sign when a basis completes a 2π rotation, returning to its original position after 4π rotation—Penrose has suggested[5] a mechanism (based on a *Gedankenexperiment* proposed by Aharonov and Susskind[8]) to test spin structure. The Aharonov–Susskind apparatus is a box with perfectly reflecting walls which is divided into two identical compartments (by an impenetrable partition which may be open or closed off by a shutter) and which contains an electron. A relative (quasistatic) 2π rotation between the two compartments can be detected if the electron wave function in box 1 is allowed to produce an interference pattern with the wave function in box 2 after uniform equal magnetic fields have been applied to the disconnected compartments in the direction of the spin of the electron. The relation with spinor structure emerges from the following argument. The physically measurable properties of a spinor ψ^A can be determined from its corresponding null bivector (null flag). If ψ^A undergoes a (phase) rotation $\psi^A \to e^{i\theta}\psi^A$, it is clear that ψ^A changes sign; it takes a 2π rotation to bring it back to its initial value, in this rotation the null flag undergoes a 4π rotation around its attached null direction $k^a = \psi^A \psi^{A'}$. This is related to the existence of two homotopy classes in the Lorentz Lie group $\mathcal{L}$—every closed path in this group being either homotopic to the 2π rotation (element of the nontrivial homotopy class) or the 4π rotation (the trivial class)—and to the fact that the SL(2,C) Lie group is the universal covering group of $\mathcal{L}$, two elements L_B^A and M_B^A of SL(2C) giving rise to the same Lorentz group element if and only if $L_B^A = \pm M_B^A$.

[a)] Détachée du Ministère des Relations Extérieures, Paris, France.

 0022-2488/87/061364-06$02.50

Since the definition of a spinor bundle $\mathscr{S}$ over a curved space-time M involves the intertwining of the fundamental group of M with that of the Lorentz group, the role of topology is crucial and has been investigated by various authors.[6,7,9] The basic idea is to start from the principal fiber bundle $\mathscr{P}$ of oriented-time oriented orthonormal bases, to unwrap each fiber into a principal $SL(2,C)$ bundle B over M, a spinor structure being defined as a second principal fiber bundle $\mathscr{S}$ over B with a 2–1 mapping ϕ: $\mathscr{S} \to B$. If $\mathscr{P}$ is simply connected, closed curves contained within a fiber and corresponding to the nontrivial 2π-rotation homotopy class can be deformed outside the fiber into the trivial curve and the notion of a spinor field cannot be defined. It is known that $\mathscr{P}$ will fail to be simply connected if the second Stiefel–Whitney class of M vanishes. If M is noncompact, it is also known[6] that a spinor structure exists if and only if M is parallelizable.

In presence of intricated space-time topologies, M may fail to be orientable, or time oriented. In such a case $\mathscr{P}$ does not exist and the usual notion of spinors cannot be defined. If M is orientable and time oriented, $\mathscr{P}$ can be unwrapped into B without unwrapping M provided $\pi_1(\mathscr{P}) = \pi_1(M) \times Z_2$; if this condition is not satisfied, the notion of spinor structure cannot be defined. It is also of interest to underline that the above features of the $SL(2,C)$ representation are related to the presence at the quantum level of two irreducible representations[10] which are conveniently labeled by the values of the Casimir operators m^2 (squared mass) and S^2 (squared angular momentum) of the $SL(2,C)$ Lie algebra. The results presented here might suggest that a canonical generalization in the context of quantum gravity could be provided by the squared mass and the squared angular momentum monopole.

These considerations are intended to motivate the viewpoint we would like to develop, i.e., the existence of spinor structure and orientability could have been controlled in the early universe or at the Planck length through an appropriate reshuffling of topological charges and related modification in the homology classes of the space-time manifold.

We shall take advantage of the availability of various Maxwellian features of gravity in presence of suitable nontrivial space-time topologies to propose a mechanism rather reminiscent of that designed by Aharonov and Susskind for the detection of spinor structure in presence of a Maxwellian magnetic field. Our analysis will be based on previous results[11–13] concerning the structure of magnetic mass source-free solutions to Einstein's equation.

A simplified picture of the gravitational apparatus to be investigated could be the following: the two compartments of the Aharonov–Susskind boxes are provided by the two asymptotic regions (or mirror universes) of the Kruskal–Schwarszchild black-hole solution. A reshuffling of the (electric) mass monopole into the magnetic mass monopole (which could be supported by the following geometrical interpretation: decreasing the area of the event horizon, shrinking the region of trapped surfaces, and smooth absorption into a causality violation and single asymptotic region) has been investigated in Refs. 11–13. As a result of the transformation, a nontrivial bundle structure (S^1 Hopf fibering and transition functions) is induced at conformal null infinity $\mathscr{I}$. One could thus draw an analogy between the above setup and the Aharonov–Susskind apparatus, where the opening of the shutter would correspond to the absorption of the "entropy reservoir" into new topological features, and the interference pattern to nontrivial bundle structure at $\mathscr{I}$, S^1 Hopf fibering and transition functions measuring the enclosed magnetic charge.

In the main body of this paper we shall prove that the parametrization of characteristic classes involved in the definition of magnetic mass (second homology class and first Chern class) induces a (0–1) labeling of the homotopy classes of the (asymptotic) Lorentz Lie group as well as a mechanism to keep track of their possible mingling (as one moves from fiber to fiber within the principal frame bundle) along closed space-time paths, thus providing a necessary (and in many cases sufficient) condition for the existence of spinor structure. This is strongly reminiscent of a setup proposed by Geroch[6] to keep track of spinor structure in the universe by moving Maxwellian–Aharonov–Susskind boxes along space-time tracks. However, the control here is purely gravitational, global, and uses topological charges (characteristic classes) which originate in the bundle structure at null infinity.

Since the above criterion is purely global and since we rely on the space-time asymptotic structure, we would like to propose the following viewpoint. If gravitational magnetic monopoles could have "unwinded" into black holes (electric mass monopoles) in the early universe, a mechanism to control onset of spinor structure would be provided.

Finally the above transformation can be shown[12] to be associated to an invariant (gravitational) charge C, and to the onset of a PT discrete symmetry (between two-mirror-asymptotic regions) emerging from a continuous transformation (acting on the—single—asymptotic region of a gravitational magnetic monopole solution). This leads us to comment, in our concluding remarks, on a possible (theoretical or experimental) link between the CPT theorem and the onset of spinor structure.

II. PRELIMINARY REMARKS

Let us briefly summarize some relevent properties of the spinor representation. Recall that a spinor space is a pair (W,ϵ_{AB}), where W is a two-dimensional vector space over C and $\epsilon_{AB} = -\epsilon_{BA}$, a skew-symmetric two-index tensor inducing a mapping from W into its dual W^*, hence playing the role of a nondegenerate metric. The corresponding group of transformations is that of linear mappings L: $W \to W$ which are metric preserving:

$$\epsilon_{CD} = L^{A}_{C} L^{B}_{D} \epsilon_{AB}, \tag{1}$$

i.e., elements of $SL(2,C)$.

A four-dimensional complex vector space Y can be associated to W, which admits a real section V with Lorentzian metric (signature $+,-,-,-$) given by

$$g_{AA'BB'} = \epsilon_{AB} \bar{\epsilon}_{A'B'}. \tag{2}$$

It can be checked that

$$\Lambda^{AA'}_{BB'} = L^{A}_{B} \bar{L}^{A'}_{B'} \tag{3}$$

is metric preserving (a proper Lorentz transformation):

$$g_{CC'DD'} = \Lambda^{AA'}_{CC'}\Lambda^{BB'}_{DD'}g_{AA'BB'}. \tag{4}$$

Furthermore L^{A}_{B} and M^{A}_{B} give rise to the same Lorentz element iff $L^{A}_{B} = \pm M^{A}_{B}$, reflecting the fact that SL(2,C) is the universal covering group of the Lorentz group (two-valuedness of the spin representation). It is the condition that this two-valuedness can be consistently eliminated over an entire space-time which enables the definition of global spinor fields, reflecting the presence of a spinor structure.

Since the Lorentz Lie group $\mathscr{L}$ is doubly connected—a path representing a 2π rotation cannot be continuously deformed to the identity whereas a 4π rotation can be—a spin structure must be able to keep track of the presence of these two homotopy classes, providing a homomorphism from $\pi_1(\mathscr{L})$ into Z_2, taking the value 0 on the trivial class (containing the 4π rotation), unity on the other class. Following a geometrical argument proposed by Penrose,[5] we suggest that such a homomorphism could be associated to the onset of topological charges or equivalently specific homology classes. Recall that to every one-spinor $\xi^A \in W$ one can associate a bivector $F_{AA'BB'}$ defined via

$$F_{[ab]} \leftrightarrow F_{AA'BB'} = \epsilon_{AB}\bar{\phi}_{A'B'} + \bar{\epsilon}_{A'B'}\phi_{AB}, \tag{5}$$

where $\phi_{AB} = \xi_A\xi_B$. This bivector (or null flag) is attached to a pole represented by the direction of the null vector $\xi^A\bar{\xi}^{A'}$. Let us denote by f_θ the (phase) rotation defined via

$$f_\theta(\xi^A) = e^{i\theta}\xi^A. \tag{6}$$

The induced transformation on F_{ab} is

$$f_\theta(F_{ab}) = F_{ab}\cos 2\theta - {*F_{ab}}\sin 2\theta. \tag{7}$$

On another hand, a basis $(\xi^A,\eta^A,\xi\cdot\eta = 1)$ being chosen in W, F_{ab} can be expressed as

$$F_{ab} = 2\xi_{[a}w_{b]} \quad (\xi\cdot w = 0), \tag{8a}$$

the (phase) rotation inducing a rotation of the two-flat defined by

$$f_\theta w_a = w_a\cos 2\theta + v_a\sin 2\theta, \tag{8b}$$

where (ξ^a,w^a,v^a) is an orthonormal triad in Minkowski space. Hence f_θ induces a reversal of the spinor sign ($f_\pi\xi^A = -f_0\xi^A$), and a winding of the null flag around its pole ξ^a, a 4π rotation of this flag being required to bring back ξ^A to its original value. Thus the flag and its null pole have the availability to keep track of the double valuedness of the spin representation. Let S_2 denote a two-sphere surrounding the point at spatial infinity in Minkowski space, and let us assume that F_{ab} is a (plane-wave) source-free Maxwell field. Denote by

$$C_1 = \int_{S_2} {*F_{ab}}\, dS^{ab} \tag{9}$$

its electric charge, and by

$$C_2 = \int_{S_2} F_{ab}\, dS^{ab} \tag{10}$$

its magnetic charge. It is straightforward to check that

$$f_\theta F_{ab} = F_{ab}\cos 2\theta - {*F_{ab}}\sin 2\theta, \tag{11}$$

$$f_\theta {*F_{ab}} = F_{ab}\sin 2\theta + {*F_{ab}}\cos 2\theta. \tag{12}$$

The quantity

$$C = C_1^2 + C_2^2 \tag{13}$$

is an invariant of the transformation, its vanishing (nonvanishing) being governed by the second homology group of the space-time manifold.

A gravitational analog of the transformation f_θ and related reshuffling of charges is available. Its topological implications have been mentioned in previous papers,[11–13] and will be crucial when we shall derive criteria for the existence and possible observability of spinor structure. The role of spinor structure has been carefully underlined[14] in the investigation of the asymptotic behavior of zero rest-mass fields. Let us briefly summarize the situation.

From now onwards, (M,g_{ab},ξ^a) will denote a solution of Einstein's equation $R_{ab} = 0$, with Killing vector field ξ^a. The norm and twist of ξ^a are, respectively, $-\lambda = \xi^a\xi_a$ and $\omega_a = \epsilon_{abcd}\xi^b\nabla^c\xi^d = \text{grad}\,\omega$. Let $h_{ab} = g_{ab} + \lambda^{-1}\xi_a\xi_b$ be the induced metric on T, the manifold of orbits of ξ^a. Assuming $\lambda \neq 0$, a rescaled metric $\tilde{h}_{ab}$ and complex potential τ can be defined[15] on T:

$$\tilde{h}_{ab} = \lambda h_{ab}, \tag{14}$$

$$\tau = \omega + i\lambda. \tag{15}$$

A transformation (which was initially introduced[15–18] to generate circle families of explicit, exact, source-free solutions to Einstein's equation with one Killing vector field, starting from one of them) is given by

$$G_\theta\tau = (\tau\cos\theta + \sin\theta)/(-\tau\sin\theta + \cos\theta). \tag{16}$$

The expression of the resulting solution $g_{ab}(\theta)$ is given explicitly in Ref. 16. The topological charges that can emerge from the action of G_θ are obtained from three real divergence-free vector fields V^a_i $(i = 1,2,3)$ on T, and their associated curl-free two-forms $\epsilon_{abc}V^c_i \equiv \tilde{F}^i_{ab}$ $(i = 1,2,3)$ on T. These forms admit the following pullbacks on (M,g_{ab}):

$$F^1_{ab} = \nabla_{[a}\lambda^{-1}\xi_{b]}, \tag{17}$$

$$F^2_{ab} = \nabla_{[a}\omega\lambda^{-1}\xi_{b]} - \tfrac{1}{2}\epsilon_{abcd}\nabla^c\xi^d, \tag{18}$$

$$F^3_{ab} = \nabla_{[a}(\lambda^{-1}(\omega^2+\lambda^2)\xi_{b]} - 2\lambda\nabla_a\xi_b - \omega\epsilon_{abcd}\nabla^c\xi^d. \tag{19}$$

Integrating $\tilde{F}^i_{ab}$ on a two-sphere S^∞_2 surrounding the point at spacelike infinity on T leads to various conserved quantities:

$$Q_i = \int_{S^\infty_2} \tilde{F}^i_{ab}\, dS^{ab}, \quad i = 1,2,3. \tag{20}$$

Furthermore, under the action of the circle group, these charges are reshuffled according to

$$G_\theta V^a_1 = V^a_1\cos^2\theta - V^a_2\sin 2\theta + V^a_3\sin^2\theta, \tag{21}$$

$$G_\theta V^a_2 = \tfrac{1}{2}V^a_1\sin 2\theta + V^a_2\cos 2\theta - \tfrac{1}{2}V^a_3\sin 2\theta, \tag{22}$$

$$G_\theta V^a_3 = V^a_1\sin^2\theta + V^a_2\sin 2\theta + V^a_3\cos^2\theta. \tag{23}$$

It is easily checked that $Q = Q_1Q_3 - Q_2^2$ is an invariant of the transformation G_θ, a gravitational analog of the charge C.

The reshuffling of Q_i's under the action of the cyclic group (phase rotation) can be related to a modification in the structure of characteristic classes and consequently decide on the existence of spinor structure. This will be investigated in the following sections.

III. MAGNETIC MASS AND CHARACTERISTIC CLASSES

Let us briefly review a definition of the magnetic mass based on various results[11–13] concerning the asymptotic structure of the NUT (gravitational magnetic monopole) source free solution to Einstein's equation. For the exact NUT solution it has been shown elsewhere[11] that this conserved quantity reduces to the NUT parameter (angular momentum monopole).

An asymptotically NUT gravitational magnetic monopole source-free solution to Einstein's equation will be a space-time with conformal null boundary $\mathscr{I}$ exhibiting the topology of an S^1 (Hopf fibering) bundle over S^∞, the two-sphere of orbits of n^a (the null normal to $\mathscr{I}$). The structure of this bundle has been analyzed in Refs. 11–13. The key point here is the availability (and expression) of the bundle curvature two-form Ω_{ab} defined by the pullback to $\mathscr{I}$ of

$$\hat{F}^1_{ab} = \Omega^{-1} {}^*C^m{}_c{}^n{}_d n^c n^d l_m \epsilon_{nab}, \tag{24}$$

where $\Omega^{-1}C_{abcd}$ denotes the rescaled Weyl tensor, (l^a, n^a, m^a, $\bar{m}^a$) is the usual Newman–Penrose null tetrad in the neighborhood of $\mathscr{I}$ ($l\cdot n = -1$), and $\epsilon_{abc} = \epsilon_{abcd} l^d$, $\epsilon^{abc} = \epsilon^{abcd} n_d$. We know also from Ref. 13 that Ω_{ab} is related to the unphysical Riemann curvature via

$$\Omega_{ab} = D_{[a} S^c_{b]} l_c, \tag{25}$$

where $S_{ab} = R_{ab} - \frac{1}{6} R g_{ab}$.

If a timelike Killing vector field is available on the physical space-time, we also know[13] that Ω_{ab} is (essentially) the pullback to $\mathscr{I}$ of $F^1_{ab} = \nabla_{[a} \lambda^{-1} \xi_{b]}$.

Finally Ω_{ab} is the lift to $\mathscr{I}$ of a closed two-form $\tilde{\Omega}_{ab}$ on S^∞, an element of $H^\infty_2(M)$, second homology group (in the asymptotic region). The fact that this $\tilde{\Omega}_{ab}$ is not globally exact on S^∞ (discontinuities in its potential) gives rise to the magnetic mass.

Definition: The magnetic mass of an asymptotically NUT (source-free) gravitational magnetic monopole solution M, is defined as the value of an element $\tilde{\Omega}_{ab}$ of $H^2_\infty(M)$—second homology group[19] of the asymptotic region of M—on the two-chain S^∞.

Corollary 1: The magnetic mass can be conveniently labeled by

$$C_1 = \int_{S^\infty} \tilde{\Omega}_{ab}\, dS^{ab},$$

the first Chern class of the $\mathscr{I}$ bundle. The parametrization of $H^2(M,R)$ provides the (integer) number of twists of $\mathscr{I}$ around its S^1 Hopf fiber.

Corollary 2: If $H^2(M,R)$ is trivial, $\tilde{\Omega}_{ab}$ is a closed and globally exact two-form: the magnetic mass vanishes, implying that $\mathscr{I}$ is a trivial $S^2 \times R$ bundle (zero twist).

As we shall see, in the next section, the value of the magnetic mass provides a homomorphism from $H^2(M)$ into the first homotopy group of the asymptotic Lorentz Lie group which enables us to keep track of the homotopy classes as one moves within a (possibly multiply connected) space-time. This is precisely what is expected from a criterion for the existence of spinor structure.

IV. CONFORMAL NULL BOUNDARY AND AHARONOV-SUSSKIND APPARATUS

We want to show that the available structure at conformal null infinity (for the source-free nonradiative solutions considered in the previous section) could be used to design a device for the detection of spinor structure. As was already mentioned in the Introduction, the resulting setup could be viewed as a gravitational analog of that proposed by Aharonov–Susskind.

Recall the existence of an exact sequence of homotopy groups[20–23]:

$$\pi_n(M,x) \xrightarrow{h_n} \pi_{n-1}(G_x) \xrightarrow{i_{n-1}} \pi_{n-1}(G,u)$$

$$\to \pi_{n-1}(M,x) \to \cdots \to \pi_0(M,x) \to 0,$$

where M is the base space of a principal bundle with fiber G, x a point of M, G_x the corresponding fiber, and each homomorphism in the sequence is the kernel of the next. [This sequence reflects in particular the intertwining of the fundamental group of G,($\pi_1(G)$) with that of the base space ($\pi_1(M)$).] If M is the space-time manifold and $G = \mathscr{L}$ (the general Lorentz Lie group), $h_2 \equiv 0$ is a necessary and sufficient condition for i_1 to be a one to one mapping, i.e., for the existence[6] on (a noncompact) M of a global system of (orthonormal) tetrads. This in turn provides a necessary and, in many cases,[9] sufficient condition for the existence of spinor structure. Since a spinor structure is associated to a homomorphism from $\pi_1(L)$ into Z_2, we shall search for a mechanism to detect a charge taking the value zero or unity after a physical object (e.g., an orthonormal tetrad) has been transported around closed paths in $\mathscr{L}$ or M. If a spinor structure is to be available, one expects such a mechanism to be associated to a mapping from $\pi_2(M,x)$ to the trivial homotopy class of $\mathscr{L}_x - x$ chosen in the asymptotic region. Along these lines, an argument has already been proposed,[7] which relates the space-time curvature in the neighborhood of a point p to an upper bound on the length of those closed paths in $\mathscr{L}_p$, which should be contractible to a point. Since the topology of the underlying space-time manifold determines the existence of spinor structure, sufficiently "twisted" manifolds should be such that any metric that could be defined on them would contain a minimum amount of curvature that would prevent the existence of spinor structure. We shall show that the space-time topology (second homology class) or equivalently the bundle structure available at conformal null infinity, not only controls this amount of curvature but provides a necessary (and in many cases sufficient) condition for the existence of spinor structure. Our result will be the following: the nonvanishing of the first Chern class—a measurement of the magnetic mass enclosed within an asymptotic region, as related to a parametrization of the second homology class of the space-time manifold—provides a nonzero mapping h_2: $\pi_2(M,x) \to \pi_1(\mathscr{L}_x)$, thus preventing the existence of a one-to-one mapping between the

homotopy classes of the fibers of the frame bundle, and consequently preventing the existence of a spinor structure. If the magnetic mass vanishes (trivial bundle structure at conformal null infinity), $h_2 = 0$, i_2 is one to one, and, with the exception of situations involving a simply connected principal frame bundle, a spinor structure will be available.

Let p be a point in the neighborhood of infinity, $\mathscr{L}_p$ the Lorentz Lie group at p, and $\mathscr{C} \subset \mathscr{L}_p$, a closed path through p, i.e., a one-parameter family $R^{\,b}_{a}(s) 0 \leqslant s \leqslant 1$ of Lorentz rotations at p. We want to evaluate the length of $\mathscr{C}$; thus we need to introduce a metric on the rotation group. Choose in M a spacelike two-sphere S through p, which surrounds the point at spacelike infinity. Denote by $\gamma_s(u)$, $0 \leqslant s \leqslant 1$, $\gamma_s(0) = \gamma_s(1) = p$, a one-parameter family of loops on S, originating and ending at p, and spanning S (all points of S, except p, lie on exactly one curve). Let t^a denote an arbitrary unit timelike vector field on $M (t \cdot t = -1)$, adjust S so that t^a and S are orthogonal at p, and introduce at p a triad x^a_λ, $\lambda = 1,2,3$, which, together with t^a, defines an orthonormal tetrad. For each value of s, $0 < s < 1$, we shall (Fermi) transport the tetrad around γ_s according to

$$u^m \nabla_m x^a_\lambda = + t^a (t_c u^m \nabla_m x^c_\lambda), \quad \lambda = 1,2,3, \tag{26}$$

where u^m denotes the tangent to $\gamma_s(u)$ $(u^m \nabla_m u = 1)$. In this transport the tetrad remains an orthonormal tetrad although (in general) after a complete tour on γ_s, x^a_λ $(\lambda = 1,2,3)$ might not be tangential to S. As a result of this transport, we define at p an element $R^{\,a}_{b}(s)$ of the Lorentz Lie group $\mathscr{L}_p$: $x^a_\lambda|_{u=1} = R^{\,a}_{b}(s)(x^b_\lambda|_{u=0})$. As s varies from 0 to 1, γ_s spans the two sphere S and $R^{\,a}_{b}(s)$ $(0 \leqslant s \leqslant 1)$ describes a closed path $\mathscr{C}$ in $\mathscr{L}_p$ [$R^{\,a}_{b}(0) = R^{\,a}_{b}(1) =$ identity]. We want to evaluate the length of $\mathscr{C}$ and decide on its homotopy class. If a spinor structure is to exist, this class should be the trivial one (connected to the identity), (i.e., $\mathscr{C}$ should be contractible to a point). Here $\tau^a_b = (d/ds) R^{\,a}_{b}(s)$, the tangent to $\mathscr{C}$, is associated to an infinitesimal rotation (a generator $F_{[ab]}$ of the Lorentz Lie algebra) and its length is proportional to the amount of curvature enclosed on S, within the two-loops γ_s and γ_{s+ds}. Since S is (in the asymptotic region) surrounding the point at spacelike infinity, the leading term is provided by

$$\int_\sigma \Omega_{ab}\, dS^{ab}, \tag{27}$$

where Ω_{ab} is the two-form, introduced in Sec. III, whose pullback to $\mathscr{I}$ provides the bundle curvature two-form, and where σ is, on S^∞, the area enclosed within γ_s^+ and γ_{s+ds}^+. (Here S^∞ is the two-sphere of null generators of $\mathscr{I}$.) Thus the total length of $\mathscr{C}$ is given by

$$C_1 = \left| \int_{S_\infty} \Omega_{ab}\, dS^{ab} \right|, \tag{28}$$

i.e., the first Chern class of the $\mathscr{I}$ bundle (or equivalently the enclosed magnetic mass). A convenient labeling of C_1 is provided by the second homology class of M, hence the following theorems.

Theorem 1: If a space-time with one asymptotic region is enclosing a nonvanishing magnetic mass, a global system of orthonormal tetrads cannot be defined, which rules out spinor structure.

Theorem 2: The nonvanishing of the first Chern class or, equivalently, the existence of S_1 (Hopf fibering) nontrivial bundle structure over S^2 at conformal null infinity, prevents the existence of global spinor fields on the corresponding (space-time) asymptotic region.

Theorem 3: The vanishing of the first Chern class provides a necessary (and in many cases sufficient) condition for the existence of spinor structure.

For an example where the space-time topology prevents the condition from being sufficient, see, e.g., the Appendix in Ref. 9.

V. CONCLUDING REMARKS

(a) It was shown in Ref. 7 that a necessary and sufficient condition for a noncompact space-time M to have spinor structure is that M may be given a global system of orthonormal tetrads. It was also underlined that this condition might not be the most convenient one to decide on this issue. In Sec. IV we related the criterion to the global topological structure of the space-time manifold, reflected by the topology of its conformal null boundary, and reduced the test for existence of spinor structure to that of detection of a vanishing or nonvanishing topological charge (the magnetic mass) conveniently labeled by the value of suitable homology classes. Thus, although curvature plays a role [e.g., in the labeling of the homotopy classes of the (asymptotic) Lorentz Lie group], it is rather its intertwining with the underlying topology that is crucial. The existence of spinor structure reflects the global structure of the space-time manifold, and originates in the topology.

(b) The existence of spinor structure has also been characterized[7] in terms of the type of the Weyl tensor C_{abcd}, a sufficient condition being that this tensor be *everywhere* type [1,1,1,1], [2,1,1], [3,1], or [4]. Here again the criterion involves the existence of continuous transport of orthonormal tetrads related to the principal null directions of C_{abcd}.

The above results suggest that a relation might exist between the algebraic type of solution to the source-free Einstein equation and its global (topological) structure as described by topological charges, or by the parametrization of suitable homology (cohomology) classes. Since these characteristic classes are conveniently used in the classification of fiber bundles, one might hope to be able to relate them to the algebraic type of the ($\mathscr{I}$-bundle) curvature two-form or to the algebraic type[24] of the Weyl curvature.

(c) We have proved elsewhere that gravitational magnetic monopoles exhibit causality violations—continuous time reversal (e.g., if an everywhere timelike and complete Killing vector field is available, its orbits must be closed, or in absence of isometry, nontrivial bundle structure with S^1 Hopf fibering over S^2 at conformal null infinity). Such solutions thus could illustrate a situation where parity reversal becomes a continuous transformation, PT invariance being a reasonable statement, e.g., when the principle frame bundle over the space-time manifold is simply connected. The "unwinding" of gravitational magnetic monopoles (i.e., NUT solution) into (black-hole) electric mass monopoles (i.e.,

Schwarszchild black hole) could be viewed as an onset of two asymptotic regions, of a discrete PT symmetry and of unambiguous notions of parity and time reversal. As was already mentioned, an invariant (gravitational charge C) is also involved in the transformation.

Since nontrivial topologies and multiply connected space-times could provide observational evidence of their existence[25] at some stage of the evolution of the universe (periodicity in the distribution of quasar red shifts?) we hope that the above considerations might consequently shed light on a possible (experimental) confirmation of a link between the CPT theorem and the existence of spinor structure.[26,27]

ACKNOWLEDGMENTS

I would like to thank Professor D. Boulware, Professor E. T. Newman, and Professor R. Wald for useful conversations. Special thanks are due to Professor J. Hartle for his very kind help. I also want to thank J. Reynolds for his prompt processing of this manuscript at the Institute for Theoretical Physics, Santa Barbara.

This research was supported in part by the National Science Foundation under Grant No. PHY82-17853, supplemented by funds from the National Aeronautics and Space Administration, at the University of California at Santa Barbara.

[1]Ya.B. Zel'dovich and I. D. Novikov, JETP Lett. **1967**, 2136; I. D. Novikov, A. G. Polnarev, A. A. Starobinsky, and Ya.B. Zel'dovich, Astron. Astrophys. **80**, 104 (1979); Ya.B. Zel'dovich and I. D. Novikov, *Structure and Evolution of the Universe* (U. Chicago P., Chicago, 1983), revised edition.

[2]M. Suveges, Acta Phys. Hung. **20**, 273 (1966).

[3]A. S. Eddington, *Space-Time and Gravitation: An Outline of the Theory of Gravitation* (Cambridge U.P., Cambridge, 1959).

[4]R. Penrose, *Structure of Space-Time, Battelle Rencontres 1967* (Benjamin, New York, 1968).

[5]R. Penrose, "Null hypersurface initial data for classical fields of arbitrary spin and for General Relativity," in report ARL 63-56 USAF, 1963.

[6]R. Geroch, J. Math. Phys. **9**, 1739 (1968).

[7]R. Geroch, J. Math. Phys. **11**, 343 (1970).

[8]Y. Aharonov and L. Susskind, Phys. Rev. **158**, 1237 (1967).

[9]C.T.S. Clarke, Gen. Relativ. Gravit. **2**, 43 (1971).

[10]E. P. Wigner, Ann. Math. **40**, 149 (1939).

[11]A. Magnon, J. Math. Phys. **27**, 1059 (1986).

[12]A. Magnon, "Mass, dual mass and gravitational entropy," J. Math. Phys., submitted for publication.

[13]A. Magnon, J. Math. Phys. **27**, 1066 (1986).

[14]R. Penrose, Proc. R. Soc. London Ser. A **284**, 159 (1965).

[15]R. Geroch, J. Math. Phys. **12**, 6 (1971).

[16]M. Buchdahl, Quart. J. Math. **5** (1954).

[17]J. Ehlers, in "Les Théories relativistes de la gravitation," CNRS, Paris, 1959.

[18]B. K. Harrison, J. Math. Phys. **9**, 1744 (1968).

[19]C. Nash and S. Sen, *Topology and Geometry for Physicists* (Academic, New York, 1983).

[20]N. Steenrod, *The topology of Fiber Bundles* (Princeton U.P., Princeton, NJ, 1951).

[21]J. Milnor, "Lectures on characteristic classes," Princeton University, 1957 (mimeographed notes).

[22]Sz. T. Hu, Homotopy Theory (Academic, New York, 1959).

[23]F. Hirzebuch, *Neue Topologische Methoden in der Algebraichschen Geometric* (Springer, Berlin, 1956).

[24]R. Penrose, Ann. Phys. (NY) **10**, 171 (1960).

[25]L. Z. Fang and H. Sato, "Is the periodicity in the distribution of quasar redshifts an evidence of a multiply connected universe?," 1985 first award-winning essay, Gravity Research Foundation, Gloucester, MA 01930.

[26]R. Penrose and W. Rindler, *Spinors and Space-Time* (Cambridge U.P., Cambridge, 1984), Vol. 1.

[27]R. Penrose, W. Rindler, *Spinors and Space-Time* (Cambridge U.P., Cambridge, 1986), Vol. 2.

Einstein–Maxwell equations and the conformal Ricci collineations

Abbas M. Faridi
Department of Physics, University of California, Santa Barbara, California 93206

(Received 1 October 1986; accepted for publication 11 February 1987)

The Einstein–Maxwell field equations for non-null electromagnetic fields are studied under the assumption of admitting a conformal Ricci collineation. It is shown that a non-null electromagnetic field does not admit any conformal Ricci collineation, unless the generators of the symmetry groups are Killing vector fields. Furthermore, it is shown that the energy-momentum tensor of a non-null electromagnetic field can admit a conformal Ricci collineation, if and only if the collineation is homothetic. The restrictions on non-null Maxwell field, its sources, and its invariants implied by the symmetry condition are calculated. An example of a space-time satisfying the Einstein–Maxwell equations, and admitting a homothetic conformal vector field is also given.

I. INTRODUCTION

The connection between the inherent symmetries of a system, and their corresponding conservation laws has been well established since the early observation of Nöther.[1] In particular, the importance of groups of motions, generated by Killing vector fields of a space-time, and their relation to the conservation laws of energy, momentum, and angular momentum is well known.[2,3] Collineations other than groups of motions have also been studied to a limited extent. It has been shown, for example, that for space-times with zero Ricci tensor, the more familiar symmetries such as motions and conformal and homothetic motions are subcases of a more general symmetry requirement known as curvature collineations.[4] Moreover, it has been shown that the allowable conformal symmetries admitted by a nonflat empty space-time (except in the case of specific type-N metrics) are the homothetic motions.[5] The significance of homothetic motions in general relativity is yet to be elaborated; though they have been used by a number of authors. For example, homothetic motions have been utilized in the analysis of spherically symmetric, and plane symmetric self-similar space-times[6,7]; homothetic Weyl space-times,[8] as well as in certain self-similar cosmologies.[9]

In this paper we are concerned with Einstein–Maxwell field equations

$$R_{\mu\nu} = f_{\mu\sigma} f^{\sigma}_{\nu} - \tfrac{1}{4} g_{\mu\nu} f_{\alpha\beta} f^{\alpha\beta}, \tag{1.1}$$

$$f^{\mu\nu};_{\nu} = j^{\mu}, \tag{1.2}$$

$$*f^{\mu\nu};_{\nu} = 0, \tag{1.3}$$

admitting a conformal motion

$$L_{\zeta} g_{\mu\nu} = 2\phi g_{\mu\nu}, \tag{1.4}$$

where L denotes the Lie derivative, $g_{\mu\nu}$ is the metric of the space-time, and $\phi = \frac{1}{4}\zeta^{\mu};_{\mu}$ is a scalar function. The electromagnetic field $f_{\mu\nu}$, and its dual $*f_{\mu\nu}$ satisfy the identities

$$f_{\mu\sigma} f^{\sigma\nu} - *f_{\mu\sigma} *f^{\sigma\nu} = \tfrac{1}{2}(f_{\alpha\beta} f^{\beta\alpha})\delta^{\nu}_{\mu}, \tag{1.5}$$

$$f_{\mu\sigma} *f^{\sigma\nu} = *f_{\mu\sigma} f^{\sigma\nu} = \tfrac{1}{4}(f_{\alpha\beta} *f^{\beta\alpha})\delta^{\nu}_{\mu}. \tag{1.6}$$

Throughout this paper, the Greek indices are tensor indices and range over the values 1, 2, 3, 4; and the lightface Latin a, b, c,...denote tetrad indices.

It has been shown (see, e.g., Ref. 4) that every curvature collineation

$$L_{\zeta} R^{\mu}\nu\rho_{\sigma} = 0, \tag{1.7}$$

generated by a conformal vector field ζ^{μ} [Eq. (1.4)] is also a Ricci collineation,

$$L_{\zeta} R_{\mu\nu} = 0, \tag{1.8}$$

provided the scalar function ϕ satisfies

$$\phi;\mu\nu = 0. \tag{1.9}$$

The problem that we wish to investigate may be posed in the following context. Suppose that the field equations (1.1)–(1.3) are satisfied for a non-null electromagnetic field, and the space-time admits a conformal Ricci collineation. What restrictions are then imposed on the Maxwell field $f_{\mu\nu}$, and its source j^{μ}? In particular, we want to determine whether or not these limitations depend on the form of the scalar function ϕ, as defined by Eq. (1.4).

II. SOME GEOMETRIC RELATIONS DUE TO RICCI CONFORMAL SYMMETRY CONDITIONS

In this section we make use of the proposed symmetry condition, and derive relations to be fulfilled by the field equations. These results will be used in the subsequent section to simplify the computations.

It is well known that a non-null electromagnetic field $f_{\mu\nu}$, and its dual $*f_{\mu\nu}$, can be expressed in the form

$$f_{\mu\nu} = 2\phi_0(n_{\mu} l_{\nu} - n_{\nu} l_{\mu}) + 2i\psi_0(m_{\mu}\bar{m}_{\nu} - \bar{m}_{\mu} m_{\nu}), \tag{2.1}$$

$$*f_{\mu\nu} = 2\psi_0(n_{\mu} l_{\nu} - n_{\nu} l_{\mu}) - 2i\phi_0(m_{\mu}\bar{m}_{\nu} - \bar{m}_{\mu} m_{\nu}), \tag{2.2}$$

where ϕ_0 and ψ_0 are the real and imaginary parts of the scalar

$$\Phi = \tfrac{1}{2} f_{\mu\nu}(l^{\mu} n^{\nu} + \bar{m}^{\mu} m^{\nu}) = \phi_0 + i\psi_0. \tag{2.3}$$

The null vectors l^{μ}, n^{μ} are real, m^{μ} complex, and satisfy

$$l_{\mu} n^{\mu} = -m_{\mu}\bar{m}^{\mu} = 1, \tag{2.4}$$

with all other contractions being zero.

The two invariants of the electromagnetic fields are

$$f_{\mu\nu} f^{\nu\mu} = 4(\Phi^2 + \bar{\Phi}^2) = 8(\phi_0^2 - \psi_0^2), \tag{2.5}$$

 0022-2488/87/061370-07$02.50

$$f_{\mu\nu}\,{}^*f^{\nu\mu} = 4i(\Phi^2 - \bar{\Phi}^2) = 16\phi_0\psi_0. \tag{2.6}$$

Einstein–Maxwell equations may now be written in terms of the above invariants and the tetrad of null vectors l^μ, n^μ, and m^μ and $\bar{m}^\mu$. We have[10–12]

$$\begin{aligned} R_{\mu\nu} &= \Lambda(l_\mu n_\nu + n_\mu l_\nu + m_\mu \bar{m}_\nu + \bar{m}_\mu m_\nu) \\ &= 2\Lambda(l_\mu n_\nu + n_\mu l_\nu) - \Lambda g_{\mu\nu}, \end{aligned} \tag{2.7}$$

where Λ is defined by the relation

$$\Lambda = 4\Phi\bar{\Phi} = 4(\phi_0^2 + \psi_0^2), \tag{2.8}$$

and the metric tensor has the form

$$g_{\mu\nu} = l_\mu n_\nu + n_\mu l_\nu - \bar{m}_\mu m_\nu - m_\mu \bar{m}_\nu. \tag{2.9}$$

We may express Λ in terms of the invariants by forming the expression

$$R_{\mu\sigma}R^{\sigma\nu} = \Lambda^2\delta^\nu_\mu = \tfrac{1}{4}[(f_{\alpha\beta}f^{\beta\alpha})^2 + (f_{\alpha\beta}\,{}^*f^{\beta\alpha})^2]\delta^\nu_\mu, \tag{2.10}$$

where use has been made of Eqs. (1.5) and (1.6). We therefore have

$$\Lambda^2 = \tfrac{1}{4}(R_{\mu\sigma}R^{\sigma\mu}) = \tfrac{1}{4}[(f_{\mu\nu}f^{\nu\mu})^2 + (f_{\mu\nu}\,{}^*f^{\nu\mu})^2]. \tag{2.11}$$

From Eqs. (1.8) and (2.11) we obtain the restriction imposed on the invariant of the electromagnetic field, and may be written in the form

$$L_\xi\Lambda = -2\phi\Lambda. \tag{2.12}$$

Lie differentiation of Eq. (2.7) yields

$$\begin{aligned} &l_\mu L_\xi(n_\nu) + n_\nu L_\xi(l_\mu) + n_\mu L_\xi(l_\nu) + l_\nu L_\xi(n_\mu) \\ &\quad - 2\phi(l_\mu n_\nu - n_\mu l_\nu) = 0. \end{aligned} \tag{2.13}$$

The contraction of Eq. (2.13) with l^μ, and n^μ, gives, respectively,

$$L_\xi(l_\mu) = 2\phi l_\mu - l_\mu l^\alpha L_\xi(n_\alpha), \tag{2.14}$$

$$L_\xi(n_\mu) = 2\phi n_\mu - n_\mu n^\alpha L_\xi(l_\alpha). \tag{2.15}$$

Equations (2.14) and (2.15) may be simplified if use is made of the normalization condition Eq. (2.4), which after differentiation takes the form

$$n^\mu L_\xi(l_\mu) + l^\mu L_\xi(n_\mu) = 2\phi. \tag{2.16}$$

Eliminating the quantity $n^\alpha L_\xi(l_\alpha)$ by means of Eq. (2.16), we may write

$$L_\xi(l_\mu) = (\phi + \lambda)l\mu, \tag{2.17}$$

$$L_\xi(n_\mu) = (\phi - \lambda)n_\mu, \quad \lambda = \phi - l^\alpha L_\xi(n_\alpha). \tag{2.18}$$

Similarly, making use of Eqs. (1.4), (2.9), (2.17), and (2.18), we obtain

$$m^\mu L_\xi(\bar{m}_\mu) + \bar{m}^\mu L_\xi(m_\mu) = -2\phi, \tag{2.19}$$

$$L_\xi(m_\alpha) = (\phi - \delta)m_\alpha, \tag{2.20}$$

$$L_\xi(\bar{m}_\alpha) = (\phi + \delta)\bar{m}_\alpha, \quad \delta = \phi + \bar{m}^\sigma L_\xi(m_\sigma). \tag{2.21}$$

We are now in a position to write the restrictions implied on Maxwell field, due to conformal Ricci collineation. We obtain, after some algebra

$$\begin{aligned} L_\xi f_{\mu\nu} &= 2L_\xi\phi_0(n_\mu l_\nu - n_\nu l_\mu) \\ &\quad + 2iL_\xi\psi_0(m_\mu\bar{m}_\nu - \bar{m}_\mu m_\nu) + 2\phi f_{\mu\nu}, \end{aligned} \tag{2.22}$$

$$\begin{aligned} L_\xi\,{}^*f_{\mu\nu} &= 2L_\xi\psi_0(n_\mu l_\nu - n_\nu l_\mu) \\ &\quad - 2iL_\xi\phi_0(m_\mu\bar{m}_\nu - m_\nu\bar{m}_\mu) + 2\phi\,{}^*f_{\mu\nu}. \end{aligned} \tag{2.23}$$

Equations (2.22) and (2.23) may be transformed into a more useful form. We may write these equations in the form

$$L_\xi f_{\mu\nu} = A f_{\mu\nu} + B\,{}^*f_{\mu\nu}, \tag{2.24}$$

$$L_\xi\,{}^*f_{\mu\nu} = A\,{}^*f_{\mu\nu} - Bf_{\mu\nu}, \tag{2.25}$$

where

$$A = (\phi_0 L_\xi\phi_0 + \psi_0 L_\xi\psi_0)/(\phi_0^2 + \psi_0^2) + 2\phi, \tag{2.26}$$

$$B = (\psi_0 L_\xi\phi_0 - \phi_0 L_\xi\psi_0)/(\phi_0^2 + \psi_0^2). \tag{2.27}$$

Taking the covariant divergence of Eqs. (2.24) and (2.25), and making use of Maxwell equations, we obtain for the right-hand side of these equations, the relations

$$\gamma^{\mu\nu};_\nu = A_\nu f^{\mu\nu} + B\,{}^*_\nu f^{\mu\nu} + Aj^\mu, \tag{2.28}$$

$$\gamma^{*\mu\nu};_\nu = A_\nu\,{}^*f^{\mu\nu} - B_\nu f^{\mu\nu} - Bj^\mu, \tag{2.29}$$

where

$$\gamma_{\mu\nu} = L_\xi f_{\mu\nu}. \tag{2.30}$$

Finally, applying the identities Eqs. (A3) and (A5) and the compatibility condition Eq. (A4) for the vector field ξ^μ, we obtain for the left-hand side of Eqs. (2.28) and (2.29) the expressions

$$\gamma^{\mu\nu};_\nu = L_\xi(j^\mu) + 4\phi j^\mu, \tag{2.31}$$

$$\gamma^{*\mu\nu};_\nu = 0. \tag{2.32}$$

The implication of the equations derived in this section on the functions A, B, and ϕ as defined by Eqs. (2.26) and (2.27) are sought in the next section.

III. DETERMINATION OF THE FUNCTIONS A, B, AND ϕ

It will be shown here that in order for the system of equations (1.1)–(1.3) to be consistent for a nontrivial, non-null electromagnetic field under Ricci collineation, Eq. (1.8), the allowable conformal vector field must be homothetic,

$$\phi = \text{const}. \tag{3.1}$$

There are two cases to be treated separately; namely we can either have a conformally invariant non-null electromagnetic field

$$\text{case (a): } L_\xi f_{\mu\nu} = 0 \tag{3.2}$$

or

$$\text{case (b): } L_\xi f_{\mu\nu} = \gamma_{\mu\nu} \neq 0. \tag{3.3}$$

Case (a): In this case we obtain from Eqs. (1.1), (2.24), and (2.25), the simple relation

$$(A - 2\phi)R_{\mu\nu} = 0, \tag{3.4}$$

which implies either (i) $A - 2\phi \neq 0$, $R_{\mu\nu} = 0$, or (ii) $A = 2\phi$, $R_{\mu\nu} \neq 0$. The results for the case of vanishing Ricci tensor have been obtained (see, e.g., Ref. 5). On the other hand, if $R_{\mu\nu} \neq 0$, Eqs. (2.8) and (2.26) yield

$$L_\xi\Lambda = 0, \quad \Lambda = 4(\phi_0^2 + \psi_0^2). \tag{3.5}$$

According to Eq. (2.12), a non-null electromagnetic field satisfies Eq. (3.5) if and only if

$$\phi = 0. \tag{3.6}$$

The condition $\phi = 0$, for a nonvanishing energy-momentum tensor of a non-null electromagnetic field, implies by virtue of Eq. (1.4), that the vector field ζ^μ must be the generator of an isometry group. The details of this case have also been worked out by Wooley,[13,14] in his analysis of Einstein–Maxwell equations, under the assumptions $L_\zeta R_{\mu\nu} = 0, \zeta_{(\mu;\nu)} = 0$. The results for case (a) may be summarized in the following theorem.

Theorem 1: In the combined Einstein–Maxwell theory, a non-null electromagnetic field, and the Ricci tensor, cannot simultaneously be conformally invariant.

An immediate consequence of Theorem 1 is that the conformal invariance of Ricci tensor may be shared by the energy-momentum tensor of a non-null electromagnetic field rather than the field itself. This in fact turns out to be the case, and embodies the main content of the case (b), specified by Eq. (3.3).

Case (b): In this case Eqs. (2.28)–(2.32), and the vanishing of the covariant divergence of the energy-momentum tensor yield

$$B_\nu\, {}^*f^{\mu\nu} = L_\zeta(j^\mu) + (4\phi - A)j^\mu - A_\nu f^{\mu\nu}, \tag{3.7}$$

$$B_\nu f^{\mu\nu} = A_\nu\, {}^*f^{\mu\nu} - Bj^\mu, \tag{3.8}$$

$$f^{\mu\nu} j^\nu = 0. \tag{3.9}$$

The analysis of these equations depends on whether the electromagnetic field is source-free, $j^\mu = 0$, which we label it as case (1b), or otherwise, $j^\mu \neq 0$, which we denote it as case (2b). Consequently, the behavior of the functions A, B, and ϕ, if they exist at all, may depend on quite different conditions. Similarly, by virtue of Eqs. (2.24)–(2.27), the behavior of a nontrival Maxwell field may likewise be different.

Case (1b): Let us consider the source-free case. In this case, Eqs. (3.7)–(3.9) become

$$B_\nu f^{\mu\nu} = A_\nu\, {}^*f^{\mu\nu}, \quad j^\mu = 0, \tag{3.10}$$

$$B_\nu\, {}^*f^{\mu\nu} = -A_\nu f^{\mu\nu}. \tag{3.11}$$

Inserting in these equations, the expressions for the field and its dual from Eqs. (2.1) and (2.2) and contracting the resulting two equations with the null vectors l^μ, n^μ, m^μ, and $\bar{m}^\mu$, we find that for a non-null electromagnetic field, we must have

$$A_\mu = 0, \quad B_\mu = 0. \tag{3.12}$$

The constant value of A may be obtained from Eqs. (2.12) and (2.26). In fact we have

$$A = \phi = \text{const}. \tag{3.13}$$

The constant B is then restricted by either one of the following relations, obtained from Eqs. (2.26) and (2.27):

$$L_\zeta \phi_0 = -\phi\phi_0 + B\psi_0, \quad j^\mu = 0, \tag{3.14}$$

$$L_\zeta \psi_0 = -\phi\psi_0 - B\phi_0, \quad j^\mu = 0. \tag{3.15}$$

The corresponding relations for the Maxwell field, given by Eqs. (2.24) and (2.25), assume the form

$$L_\zeta f_{\mu\nu} = \phi f_{\mu\nu} + B\, {}^*f_{\mu\nu}, \quad j^\mu = 0, \tag{3.16}$$

$$L_\zeta\, {}^*f_{\mu\nu} = \phi\, {}^*f_{\mu\nu} - Bf_{\mu\nu}. \tag{3.17}$$

Case (2b): In this case, where $j^\mu \neq 0$, Eq. (3.9) implies that for a nontrival j^μ to exist, the determinant of the admissible non-null electromagnetic field must vanish. The vanishing of the determinant is equivalent to the vanishing of the invariant[15]

$$f_{\mu\nu}\, {}^*f^{\nu\mu} = 0. \tag{3.18}$$

The physical implication of Eq. (3.18) is the orthogonality of the electric and the magnetic fields in a local Minkowskian frame. From Eqs. (2.6) and (3.18) we have either

$$\psi_0 = 0, \quad \phi_0 \neq 0, \quad j^\mu \neq 0, \tag{3.19}$$

or

$$\phi_0 = 0, \quad \psi_0 \neq 0. \tag{3.20}$$

In either case, the function B, as given by Eq. (2.27) is equal to zero. Equations (3.7) and (3.8) and the Lie differentiation of (3.9) give us

$$L_\zeta j^\mu + (4\phi - A)j^\mu - A_\nu f^{\mu\nu} = 0, \tag{3.21}$$

$$A_\nu f\, {}^{*\mu\nu} = 0, \tag{3.22}$$

$$f_{\nu\mu} L_\zeta j^\mu = 0. \tag{3.23}$$

Multiplying Eq. (3.21) by $f_{\sigma\mu}$, and making use of Eqs. (1.5), (1.6), and (3.22), we obtain

$$(f\cdot f)A_\mu = 0, \quad (f\cdot f) = f_{\mu\nu} f^{\nu\mu}. \tag{3.24}$$

Since $(f\cdot f)$ is the only nonvanishing invariant of the electromagnetic field in this case, we must have $A_\mu = 0$, or A is at most a constant. With A being a constant, Eqs. (2.12) and (2.26) give the results

$$A = \phi = \text{const}, \tag{3.25}$$

$$L_\zeta(\phi_0) = -\phi\phi_0, \quad \phi_0 \neq 0, \quad \psi_0 = 0, \quad j^\mu \neq 0, \tag{3.26}$$

$$L_\zeta(\psi_0) = -\phi\psi_0, \quad \psi_0 \neq 0, \quad \phi_0 = 0. \tag{3.27}$$

The corresponding restrictions on the electromagnetic field and its source are obtained from Eqs. (2.24), (2.25), and (3.21). They will take the form

$$L_\zeta f_{\mu\nu} = \phi f_{\mu\nu}, \tag{3.28}$$

$$L_\zeta\, {}^*f_{\mu\nu} = \phi\, {}^*f_{\mu\nu}, \quad j^\mu \neq 0, \tag{3.29}$$

$$L_\zeta j^\mu = -3\phi j^\mu. \tag{3.30}$$

The results obtained for case (b) may be stated in the following theorem.

Theorem 2: In the combined Einstein–Maxwell theory, the energy momentum tensor of a non-null electromagnetic field, with or without a source, and the Ricci tensor, are simultaneously conformally invariant, if and only if the conformal vector field is homothetic.

IV. A PARTICULAR SOLUTION

In this section we construct a special source-free ($j^\mu = 0$) solution, satisfying the conditions (a) the principal null directions l^μ, and n^μ are geodesics, (b) the null tetrad of vector l^μ, n^μ, m^μ, and $\bar{m}^\mu$ are parallelly propagated along l^μ and n^μ, and (c) the null geodesics with tangent vectors l^μ and n^μ are twist-free.

The particular solution to Einstein–Maxwell equations satisfying the above requirements has been found, using the Cartan's equations of structure. We will not, however, present the details of the calculations here, since it has also been

worked out by Tariq and Tupper[16,17] in Newman–Penrose formalism (Debney and Zund[18,19] also give some useful results shared by the principal null geodesics, and when the tetrad vectors are parallelly propagated along them). Our problem therefore reduces to determine whether or not Eq. (1.4) admits a nontrivial homothetic solution for the given metric obtained from the field equations.

The solution to Einstein–Maxwell equations under the assumed conditions (a), (b), and (c), may be specified by the relations

$$l^{\mu}=\delta^{\mu}_{r},\quad n^{\mu}=\delta^{\mu}_{u},\quad m^{\mu}=\lambda_1\delta^{\mu}_{y}+\lambda_2\delta^{\mu}_{z}, \tag{4.1}$$

where the coordinates are x^{μ}: $(x^1=u,\ x^2=r,\ x^3=y,\ x^4=z)$, and the functions λ_1, and λ_2 are, respectively,

$$\lambda_1=(1/\sqrt{2})u^n r^m,\quad \lambda_2=(i/\sqrt{2})u^m r^n, \tag{4.2}$$

with $m=(\sqrt{3}-1)/4$, and $n=-(\sqrt{3}+1)/4$. The metric is defined by Eq. (2.9) and the Maxwell field by Eqs. (2.1)–(2.3), with

$$\Phi\bar{\Phi}=\phi_0^2+\psi_0^2=1/8ur. \tag{4.3}$$

The only nonvanishing spin coefficients are

$$\rho=\bar{\rho}=-1/4r,\quad \mu=\bar{\mu}=1/4u, \tag{4.4}$$

$$\sigma=\bar{\sigma}=\sqrt{3}/4r,\quad \lambda=\bar{\lambda}=\sqrt{3}/4u, \tag{4.5}$$

having the intrinsic derivatives

$$D\rho=4\rho^2,\quad D\sigma=4\rho\sigma,\quad \Delta\mu=-4\mu^2,\quad \Delta\lambda=-4\lambda\mu, \tag{4.6}$$

with all other intrinsic derivatives being zero.

With these spin coefficients, the commutation relations Eqs. (B10a)–(B10d), the tetrad components of the homothetic vector field ζ^a $(a=1,2,3,4)$, Eqs. (B5a)–(B5j), and its corresponding compatibility conditions, Eqs. (B13)–(B17), assume the form, respectively;

$$(\Delta D-D\Delta)\psi=0, \tag{4.7a}$$

$$(\delta\bar{\delta}-\bar{\delta}\delta)\psi=0, \tag{4.7b}$$

$$(\delta D-D\delta)\psi=-\rho\delta\psi-\sigma\bar{\delta}\psi, \tag{4.7c}$$

$$(\delta\Delta-\Delta\delta)\psi=\mu\delta\psi+\lambda\bar{\delta}\psi, \tag{4.7d}$$

$$D\zeta_1=0, \tag{4.8a}$$

$$\Delta\zeta_2=0, \tag{4.8b}$$

$$\delta\zeta_3=\lambda\zeta_1-\sigma\zeta_2, \tag{4.8c}$$

$$\Delta\zeta_1+D\zeta_2=2\phi, \tag{4.8d}$$

$$\delta\zeta_1+D\zeta_3=-\rho\zeta_3-\sigma\zeta_4, \tag{4.8e}$$

$$\delta\zeta_2+\Delta\zeta_3=\mu\zeta_3+\lambda\zeta_4, \tag{4.8f}$$

$$\bar{\delta}\zeta_3+\delta\zeta_4=-2\phi+2\mu\zeta_1-2\rho\zeta_2, \tag{4.8g}$$

$$\bar{\delta}\zeta_4=\lambda\zeta_1-\sigma\zeta_2, \tag{4.8h}$$

$$\bar{\delta}\zeta_1+D\zeta_4=-\rho\zeta_4-\sigma\zeta_3, \tag{4.8i}$$

$$\bar{\delta}\zeta_2+\Delta\zeta_4=\mu\zeta_4+\lambda\zeta_3, \tag{4.8j}$$

$$\zeta_1\Delta\psi_0+\zeta_2D\psi_0-\zeta_3\bar{\delta}\psi_0-\zeta_4\delta\psi_0+2\psi_0(D\zeta_2-\delta\zeta_4)=2\psi_0(\phi+\rho\zeta_2-\mu\zeta_1), \tag{4.9a}$$

$$\psi_0\Delta\zeta_4+\psi_2(2D\zeta_3-\delta\zeta_1)=2\psi_0(\lambda\zeta_3+\mu\zeta_4)-2\psi_2(\rho\zeta_3+\sigma\zeta_4), \tag{4.9b}$$

$$\zeta_1\Delta\psi_2+\zeta_2D\psi_2-\zeta_3\bar{\delta}\psi_2-\zeta_4\delta\psi_2=-2\phi\psi_2, \tag{4.9c}$$

$$\psi_4D\zeta_3-\psi_2(\bar{\delta}\zeta_2-2\Delta\zeta_4)=2\psi_2(\lambda\zeta_3+\mu\zeta_4)-\psi_4(\rho\zeta_3+\sigma\zeta_4), \tag{4.9d}$$

$$\zeta_1\Delta\psi_4+\zeta_2D\psi_4-\zeta_3\bar{\delta}\psi_4-\zeta_4\delta\psi_4+2\psi_4(\Delta\zeta_1-\bar{\delta}\zeta_3)=2\psi_4(\phi-\mu\zeta_1+\rho\zeta_2), \tag{4.9e}$$

where

$$\psi_0=2\rho\sigma,\quad \psi_2=2\rho\mu,\quad \psi_4=2\lambda\mu, \tag{4.10}$$

are the nonvanishing scalars associated with the Weyl tensor.

The compatibility conditions, Eqs. (4.9a)–(4.9e), and the commutation relations Eqs. (4.7a)–(4.7d), when use is made of Eq. (4.10), give us the results

$$\delta\zeta_1=\bar{\delta}\zeta_1=\delta\zeta_2=\bar{\delta}\zeta_2=0, \tag{4.11}$$

$$\Delta\zeta_1=4\mu\zeta_1, \tag{4.12}$$

$$D\zeta_2=-4\rho\zeta_2, \tag{4.13}$$

$$\Delta\zeta_3=\mu\zeta_3+\lambda\zeta_4, \tag{4.14}$$

$$\delta\zeta_3=\lambda\zeta_1-\sigma\zeta_2, \tag{4.15}$$

$$D\zeta_4=-\rho\zeta_4-\sigma\zeta_3, \tag{4.16}$$

$$\Delta\zeta_4=\mu\zeta_4+\lambda\zeta_3. \tag{4.17}$$

The integration of Eqs. (4.8a)–(4.8j) gives for the tetrad components of ζ^a,

$$\zeta_1=cu, \tag{4.18}$$

$$\zeta_2=(2\phi-c)r, \tag{4.19}$$

$$\zeta_3=d_1yu^{-n}r^{-m}-id_2zu^{-m}r^{-n}+F_1(u,r), \tag{4.20}$$

$$\zeta_4=d_1yu^{-n}r^{-m}+id_2zu^{-m}r^{-n}+F_2(u,r), \tag{4.21}$$

where c, d_1, and d_2 are constants, and F_1, F_2 are two arbitrary integration constants. The constants d_1 and d_2, may be expressed in terms of the constant c, and the homothetic constant ϕ,

$$d_1=(\alpha-\sqrt{2}\phi)/2,\quad d_2=\alpha/2,\qquad \alpha=\sqrt{6}c/2+\sqrt{2}(1-\sqrt{3})\phi/2. \tag{4.22}$$

The arbitrary constants of integrations F_1, F_2 must satisfy

$$\begin{aligned} &F_{1,r}-(1/4r)F_1+(\sqrt{3}/4r)F_2=0,\\ &F_{1,u}-(1/4u)F_1-(\sqrt{3}/4u)F_2=0,\\ &F_{2,r}-(1/4r)F_2-(\sqrt{3}/4r)F_1=0,\\ &F_{2,u}-(1/4u)F_2-(\sqrt{3}/4u)F_1=0. \end{aligned} \tag{4.23}$$

The only admissible solution to the system of equations in (4.23) is the trivial solution

$$F_1(u,r)=F_2(u,r)=0. \tag{4.24}$$

The complete particular solution to the Eqs. (1.1)–(1.9) with respect to the system of coordinates (u,r,y,z) may be expressed in the following form.

For the metric we have

$$g_{\mu\nu}=\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&-u^{-2n}r^{-2m}&0\\0&0&0&-u^{-2m}r^{-2n}\end{pmatrix},$$
$$g^{\mu\nu}=\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&-u^{2n}r^{2m}&0\\0&0&0&-u^{2m}r^{2n}\end{pmatrix}, \tag{4.25}$$

where $m=(\sqrt{3}-1)/4$, and $n=-(\sqrt{3}+1)/4$. The Maxwell field becomes

$$f_{\mu\nu}=\begin{pmatrix}0&-[(\cos\epsilon)/\sqrt{2}](ur)^{-1/2}&0&0\\[(\cos\epsilon)/\sqrt{2}](ur)^{-1/2}&0&0&0\\0&0&0&(\sin\epsilon)/\sqrt{2}\\0&0&-(\sin\epsilon)/\sqrt{2}&0\end{pmatrix}, \tag{4.26}$$

where ϵ is an arbitrary real constant. The homothetic vector field ζ^μ takes the form

$$\zeta^u=cu,\quad \zeta^r=(2\phi-c)r,$$
$$\zeta^y=-c_1 y,\quad \zeta^z=c_2 z, \tag{4.27}$$

where the constants c_1 and c_2 are

$$c_1=\sqrt{3}(c-\phi)/2,\quad c_2=\sqrt{3}c/2+(1-\sqrt{3})\phi/2. \tag{4.28}$$

From Eq. (4.26), we see that both electric and magnetic fields are in the radial direction with respect to the adopted coordinate system (u,r,y,z). In addition to the homothetic motion, the solution admits a three-parameter group of motions,

$$\eta_1^\mu=\delta_y^\mu,\quad \eta_2^\mu=\delta_z^\mu,$$
$$\eta_3^\mu=u\delta_u^\mu-r\delta_r^\mu-dy\delta_y^\mu+dz\delta_z^\mu,$$

where the constant $d=\sqrt{3}c/2$ may be obtained from Eq. (4.28), by letting $\phi=0$.

APPENDIX A: CONFORMAL MOTION

We summarize here some of the relevant relations concerning a conformal motion. A space-time is said to admit a conformal motion if there exists a vector field ζ^μ, such that

$$L_\zeta g_{\mu\nu}=\zeta_{\mu;\nu}+\zeta_{\nu;\mu}=2\phi g_{\mu\nu}, \tag{A1}$$

where the symbol L_ζ denotes Lie differentiation with respect to the vector field ζ^μ, and ϕ is a scalar function satisfying

$$\phi=\tfrac{1}{4}\zeta^\mu{}_{;\mu}. \tag{A2}$$

Every conformal motion must satisfy

$$L_\zeta\Gamma^\mu_{\nu\sigma}=\phi;\sigma\delta^\mu_\nu+\phi;\nu\delta^\mu_\sigma-g_{\nu\sigma}g^{\mu\alpha}\phi;\alpha. \tag{A3}$$

The integrability condition for (A1) can be shown to have the form

$$\zeta_{\mu;\nu;\sigma}=\zeta^\alpha R_{\alpha\sigma\nu\mu}+\phi;\sigma g_{\mu\nu}+\phi;\nu g_{\mu\sigma}-\phi;\mu g_{\nu\sigma}. \tag{A4}$$

For an arbitrary tensor $K^{\mu\nu}$, we have the identity[20]

$$L_\zeta(K^{\mu\nu};\alpha)-(L_\zeta K^{\mu\nu});\alpha=(L_\zeta\Gamma^\mu_{\alpha\beta})K^{\beta\nu}+(L_\zeta\Gamma^\nu_{\alpha\beta})K^{\mu\beta}. \tag{A5}$$

The Lie differentiation of the Riemann tensor can be expressed in the form

$$L_\zeta R^\alpha_{\mu\nu\sigma}=(L_\zeta\Gamma^\alpha_{\sigma\mu});\nu-(L_\zeta\Gamma^\alpha_{\nu\mu});\sigma. \tag{A6}$$

The curvature collineation is defined by the relation

$$L_\zeta R^\alpha_{\mu\nu\sigma}=0, \tag{A7}$$

which by virtue of Eqs. (A2) and (A3) gives the restriction on ϕ

$$\phi;\mu\nu=0. \tag{A8}$$

The contraction of Eq. (A7) yields the Ricci collineation

$$L_\zeta R_{\mu\nu}=0. \tag{A9}$$

The integrability conduction for (A1) is generally expressed in terms of the vanishing of the Lie derivative of the conformal Weyl tensor

$$L_\zeta C^\alpha_{\mu\nu\sigma}=0. \tag{A10}$$

APPENDIX B: CONFORMAL MOTION AND ITS COMPATIBILITY CONDITIONS IN TETRAD REPRESENTATION

In the following we write the tetrad components equations of Eqs. (A1) and (A10). We choose a tetrad of null vector l^μ, n^μ, m^μ, $\bar{m}^\mu$, with l^μ and n^μ real and m^μ complex. The only nonvanishing contractions are

$$l_\mu n^\mu=-m_\mu\bar{m}^\mu=1. \tag{B1}$$

A frame defined by the inner product

$$e^\mu_a e^b_\mu=\delta^b_a,\quad e^\mu_a e^a_\nu=\delta^\mu_\nu, \tag{B2}$$

where $e^\mu_a(l^\mu,n^\mu,m^\mu,\bar{m}^\mu)$, induces a metric of the form

$$\eta_{ab}=\eta^{ab}=\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&0&-1\\0&0&-1&0\end{pmatrix},\quad a,b=1,2,3,4. \tag{B3}$$

Equation (A1) in tetrad representation takes the form

$$\zeta_{a;b}+\zeta_{b;a}=2\phi\eta_{ab}+\zeta^c(\gamma_{acb}+\gamma_{bca}), \tag{B4}$$

where γ_{abc} are the Ricci rotation coefficients. Equation (B4) is equivalent to the set of scalar equations

$$D\zeta_1 = (\epsilon + \bar{\epsilon})\zeta_1 - \bar{\kappa}\zeta_3 - \kappa\zeta_4, \tag{B5a}$$

$$\Delta\zeta_2 = -(\gamma + \bar{\gamma})\zeta_2 + \nu\zeta_3 + \bar{\nu}\zeta_4, \tag{B5b}$$

$$\delta\zeta_3 = \bar{\lambda}\zeta_1 - \sigma\zeta_2 - (\bar{\alpha} - \beta)\zeta_3, \tag{B5c}$$

$$\bar{\delta}\zeta_4 = \lambda\zeta_1 - \bar{\sigma}\zeta_2 - (\alpha - \bar{\beta})\zeta_4, \tag{B5d}$$

$$\Delta\zeta_1 + D\zeta_2 = 2\phi + (\gamma + \bar{\gamma})\zeta_1 - (\epsilon + \bar{\epsilon})\zeta_2 + (\pi - \bar{\tau})\zeta_3 + (\bar{\pi} - \tau)\zeta_4, \tag{B5e}$$

$$\delta\zeta_1 + D\zeta_3 = (\bar{\alpha} + \beta + \bar{\pi})\zeta_1 - \kappa\zeta_2 + (\epsilon - \bar{\epsilon} - \bar{\rho})\zeta_3 - \sigma\zeta_4, \tag{B5f}$$

$$\bar{\delta}\zeta_1 + D\zeta_4 = (\alpha + \bar{\beta} + \pi)\zeta_1 - \bar{\kappa}\zeta_2 + (\bar{\epsilon} - \epsilon - \rho)\zeta_4 - \bar{\sigma}\zeta_3, \tag{B5g}$$

$$\delta\zeta_2 + \Delta\zeta_3 = \bar{\nu}\zeta_1 - (\bar{\alpha} + \beta + \tau)\zeta_2 + (\mu + \gamma - \bar{\gamma})\zeta_3 + \bar{\lambda}\zeta_4, \tag{B5h}$$

$$\bar{\delta}\zeta_2 + \Delta\zeta_4 = \nu\zeta_1 - (\alpha + \bar{\beta} + \bar{\tau})\zeta_2 + (\bar{\mu} + \bar{\gamma} - \gamma)\zeta_4 + \lambda\zeta_3, \tag{B5i}$$

$$\bar{\delta}\zeta_3 + \delta\zeta_4 = -2\phi + (\mu + \bar{\mu})\zeta_1 - (\rho + \bar{\rho})\zeta_2 + (\alpha - \bar{\beta})\zeta_3 + (\bar{\alpha} - \beta)\zeta_4, \tag{B5j}$$

where the spin coefficients $\alpha\beta\gamma\cdots$ are related to the Ricci rotations and may be expressed in the form

$$\gamma_{131} = \kappa, \quad \gamma_{132} = \tau, \quad \gamma_{133} = \sigma, \quad \gamma_{134} = \rho, \tag{B6}$$

$$\gamma_{241} = -\pi, \quad \gamma_{242} = -\nu, \quad \gamma_{243} = -\mu, \quad \gamma_{244} = -\lambda, \tag{B7}$$

$$\gamma_{121} = \epsilon + \bar{\epsilon}, \quad \gamma_{122} = \gamma + \bar{\gamma}, \quad \gamma_{123} = \beta + \bar{\alpha}, \quad \gamma_{124} = \alpha + \bar{\beta}. \tag{B8}$$

The intrinsic derivatives are defined according to the relations

$$D\phi = \phi;\mu l^\mu, \quad \Delta\phi = \phi;\mu n^\mu, \quad \delta\phi = \phi;\mu m^\mu, \quad \bar{\delta}\phi = \phi;\mu\bar{m}^\mu, \tag{B9}$$

and are satisfied by the commutation relation

$$\Delta D - D\Delta = (\gamma + \bar{\gamma})D + (\epsilon + \bar{\epsilon})\Delta - (\tau + \bar{\pi})\bar{\delta} - (\bar{\tau} + \pi)\delta, \tag{B10a}$$

$$\delta D - D\delta = (\bar{\alpha} + \beta - \bar{\pi})D + \kappa\Delta - \sigma\bar{\delta} - (\bar{\rho} + \epsilon - \bar{\epsilon})\delta, \tag{B10b}$$

$$\delta\Delta - \Delta\delta = -\bar{\nu}D + (\tau - \bar{\alpha} - \beta)\Delta + \bar{\lambda}\bar{\delta} + (\mu - \gamma + \bar{\gamma})\delta, \tag{B10c}$$

$$\bar{\delta}\delta - \delta\bar{\delta} = (\bar{\mu} - \mu)D + (\bar{\rho} - \rho)\Delta - (\bar{\alpha} - \beta)\bar{\delta} - (\bar{\beta} - \alpha)\delta. \tag{B10d}$$

Equation (A10) can be written in the form

$$C_{abcd;p}\zeta^p + C_{pbcd}\zeta^p;a + C_{apcd}\zeta^p;b + C_{abpd}\zeta^p;c + C_{abcp}\zeta^p;d = 2\phi C_{abcd} + \zeta^p[C_{rbcd}(\gamma_a{}^r{}_p + \gamma^r{}_{pa}) + C_{arcd}(\gamma_b{}^r{}_p + \gamma^r{}_{pb}) + C_{abrd}(\gamma_c{}^r{}_p + \gamma^r{}_{pc}) + C_{abcr}(\gamma_d{}^r{}_p + \gamma^r{}_{pd})]. \tag{B11}$$

The independent components of conformal curvature tensor may be expressed in terms of the five complex scalars ψ_0, $\psi_1,\ldots,\psi_4$. We have

$$C_{1212} = \psi_2 + \bar{\psi}_2, \quad C_{1213} = \psi_1, \quad C_{1223} = -\bar{\psi}_3, \quad C_{1234} = -\psi_2 + \bar{\psi}_2, \quad C_{1313} = \psi_0, \quad C_{1324} = -\psi_2, \quad C_{1334} = -\psi_1, \quad C_{2323} = \bar{\psi}_4, \quad C_{2334} = -\bar{\psi}_3, \quad C_{3434} = \psi_2 + \bar{\psi}_2, \quad C_{1314} = C_{1323} = C_{2324} = 0. \tag{B12}$$

Making use of Eqs. (B5a)–(B10d), and (B12), Eq. (B11) becomes

$$\zeta_1\Delta\psi_0 + \zeta_2 D\psi_0 - \zeta_3\bar{\delta}\psi_0 - \zeta_4\delta\psi_0 + 2\psi_0(D\zeta_2 - \delta\zeta_4) + 2\psi_1(\delta\zeta_1 - D\zeta_3) = 2\phi\psi_0 + 2\psi_0[(2\gamma - \mu)\zeta_1 + (\epsilon - \bar{\epsilon} + \bar{\rho})\zeta_2 + (\pi - 2\alpha)\zeta_3 + (\bar{\pi} - \beta - \bar{\alpha})\zeta_4] + 2\psi_1[(\beta + \bar{\alpha} - 2\tau - \bar{\pi})\zeta_1 - \kappa\zeta_2 + (\bar{\epsilon} - \epsilon + 2\rho - \bar{\rho})\zeta_3 + \sigma\zeta_4], \tag{B13}$$

$$\zeta_1\Delta\psi_1 + \zeta_2 D\psi_1 - \zeta_3\bar{\delta}\psi_1 - \zeta_4\delta\psi_1 - \psi_0\Delta\zeta_4 + \psi_1(D\zeta_2 - \delta\zeta_4) + \psi_2(\delta\zeta_1 - 2D\zeta_3) = \psi_0[(\pi + \bar{\tau})\zeta_2 - \lambda\zeta_3 + (\gamma - \bar{\gamma} - \mu)\zeta_4] + \psi_1[(2\gamma - \mu)\zeta_1 + (\epsilon - \bar{\epsilon} + \bar{\rho})\zeta_2 + (\pi - 2\alpha)\zeta_3 + (\bar{\pi} - \bar{\alpha} - \beta)\zeta_4] + \psi_2[(\bar{\alpha} + \beta - 2\bar{\pi} - 3\tau)\zeta_1 - \kappa\zeta_2 + (3\rho - \bar{\rho} - 2\epsilon + 2\bar{\epsilon})\zeta_3 + 2\sigma\zeta_4], \tag{B14}$$

$$\zeta_1\Delta\psi_2 + \zeta_2 D\psi_2 - \zeta_3\bar{\delta}\psi_2 - \zeta_4\delta\psi_2 + \psi_1(\bar{\delta}\zeta_2 - \Delta\zeta_4) + \psi_3(\delta\zeta_1 - D\zeta_3) = -2\phi\psi_2 + \psi_1[\nu\zeta_1 + (2\pi + \bar{\tau} - \alpha - \bar{\beta})\zeta_2 - \lambda\zeta_3 + (\gamma - \bar{\gamma} + \bar{\mu} - 2\mu)\zeta_4] + \psi_3[(\bar{\alpha} + \beta - 2\tau - \bar{\pi})\zeta_1 - \kappa\zeta_2 + (2\rho - \bar{\rho} - \epsilon + \bar{\epsilon})\zeta_3 + \sigma\zeta_4], \tag{B15}$$

$$\zeta_1\Delta\psi_3 + \zeta_2 D\psi_3 - \zeta_3\bar{\delta}\psi_3 - \zeta_4\delta\psi_3 - \psi_4 D\zeta_3 + \psi_3(\Delta\zeta_1 - \bar{\delta}\zeta_3) + \psi_2(\bar{\delta}\zeta_2 - 2\Delta\zeta_4) = \psi_2[\nu\zeta_1 + (3\pi + 2\bar{\tau} - \alpha - \bar{\beta})\zeta_2 - 2\lambda\zeta_3 + (2\gamma - 2\bar{\gamma} + \bar{\mu} - 3\mu)\zeta_4] + \psi_3[(\bar{\gamma} - \gamma - \bar{\mu})\zeta_1 + (\rho - 2\epsilon)\zeta_2 + (\alpha + \bar{\beta} - \bar{\tau})\zeta_3 + (2\beta - \tau)\zeta_4] + \psi_4[-(\tau + \bar{\pi})\zeta_1 + (\bar{\epsilon} - \epsilon + \rho)\zeta_3 + \sigma\zeta_4], \tag{B16}$$

$$\zeta_1\Delta\psi_4 + \zeta_2 D\psi_4 - \zeta_3\bar{\delta}\psi_4 - \zeta_4\delta\psi_4 + 2\psi_4(\Delta\zeta_1 - \bar{\delta}\zeta_3) + 2\psi_3(\bar{\delta}\zeta_2 - \Delta\zeta_4) = 2\phi\psi_4 + 2\psi_3[\nu\zeta_1 + (2\pi + \bar{\tau} - \alpha - \bar{\beta})\zeta_2 - \lambda\zeta_3 + (\gamma - \bar{\gamma} + \bar{\mu} - 2\mu)\zeta_4] + 2\psi_4[(\bar{\gamma} - \gamma - \bar{\mu})\zeta_1 + (\rho - 2\epsilon)\zeta_2 + (\alpha + \bar{\beta} - \bar{\tau})\zeta_3 + (2\beta - \tau)\zeta_4]. \tag{B17}$$

[1]E. Nöther, Nachr. Akad. Wiss. Goettingen II, Math. Phys. Kl. **1918**, 235.
[2]W. R. Davis and M. K. Moss, Nuovo Cimento **33**, 1558 (1965).
[3]G. H. Katzin and J. Levine, J. Math. Phys. **9**, 8 (1968).
[4]G. H. Katzin, J. Levine, and W. R. Davis, J. Math. Phys. **10**, 617 (1969).
[5]C. D. Collinson and C. D. Frech, J. Math. Phys. **8**, 701 (1967).
[6]M. E. Cahill and A. H. Taub, Commun. Math. Phys. **21**, 1 (1971).
[7]A. H. Taub, *General Relativity: Papers in Honour of J. L. Synge*, edited by L. O'Raifeartaigh (Oxford U. P., London, 1972), Chap. VIII, p. 133.
[8]B. B. Godfrey, Gen. Relativ. Gravit. **3**, 3 (1972).
[9]D. M. Eardley, Commun. Math. Phys. **37**, 287 (1974).
[10]C. W. Misner and J. A. Wheeler, Ann. Phys. (NY) **2**, 525 (1957).
[11]P. A. Goodinson and R. A. Newing, J. Int. Math. Appl. **6**, 212 (1970).
[12]P. A. Goodinson and R. A. Newing, J. Int. Math. Appl. **5**, 72 (1969).
[13]M. L. Woolley, Commun. Math. Phys. **31**, 75 (1973).
[14]M. L. Woolley, Commun. Math. Phys. **33**, 135 (1973).
[15]R. K. Sachs, Proc. R. Soc. London Ser. A **264**, 309 (1961).
[16]N. Tariq and B. O. J. Tupper, Gen. Relativ. Gravit. **6**, 345 (1975).
[17]N. Tariq and B. O. J. Tupper, Tensor (N.S.) **28**, 83 (1974).
[18]G. C. Debney and J. Zund, Tensor (N.S.) **22**, 333 (1971).
[19]G. C. Debney and J. Zund, Tensor (N.S.) **25**, 53 (1972).
[20]K. Yano, *The Theory of Lie Derivatives and its Applications* (North-Holland, Amsterdam, 1957).

Exact solutions for space-times with local rotational symmetry in which the Dirac equation separates

B. R. Iyer and C. V. Vishveshwara
Raman Research Institute, Bangalore 560080, India

(Received 5 February 1986; accepted for publication 18 February 1987)

The field equations for the class of perfect fluid space-times with local rotational symmetry in which the authors had earlier shown the Dirac equation separates are studied. For the vacuum and dust cases all possible solutions are exhibited. Other solutions correspond to radiation, a stiff fluid, and a fluid with negative pressure.

I. INTRODUCTION

In an earlier paper[1] (hereafter referred to as I) we investigated the problem of separability of the Dirac equation in perfect fluid space-times with local rotational symmetry and showed that separation was possible only in a certain subclass of the whole family. The geometrical properties of these space-times were also obtained but the question of the specific space-times in this subclass was left unanswered. In this paper we study the field equations for these particular space-times and attempt to isolate those exact solutions which fall in this category. For the vacuum ($p=\rho=0$) and dust ($p=0$) cases all the possible solutions are exhibited. Some exact solutions for other interesting sources like radiation ($p=\frac{1}{3}\rho$), a stiff fluid ($p=\rho$), and fluid with negative pressure ($p+\rho=0$) are also obtained. Though most of these solutions were known earlier we present a unified and systematic treatment of the different cases of particular interest as background metrics wherein our earlier separation of variables for the Dirac equation is applicable.

In the next section we set up the field equations for the relevant solutions. In Secs. III and IV we obtain all the vacuum and dust solutions, respectively. Section V contains some solutions corresponding to radiation, a stiff fluid, and a fluid with negative pressure.

II. SPACE-TIMES WITH LOCAL ROTATIONAL SYMMETRY WHEREIN THE DIRAC EQUATION SEPARATES

As demonstrated in I the space-times with local rotational symmetry in which the Dirac equation separates are of the following four types.

Case I:

$$ds^2=(1/F^2)dx^{0^2}-dx^{1^2}-Y^2(dx^{2^2}+t^2\,dx^{3^2}), \tag{1a}$$

where

$$F=F(x^1),\quad Y=Y(x^1). \tag{1b}$$

Case III:

$$ds^2=dx^{0^2}-X^2\,dx^{1^2}-Y^2(dx^{2^2}+t^2\,dx^{3^2}), \tag{2a}$$

with

$$X=X(x^0),\quad Y=Y(x^0). \tag{2b}$$

Case II a:

$$ds^2=(1/F^2)d\bar{x}^{0^2}-X^2\,d\bar{x}^{1^2}-Y^2(dx^{2^2}+t^2\,dx^{3^2}), \tag{3a}$$

where

$$F=F(\bar{x}^0),\quad X=X(\bar{x}^1),\quad Y=Y(\bar{x}^1). \tag{3b}$$

Case II b:

$$ds^2=(1/F^2)d\bar{x}^{0^2}-X^2\,d\bar{x}^{1^2}-Y^2(d\bar{x}^{2^2}+t^2\,dx^{3^2}), \tag{4a}$$

where

$$F=F(\bar{x}^0),\quad X=X(\bar{x}^1),\quad Y=Y(\bar{x}^0). \tag{4b}$$

In the above equations t is one of the four functions

$$\begin{aligned}&\text{(i) } t=\text{const},\qquad \text{(ii) } t=x^2,\\ &\text{(iii) } t=\sin(x^2),\quad \text{(iv) } t=\sinh(x^2).\end{aligned} \tag{5}$$

It is clear that the solutions corresponding to $t=$ const and $t=x^2$ are related trivially by transformations from Cartesian to cylindrical coordinates in the x^2-x^3 plane, i.e.,

$$x^{2'}=x^2\cos(x^3),\quad x^{3'}=x^2\sin(x^3). \tag{6}$$

Consequently, these two cases can be treated together. Further, in Cases II a and II b, by the following transformation of coordinates

$$x^0=\int\frac{d\bar{x}^0}{F(\bar{x}^0)},\quad x^1=\int X(\bar{x}^1)d\bar{x}^1, \tag{7}$$

the line elements become the following.

Case II a:

$$ds^2=dx^{0^2}-dx^{1^2}-Y^2(dx^{2^2}+t^2\,dx^{3^2}), \tag{8a}$$

where

$$Y=Y(x^1). \tag{8b}$$

Case II b:

$$ds^2=dx^{0^2}-dx^{1^2}-Y^2(dx^{2^2}+t^2\,dx^{3^2}), \tag{9a}$$

with

$$Y=Y(x^0). \tag{9b}$$

In this form Eq. (8) is a special case of (1) with $F=$ const while Eq. (9) is a special case of Eq. (2) with $X=$ const.

Choosing units $c=8\pi G=1$ and signature $(+,-,-,-)$ the field equations are

$$G_{ab}=T_{ab}, \tag{10a}$$

where for a perfect fluid

$$T_{ab}=(\rho+p)U_aU_b-pg_{ab}. \tag{10b}$$

Introducing ϵ such that

 0022-2488/87/061377-05$02.50

$$\epsilon = \begin{cases} 0, & \text{for } t = \text{const}, \\ +1, & \text{for } t = \sin(x^2), \\ -1, & \text{for } t = \sinh(x^2). \end{cases} \tag{11}$$

Equation (10) for Case I becomes

$$2\frac{Y_{,11}}{Y} + \frac{Y_{,1}^2}{Y^2} - \frac{\epsilon}{Y^2} = -\rho, \tag{12a}$$

$$2\frac{Y_{,1}F_{,1}}{YF} - \frac{Y_{,1}^2}{Y^2} + \frac{\epsilon}{Y^2} = -p, \tag{12b}$$

$$\frac{F_{,11}}{F} - \frac{Y_{,11}}{Y} - \frac{F_{,1}}{F}\left(2\frac{F_{,1}}{F} - \frac{Y_{,1}}{Y}\right) = -p. \tag{12c}$$

For Case III one obtains

$$2\frac{X_{,0}Y_{,0}}{XY} + \frac{Y_{,0}^2}{Y^2} + \frac{\epsilon}{Y^2} = \rho, \tag{13a}$$

$$2\frac{Y_{,00}}{Y} + \frac{Y_{,0}^2}{Y^2} + \frac{\epsilon}{Y^2} = -p, \tag{13b}$$

$$\frac{Y_{,00}}{Y} + \frac{X_{,0}Y_{,0}}{XY} + \frac{X_{,00}}{X} = -p. \tag{13c}$$

As mentioned earlier, Case II a corresponds to $F = \text{const}$ in Eqs. (12) while Case II b corresponds to $X = \text{const}$ in Eqs. (13).

The field equations should be supplemented by the equation of state for the perfect fluid which we prescribe to be of the form

$$p = (\gamma - 1)\rho. \tag{14}$$

The conservation equation for T^{ab} gives

$$T^{ab}{}_{;b} = 0. \tag{15}$$

For Case I Eq. (15) gives

$$\rho_{,0} = p_{,2} = p_{,3} = 0 \tag{16a}$$

and

$$pY^4F^{\gamma/(\gamma-1)} = \text{const}, \tag{16b}$$

while for Case III we have

$$\rho(XY^2)^{\gamma} = \text{const} = \rho_0, \tag{17a}$$

$$p_{,1} = p_{,2} = p_{,3} = 0. \tag{17b}$$

Though not useful for the vacuum and dust cases the above "first integrals" are useful in the other cases.

III. VACUUM SPACE-TIMES ($\gamma = 1; p = \rho = 0$)

Case I: If $F = \text{const}$, Eqs. (12) become

$$Y_{,1}^2 = \epsilon, \quad Y_{,11} = 0. \tag{18}$$

Thus for $\epsilon = 0$, one obtains a flat space-time in Cartesian coordinates while for $\epsilon = +1$ one finds $Y^2 = (x^1)^2$, which is a flat space-time in spherical polar coordinates. There is no solution for $\epsilon = -1$.

From Eqs. (12a) and (12b) $Y = \text{const}$ is possible only if $\epsilon = 0$. In this case Eq. (12c) gives

$$F_{,11}/F - 2F_{,1}^2/F^2 = 0, \tag{19}$$

which on integration yields

$$1/F^2 = (x^1)^2 \tag{20}$$

(here and in later parts all trivial integration constants are transformed away by a suitable translation or scaling). This is just a Minkowski space-time in Rindler coordinates

$$\bar{x}^0 = x^1 \sinh(x^0), \quad \bar{x}^1 = x^1 \cosh(x^0). \tag{21}$$

In general (i.e., if $F_{,1} \neq 0$, $Y_{,1} \neq 0$) by adding Eqs. (12a) and (12b) and integrating we obtain

$$FY_{,1} = C_1. \tag{22}$$

Equation (12a) can be rewritten as

$$2Y_{,11}/(Y_{,1}^2 - \epsilon) + 1/Y = 0, \tag{23}$$

which on integration gives

$$Y(Y_{,1}^2 - \epsilon) = C_2. \tag{24}$$

Solutions of (22) and (24) satisfy (12c) identically. Hence a solution of Eq. (24) yields a solution of the field equation. From Eq. (24) for $\epsilon = 0$, we obtain

$$Y^2 = (x^1)^{4/3}, \quad F^2 = (x^1)^{2/3}, \tag{25}$$

which is the plane symmetric Taub solution[2]

$$ds^2 = z^{-1/2}(dT^2 - dz^2) - z(dx^2 + dy^2); \quad z > 0, \tag{26a}$$

as follows by the transformations

$$T = (\tfrac{3}{4})^{1/3}x^0, \quad z = (\tfrac{3}{4})^{4/3}(x^1)^{4/3}, \qquad x = (\tfrac{3}{4})^{-2/3}x^2, \quad y = (\tfrac{3}{4})^{-2/3}x^3. \tag{26b}$$

For $\epsilon = 1$, the solution may be implicitly given as

$$Y = -(c_2/2)(1 + \cosh(2\rho)), \qquad F = \pm c_1 \coth\rho, \qquad \pm x^1 + c_3 = -(c_2/2)(\sinh(2\rho) + 2\rho). \tag{27}$$

On transforming to coordinates (x^0, ρ, x^2, x^3), one obtains

$$ds^2 = \tanh^2\rho\, dx^{0^2} - 4c_2^2 \cosh^4\rho\, d\rho^2 - c_2^2 \cosh^4\rho(dx^{2^2} + \sin^2 x^2\, dx^{3^2}). \tag{28}$$

This is just a Schwarzschild solution of mass $c_2/2$ as can be seen by transforming to coordinate r:

$$r = c_2 \cosh^2(\rho). \tag{29}$$

It is also one of the Levi-Civita degenerate static vacuum solution type AI (Ref. 3).

For $\epsilon = -1$ one obtains

$$Y = c_2 \sin^2(\rho), \quad F = \pm c_1 \tan(\rho), \qquad \pm x^1 + c_3 = -(c_2/2)(\sin(2\rho) - 2\rho), \tag{30}$$

which in coordinates (x^0, ρ, x^2, x^3) give

$$ds^2 = \cot^2\rho\, dx^{0^2} - c_2^2 \sin^4\rho(4\, d\rho^2 + dx^{2^2} + \sinh^2 x^2\, dx^{3^2}). \tag{31}$$

This is the degenerate static vacuum solution due to Levi-Civita[3] which in the classification of Ehlers and Kundt[3] is class AII. In terms of coordinates

$$z = c_2 \sin^2\rho, \tag{32}$$

$$ds^2 = (c_2/z - 1)dx^{0^2} - ((c_2/z) - 1)^{-1}\, dz^2 - z^2(dx^{2^2} + \sinh^2 x^2\, dx^{3^2}). \tag{33}$$

Case III: Let us now turn to Eqs. (13). If $X = \text{const}$ they become

$$Y_{,0}^2 = -\epsilon, \quad Y_{,00} = 0. \tag{34}$$

As for Eqs. (12) for $\epsilon = 0$ one has a flat space-time in Cartesian coordinates while for $\epsilon = -1$ one obtains $Y^2 = x^{0^2}$. This is just a Milne universe: a flat space-time in Rindler-like coordinates, as can be seen by the transformations

$$\begin{aligned} &\bar{x}^0 = x^0 \cosh x^2, \quad \bar{x}^1 = x^1, \\ &\bar{x}^2 = x^0 \sinh x^2 \cos x^3, \quad \bar{x}^3 = x^0 \sinh x^2 \sin x^3. \end{aligned} \tag{35}$$

From Eqs. (13a) and (13b) $Y = \text{const}$ is possible only if $\epsilon = 0$. In this case Eq. (13c) gives

$$X_{,00} = 0, \quad \text{i.e., } X = x^0. \tag{36}$$

This again is a flat space-time in Rindler-like coordinates

$$\bar{x}^0 = x^0 \cosh x^1, \quad \bar{x}^1 = x^0 \sinh x^1. \tag{37}$$

We now consider cases when neither X nor Y is constant. As before, taking the difference of Eqs. (13a) and (13b) and integrating we obtain

$$X = c_1 Y_{,0}. \tag{38}$$

Equation (13b) on integration gives

$$Y(Y_{,0}^2 + \epsilon) = c_2. \tag{39}$$

Solutions of Eqs. (38) and (39) satisfy Eq. (13c) identically. For $\epsilon = 0$ one obtains

$$Y = x^{0^{2/3}}, \quad X = (x^0)^{-1/3}, \tag{40}$$

which is a Kasner space-time with local rotational symmetry. The Dirac equation in this case is treated in more detail elsewhere.[4]

For $\epsilon = +1$ we obtain

$$\begin{aligned} &Y = c_2 \sin^2 T, \quad X = \pm c_1 \cot T, \\ &\pm x^0 + c_3 = (c_2/2)(2T - \sin 2T), \end{aligned} \tag{41}$$

which in terms of coordinates (T, x^1, x^2, x^3) gives

$$\begin{aligned} ds^2 = {} & 4c_2^2 \sin^4 T\, dT^2 - c_1^2 \cot^2 T\, dx^{1^2} \\ & - c_2^2 \sin^4 T(dx^{2^2} + \sin^2 x^2\, dx^{3^2}). \end{aligned} \tag{42}$$

Transforming to

$$\bar{T} = c_2 \sin^2 T, \; r = c_1 x^1 \tag{43}$$

yields the "inner" sector of the Schwarzschild solution, i.e. $(r < c_2)$

$$\begin{aligned} ds^2 = {} & (c_2/\bar{T} - 1)^{-1} d\bar{T}^2 - (c_2/\bar{T} - 1) dr^2 \\ & - \bar{T}^2(dx^{2^2} + \sin^2 x^2\, dx^{3^2}). \end{aligned} \tag{44}$$

For $\epsilon = -1$, on the other hand,

$$\begin{aligned} &Y = -c_2 \cosh^2 T, \quad X = \mp c_1 \tanh T, \\ &\pm x^0 + c_3 = (c_2/2)(\sinh 2T + 2T), \end{aligned} \tag{45}$$

which in terms of (T, x^1, x^2, x^3) yields

$$\begin{aligned} ds^2 = {} & 4c_2^2 \cosh^4 T\, dT^2 - c_1^2 \tanh^2 T\, dx^{1^2} \\ & - c_2^2 \cosh^4 T(dx^{2^2} + \sinh^2 x^2\, dx^{3^2}). \end{aligned} \tag{46}$$

Once again going over to

$$\bar{T} = c_2 \cosh^2 T, \; r = c_1 x^1, \tag{47}$$

we obtain

$$\begin{aligned} ds^2 = {} & (1 - c_2/\bar{T})^{-1} d\bar{T}^2 - (1 - c_2/\bar{T}) dr^2 \\ & - \bar{T}^2(dx^{2^2} + \sinh^2 x^2\, dx^{3^2}). \end{aligned} \tag{48}$$

This solution is to the Levi-Civita static solution AII (Ref. 3), the analog of the $R < 2M$ region of the Schwarzschild solution.

IV. THE DUST SOLUTIONS ($\gamma = 1, p = 0$)

In Eq. (12) corresponding to Case I for dust, if $F_{,1} = 0$, then Eqs. (12b) and (12c) become

$$Y_{,1}^2 - \epsilon = 0, \quad Y_{,11} = 0, \tag{49}$$

which when compared with Eq. (12a), gives $\rho = 0$. Thus no dust solutions are possible in this case. Similarly, for $Y_{,1}^2 = \epsilon$ no dust solutions exist.

In general, however, Eq. (12b) gives

$$F_{,1}/F = (Y_{,1}^2 - \epsilon)/2YY_{,1}. \tag{50}$$

Differentiating (50) and substituting in Eq. (12c) one obtains

$$2Y_{,11}/Y + (Y_{,1}^2 - \epsilon)/Y^2 = 0, \tag{51}$$

which employing (12a) gives $\rho = 0$. Thus no dust solution is possible for Eq. (12). They seem to be possible only in metrics of subclass III corresponding to Eq. (13).

If $X_{,0} = 0$, Eqs. (13) yield

$$\begin{aligned} &Y_{,0}^2 + \epsilon = \rho, \\ &Y_{,00}/Y + (Y_{,0}^2 + \epsilon)/Y^2 = 0, \\ &Y_{,00}/Y = 0, \end{aligned} \tag{52}$$

which are consistent only for $\rho = 0$. Thus one does not have dust solutions with $X = \text{const}$. From Eq. (13b) $Y = \text{const}$ solutions are only possible for $\epsilon = 0$, which from (13a) implies $\rho = 0$. Thus one does not have such dust solutions either. If $X_{,0} \neq 0$, $Y_{,0} \neq 0$, Eq. (13b) can be rewritten as

$$2Y_{,0}\, Y_{,00}/(Y_{,0}^2 + \epsilon) + Y_{,0}/Y = 0, \tag{53}$$

which gives

$$Y(Y_{,0}^2 + \epsilon) = \text{const}. \tag{54}$$

For $\epsilon = 0$, Eq. (54) is solved by

$$Y = (c_1 x^0 + c_2)^{2/3}. \tag{55}$$

Replacing Y in (13c) from Eq. (55) one obtains for X, the differential equation

$$X_{,TT} - \tfrac{1}{3} X_{,T} - \tfrac{2}{9} X = 0, \tag{56a}$$

where

$$T = \log(c_1 x^0 + c^2). \tag{56b}$$

Consequently, the general solutions for X is

$$X = [c_3(c_1 x^0 + c_2) + c_4]/(c_1 x^0 + c_2)^{1/3}. \tag{57}$$

Substituting for X and Y from Eqs. (55) and (57) in Eq. (13a) yields ρ:

$$\rho = \tfrac{4}{3} c_3 c_1^2/(c_1 x^0 + c_2)[c_3(c_1 x^0 + c_2) + c_4]. \tag{58}$$

By a simple translation and scaling, the metric becomes

$$\begin{aligned} ds^2 = {} & dx^{0^2} - ((x^0 + \alpha)/x^{0^{1/3}})^2\, dx^{1^2} \\ & - x^{0^{4/3}}(dx^{2^2} + dx^{3^2}), \end{aligned} \tag{59a}$$

where

$$\alpha = c_4/c_3 \text{ and } \rho = \tfrac{4}{3}(x^{0^2} + \alpha x^0)^{-1}. \tag{59b}$$

For the special choice of $\alpha = 0$, Eqs. (59) yield

$$ds^2 = dx^{0^2} - x^{0^{4/3}}(dx^{1^2} + dx^{2^2} + dx^{3^2}), \tag{60a}$$

$$\rho = 4/(3x^{0^2}). \tag{60b}$$

This is the Einstein–de Sitter solution for dust which has homogeneous and isotropic spatial sections. However, if $\alpha \neq 0$ we obtain a more general solution which does not seem obviously equivalent to the $\alpha = 0$ case. For $\epsilon = 1$, the solution may be written in the implicit form

$$\begin{aligned} &Y = c_1 \sin^2 T, \\ &\pm x^0 + c_2 = (c_1/2)[2T - \sin 2T]. \end{aligned} \tag{61}$$

Substituting (61) in Eq. (13c) then gives

$$X_{,TT}/X - 2/\sin^2 T = 0. \tag{62}$$

By inspection $X = \cot(T)$ is a solution to the above equation. To find the other solution let

$$X = V\cot(T) \tag{63}$$

in Eq. (62) so that V satisfies

$$V_{,TT}/V_{,T} = 2\csc^2 T/\cot T. \tag{64}$$

The above equation is integrated and finally one has

$$X = c_3(1 - T\cot T) + c_4 \cot T. \tag{65}$$

Substituting (61) and (65) in Eq. (12a) we have

$$\rho = c_3/c_1^2 \sin^4 T[c_3 - \cot T(c_3 T - c_4)]. \tag{66}$$

In terms of (T,x^1,x^2,x^3) one has

$$\begin{aligned} ds^2 = {} & 4\sin^4 T\, dT^2 - [1 - \cot T(T - c)]^2\, dx^{1^2} \\ & - \sin^4 T(dx^{2^2} + \sin^2 x^2\, dx^{3^2}), \end{aligned} \tag{67a}$$

$$\rho = 1/\sin^4 T[1 - \cot T(T - C)]. \tag{67b}$$

Similarly, for $\epsilon = -1$,

$$\begin{aligned} &Y = c_1 \sinh^2 T, \\ &\pm x^0 + c_2 = (c_1/2)(\sinh 2T - 2T). \end{aligned} \tag{68}$$

Substituting into Eq. (13c) gives

$$X_{,TT}/X - 2/\sinh^2 T = 0. \tag{69}$$

As before, since $X = \coth T$ is a solution of (69) we write

$$X = V\coth T, \tag{70a}$$

V is then a solution of

$$V_{,TT}/V_{,T} = (2\,\mathrm{csch}^2 T)/(\coth T), \tag{70b}$$

and consequently

$$X = c_3(T\coth T - 1) + c_4 \coth T. \tag{71}$$

For this case

$$\rho = c_3/c_1^2 \sinh^4 T[c_3(T\coth T - 1) + c_4 \coth T]. \tag{72}$$

In terms of (T,x^1,x^2,x^3) we thus have

$$\begin{aligned} ds^2 = {} & 4\sinh^4 T\, dT^2 - [(T + \beta)\coth T - 1]^2\, dx^{1^2} \\ & - \sinh^4 T(dx^{2^2} + \sinh^2 x^2\, dx^{3^2}), \end{aligned} \tag{73a}$$

$$\rho = 1/\sinh^4 T[(T + \beta)\coth T - 1]. \tag{73b}$$

The dust solutions given by Eqs. (67) and (73) for $\epsilon = \pm 1$ are those obtained by Kantowski and Sachs.[5]

V. OTHER SOLUTIONS

If $F = \text{const}$, adding twice Eq. (12c) to (12b) one finds that the equations are consistent with (12a) only if $\rho + 3p = 0$. This case is not of physical interest. Similarly if $X = \text{const}$ adding two times (13c) to (13a) one finds that there is consistency with (13b) only for $\rho = p$, i.e., $\gamma = 2$. In this case, choosing $X = 1$ we obtain from Eq. (17)

$$\rho = c_1^2/4Y^4. \tag{74}$$

Substituting in Eq. (13a) and integrating we get

$$x^0 = \int \frac{2Y\,dY}{\sqrt{c_1^2 - 4\epsilon Y^2}}. \tag{75}$$

For $\epsilon = 0$, the solution after suitable scalings give

$$ds^2 = dx^{0^2} - dx^{1^2} - x^0(dx^{2^2} + x^{2^2}\, dx^{3^2}), \tag{76a}$$

$$\rho = c_1^2/4x^{0^2}. \tag{76b}$$

For $\epsilon = 1$, similarly,

$$ds^2 = dx^{0^2} - dx^{1^2} - (c_1^2/4 - x^{0^2})(dx^{2^2} + \sin^2 x^2\, dx^{3^2}), \tag{77a}$$

$$\rho = (c_1^2/4)(c_1^2/4 - x^{0^2})^{-2}, \tag{77b}$$

whereas for $\epsilon = -1$,

$$\begin{aligned} ds^2 = {} & dx^{0^2} - dx^{1^2} - (x^{0^2} - c_1^2/4)(dx^{2^2} \\ & + \sinh^2 x^2\, dx^{3^2}), \end{aligned} \tag{78a}$$

$$\rho = c_1^2/4/(x^{0^2} - c_1^2/4)^2. \tag{78b}$$

The solutions given by Eqs. (76), (77), (78), for a $\gamma = 2$ fluid is to our knowledge new.

Let us consider Eq. (13) for $\epsilon = 0$. Adding $(\gamma - 1)$ times Eq. (13a) to Eq. (13b) and integrating one gets

$$X = c_1(Y_{,0}^2 Y^\gamma)^{1/2(\gamma - 1)}. \tag{79}$$

Since

$$\rho = c(XY^2)^{-\gamma}, \tag{80}$$

one thus gets

$$\rho = c_\gamma = (Y_{,0})^{\gamma/(\gamma - 1)} Y^{\gamma(4 - 3\gamma)/2(\gamma - 1)}, \tag{81a}$$

where

$$c_\gamma = c/c_1^\gamma. \tag{81b}$$

Equation (13b) thus becomes

$$\begin{aligned} &2Y_{,00}/Y + Y_{,0}^2/Y^2 \\ &\quad = -(\gamma - 1)c_\gamma Y_{,0}^{\gamma/(\gamma - 1)} Y^{\gamma(4 - 3\gamma)/2(\gamma - 1)}. \end{aligned} \tag{82}$$

The above equation will now be solved for the following interesting physical cases.

(a) $\gamma = 2$ $(p = \rho)$.

For this value the right-hand side of Eq. (82) is proportional to $Y_{,0}^2/Y^2$. Thus integrating (82) yields

$$Y = (c_2 x^0 + c_3)^{1/(1 + \alpha)}, \tag{83}$$

where

$$2\alpha = 1 + c/c_1^2.$$

Then Eq. (79) gives X as

$$X = (c_1/c_2)(1+\alpha)(c_2x^0 + c_3)^{(\alpha-1)/(\alpha+1)}. \tag{84}$$

After the usual scalings one thus has

$$ds^2 = dx^{0^2} - (x^0)^{2(\alpha-1)/(\alpha+1)}\, dx^{1^2} - (x^0)^{2/(1+\alpha)}(dx^{2^2} + dx^{3^2}), \tag{85a}$$

$$\rho = c/c_1^2(1+\alpha)^2 x^{0^2}. \tag{85b}$$

This solution is identical to one of the solutions in Vajk and Eltgroth.[6]

(b) $\gamma = \frac{4}{3}$ $(p = \frac{1}{3}\rho)$.

In this case ρ is proportional to $Y_{,0}^4$ and hence Eq. (82) becomes

$$2Y_{,00}/Y + Y_{,0}^2/Y^2 = -\beta Y_{,0}^4, \tag{86a}$$

where

$$\beta = \tfrac{1}{3}cc_1^{-4/3}. \tag{86b}$$

Substituting $YY_{,0}^2 = u$ into the above equation and integrating one obtains

$$u = YY_{,0}^2 = (c_2 + \beta Y)^{-1}, \tag{87}$$

whose solution may be written as

$$Y = (c_2/\beta)\sinh^2 T,$$
$$\pm x^0 + c_3 = (c_2^2/16\sqrt{\beta^3})[\sinh 4T - 4T], \tag{88}$$
$$X = c_1c_2\sqrt{\beta}\,(\cosh^3 T)/(\sinh T).$$

In terms of (T,x^1,x^2,x^3) the space-time is described by

$$ds^2 = (c_2^4/4\beta^3)\sinh^4 2T\, dT^2 - \cosh^4 T \coth^2 T\, dx^{1^2} - \sinh^4 T(dx^{2^2} + dx^{3^2}), \tag{89a}$$

$$\rho = 16\beta^2 c_\gamma/c_2^4 \sinh^4 2T. \tag{89b}$$

Like the earlier case, this is also a particular solution from Vajk and Eltgroth.[6]

(c) $\gamma = 0$ $(p + \rho = 0)$.

For this value of γ, $\rho = \text{const} = \rho_0$ and $X = c_1Y_{,0}$. Thus Eq. (82) yields

$$2Y_{,00}/Y + Y_{,0}^2/Y^2 = \rho_0. \tag{90}$$

The above equation can be integrated by letting $Y_{,0}/Y = u$. We get

$$Y = c_3[\cosh[\sqrt{3\rho_0}(x^0 + c_2)/2]]^{2/3},$$
$$X = c_1c_3\sqrt{\rho_0/3}\,[\cosh[\sqrt{3\rho_0}(x^0 + c_2)/2]]^{-1/3} \times (\sinh[\sqrt{3\rho_0}(x^0 + c_2)/2]). \tag{91}$$

Thus the metric may be written as

$$ds^2 = dx^{0^2} - (\cosh(\sqrt{3\rho_0}x^0/2))^{-2/3}\sinh^2(\sqrt{3\rho_0}x^0/2)dx^{1^2} - (\cosh(\sqrt{3\rho_0}x^0/2))^{4/3}(dx^{2^2} + dx^{3^2}). \tag{92}$$

To the best of our knowledge Eq. (92) is a new solution. For the various values of γ dealt with above we have not been able to obtain solutions of Eq. (13) for $\epsilon = \pm 1$ or of Eq. (12) for $\epsilon = 0, +1$.

In the foregoing we have systematically obtained the various exact solutions with local rotational symmetry in which the Dirac equation is separable. As was mentioned at the outset many of them turn out to be already known solutions sometimes in terms of unconventional coordinates. Other solutions, given by Eqs. (59), (76), (77), (78), and (92), are new as far as we know. Our results, while incorporating a regular classification of these space-times would also facilitate the study of the Dirac equation in backgrounds exhibiting local rotational symmetry.

ACKNOWLEDGMENT

It is a pleasure to thank Professor M. A. H. MacCallum for making available an algebraic computational program with the help of which we have checked our calculations.

[1] B. R. Iyer and C. V. Vishveshwara, J. Math. Phys. **26**, 1034 (1985).
[2] D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Cambridge U. P., Cambridge, 1980), p. 159.
[3] Reference 2, p. 188.
[4] B. R. Iyer, Phys. Lett. A **112**, 313 (1985).
[5] R. Kantowski and R. K. Sachs, J. Math. Phys. **7**, 443 (1966).
[6] J. P. Vajk and P. G. Eltgroth, J. Math. Phys. **11**, 2212 (1970); see, also, Ref. 2, p. 146.

Comment on the two "new" classes of Bianchi type II solutions

Dieter Lorenz-Petzold
Fakultät für Physik, Universität Konstanz, D-7750 Konstanz, Federal Republic of Germany

(Received 24 June 1986; accepted for publication 11 February 1987)

Hajj-Boutros's claim [J. Math. Phys. **27**, 1592 (1986)] that two new classes of Bianchi type II solutions can be generated from Lorenz's solution [Phys. Lett. A **79**, 19 (1980)] is shown to be wrong.

In a recent paper by Hajj-Boutros[1] the locally rotationally symmetric (LRS) Bianchi type II stiff matter solution derived by Lorenz[2] (and independently by Ruban[3]) has been reconsidered. Our solution is the unique stiff matter solution of Eqs. (2.6)–(2.8).[1] This solution includes the Taub[4] vacuum solution as a special case. It is an easy matter of calculation to derive the corresponding non-LRS solution.[5] The LRS case is given by Eqs. (3.1)–(3.5).[1]

We make the following comments. The crucial equation in the paper of Hajj-Boutros[1] is given by Eq. (2.9). By taking $R = R(t)$, $S = S(t)$, and $r = \dot{R}/R$ (or $\dot{S}/S$) this equation can be considered as a Riccati equation in r. Hajj-Boutros finds that from a known solution $r_0 = \dot{R}_0/R_0$ ($\dot{S}_0/S$) some new solutions (of Bianchi type II) are given by Eqs. (2.14)–(2.17). However, this idea is entirely wrong. First of all the Bianchi type II stiff matter solution $R = R(\tau)$, $S = S(\tau)$ [see Eqs. (3.1)–(3.4)], where the temporal variable τ is related to t by the relation $dt = SR^2\, d\tau$, is not a "particular" solution of the field equations (2.4)–(2.8): it is the most general (LRS) solution of Eq. (2.9)! Equation (2.9) can be rewritten in the simple form

$$(R'/R + S'/S)' = 0\,, \tag{1}$$

where $(\ )' = d(\ )/d\tau$. The most general solution is given by

$$(SR)^2 = \exp 2(q\tau + \phi), \quad q, \phi = \text{const.} \tag{2}$$

Moreover, our solution is given by $R = R(\tau)$ and $S = S(\tau)$ and not by $R = R(t)$ and $S = S(t)$. Thus Eqs. (3.6) and (3.7) of Ref. 1 are meaningless! This is the main error made by Hajj-Boutros. By using $dt = SR^2\, d\tau$ the solutions given can be reexpressed (at least in principle) in t time.

For the sake of completeness we also mention the recent critical remarks of MacCallum[6] concerning various "new" and incorrect Bianchi type II solutions.

[1] J. Hajj-Boutros, J. Math. Phys. **27**, 1592 (1986).
[2] D. Lorenz, Phys. Lett. A **79**, 19 (1980).
[3] V. A. Ruban, Preprint 412, Leningrad Institut of Nuclear Physics B. P. Konstantinova, 1978.
[4] A. H. Taub, Ann. Math. **53**, 472 (1951).
[5] D. Lorenz, Phys. Lett. A **80**, 235 (1980).
[6] M. A. H. MacCallum, Gen. Relativ. Gravit. **17**, 659 (1985).

 0022-2488/87/061382-01$02.50

Nonexistence of static conformally flat solutions in a new scalar–tensor theory

Tarkeshwar Singh
Department of Mathematics, Shree Ramdeobaba Kamala Nehru Engineering College, Gittikhadan, Katol Road, Nagpur 440013, India

(Received 28 October 1986; accepted for publication 11 February 1987)

For the special case when the scalar field is massless and conformally invariant, it is shown that there do not exist spherically symmetric conformally flat solutions in a new scalar–tensor theory proposed by Schmidt *et al.* [Phys. Rev. D **24**, 1484 (1981)] representing disordered radiation and in the presence of a source-free electromagnetic field except for the trivial empty flat space-time of Einstein's theory. The solution in the vacuum case is also only a flat space-time of Einstein's theory.

I. INTRODUCTION

Schmidt *et al.*[1] proposed a new scalar–tensor theory of gravitation, where the gravitational constant depends on a scalar field which itself couples to the surrounding masses through the curvature scalar. The idea was to obtain a possible stable configuration as a final situation in the history of a collapsing object due to the generation of a strong scalar field. The result was contrary to what was expected. In the new theory a mass term was added and an arbitrary coupling constant β between the scalar field ϕ and the curvature invariant R was also allowed. The theory was subsequently applied to a Friedmann–Robertson–Walker universe by Banerjee and Santos.[2] Further, Singh and Singh[3] have shown that the spatially homogeneous stationary perfect fluid cosmological model in this theory cannot include the radiation-filled universe or the empty universe at the limit in the presence of a massive scalar field. Banerjee *et al.*[4] have discussed a stiff fluid Bianchi type I cosmological model in this theory by considering the cosmological constant Λ and the mass term both being equal to zero. Finally, they have considered some special cases for $\beta \geqslant 1$ and have shown that solutions for matter-free space include the one previously found by Accioly *et al.*[5] for the conformally invariant scalar field ($\beta = 1$). Very recently, the author[6] has shown that an analog of the Birkhoff theorem in general relativity exists in this new scalar–tensor theory for the special case when the scalar field is massless and independent of time.

In the present case we apply this general theory to the static spherically symmetric conformally flat space-time for the special case when the scalar field ϕ is massless and conformally invariant ($\beta = 1$) (see Refs. 7 and 8).

II. FIELD EQUATIONS AND THEIR SOLUTIONS

The gravitational field equations in the scalar–tensor theory proposed by Schmidt *et al.*[1] are given by

$$(\gamma - (\beta/12)\phi^2)G_{ij} = -\tfrac{1}{2}T_{ij} - \tfrac{1}{2}[\phi_{,i}\phi_{,j} - \tfrac{1}{2}g_{ij}(\phi_{,k}\phi^{,k} - \mu^2\phi^2)] + (\beta/12)[(\phi^2)_{;ij} - g_{ij}(\phi^2)^{;k}_{;k}], \quad (1)$$

and the wave equation is

$$\Box\phi + [\mu^2 + (\beta/6)R\,]\phi = 0. \quad (2)$$

Here μ is the mass of the scalar field, β is an arbitrary coupling constant, and $\gamma = c^2/16\pi G$ is half of the inverse gravitational constant. The effective inverse gravitational coupling in this theory becomes

$$\gamma_{\text{eff}} = \gamma - (\beta/12)\phi^2$$

and the effective mass of the scalar field is now

$$\mu_{\text{eff}} = [\mu^2 + (\beta/6)R\,]^{1/2}.$$

We consider the static spherically symmetric conformally flat metric in the form

$$ds^2 = e^{\alpha}(-dr^2 - r^2\,d\theta^2 - r^2\sin^2\theta\,d\Phi^2 + dt^2), \quad (3)$$

where α is the function of r alone. We assume that the scalar field ϕ also has spherical symmetry.

Taking ϕ as a function of r only and using (3) in (1) and (2), the explicit field equations for the new scalar–tensor theory can be written as

$$\tfrac{1}{2}e^{\alpha}T^1_1 = -(\gamma - (\beta/12)\phi^2)(3\alpha'^2/4 + 2\alpha'/r) + \tfrac{1}{4}(\phi')^2 - \tfrac{1}{4}e^{\alpha}\mu^2\phi^2 + (\beta/12)[(2/r)(\phi^2)' + 3\alpha'(\phi^2)'/2], \quad (4)$$

$$\tfrac{1}{2}e^{\alpha}T^2_2 = -(\gamma - (\beta/12)\phi^2)(\alpha'' + \alpha'^2/4 + \alpha'/r) - \tfrac{1}{4}(\phi')^2 - \tfrac{1}{4}\mu^2\phi^2e^{\alpha} + (\beta/12)[(\phi^2)'' + (\phi^2)'(\alpha' + 1/r)] = \tfrac{1}{2}e^{\alpha}T^3_3, \quad (5)$$

$$\tfrac{1}{2}e^{\alpha}T^4_4 = -(\gamma - (\beta/12)\phi^2)(\alpha'' + \alpha'^2/4 + 2\alpha'/r) - \tfrac{1}{4}(\phi')^2 - \tfrac{1}{4}\mu^2\phi^2e^{\alpha} + (\beta/12)[(\phi^2)'' + 2(\phi^2)'/r + (\phi^2)'\alpha'/2], \quad (6)$$

$$e^{-\alpha}[(\phi^2)'' + (\phi^2)'(\alpha' + 2/r)] = \frac{(\beta/6)\phi}{\gamma + \beta(\beta-1)\phi^2/12}\left\{\frac{6}{\beta}\mu^2\gamma + \frac{1}{2}T + \frac{1}{2}\mu^2\phi^2 - \frac{(1-\beta)}{2}e^{-\alpha}(\phi')^2\right\}. \quad (7)$$

Here a prime indicates differentiation with respect to r.

III. SOLUTIONS OF THE FIELD EQUATIONS

As the field equations are nonlinear, the problem becomes difficult, in the general case with the nonvanishing

 0022-2488/87/061383-02$02.50

mass term μ. So the solutions in a special case for a massless conformally invariant scalar field ($\mu = 0$, $\beta = 1$) are presented. The question of overdeterminacy is settled by satisfying all the field equations by the actual substitution of the solution obtained. Before proceeding further we simply assume $\mu = 0$ and $\beta = 1$.

A. Vacuum solutions

It can be easily verified that when the scalar field ϕ is a constant and $T_{ij} = 0$, the field equations (4)–(7) yield a solution which describes an empty flat space-time of Einstein's theory, and when $T_{ij} = 0$, the field equations (4)–(7) reduce to the vacuum case.

From Eqs. (5)–(7) we can easily obtain

$$e^{\alpha} = k/(12\gamma - \phi^2), \tag{8a}$$

$$\phi^2 = 12\gamma - \exp(k_1/kr + k_2), \tag{8b}$$

where k, k_1, and k_2 are integration constants. Solution (8) satisfies Eq. (4) only when $k_1 = 0$. This, in view of (8), gives

$$\phi^2 = \text{const}, \quad \alpha = \text{const}.$$

Thus the solution of Eqs. (4)–(7) is

$$\alpha = \text{const}, \quad \phi = \text{const}. \tag{9}$$

Hence the only spherically symmetric static conformally flat solution of the new scalar–tensor theory[1] is simply the empty flat space-time of Einstein's theory, when the scalar field is massless and conformally invariant.

B. Electrovac solution

Here we consider the energy momentum tensor for a trace-free electromagnetic field in the form

$$T_{ij} = F_{ij}F^{i}_{\ j} - \tfrac{1}{4}g_{ij}F_{\prime\prime\prime}F^{\prime\prime\prime}, \tag{10}$$

where F_{ij} is the electromagnetic field tensor satisfying

$$F^{ij}_{\ ;j} = 0 \tag{11}$$

and

$$F_{ij,k} + F_{jk,i} + F_{ki,j} = 0. \tag{12}$$

For a static charged particle the only nonzero component of the electromagnetic field F_{ij} is F_{14}. Equations (11) and (12) now lead to $F_{14} = q/r^2$, where q is a constant which can be identified with the electric charge of the particle.

With metric (3) the nonvanishing components of the energy momentum are

$$-T^1_1 = T^2_2 = T^3_3 = -T^4_4 = -\tfrac{1}{2}(q^2/r^4)e^{-2\alpha}. \tag{13}$$

Using (13), the field equations (5)–(7) admit the solution

$$e^{\alpha} = (c_1r^2 + c_3)/(\phi^2 - 12\gamma)r^2, \tag{14a}$$

$$\phi^2 = 12\gamma - \exp\left\{\frac{c_2}{(-c_1c_3)^{1/2}}\tanh^{-1}\frac{r(-c_1c_2)^{1/2}}{c_3} + \phi_0\right\}, \tag{14b}$$

where ϕ_0, c_1, and c_2 are constants of integration and c_3 is set equal to $-3q^2$. On actual verification it was found that solution (14) satisfies each of the field equations only when $c_2 = 0$ and $c_3 = 0$, which, in turn reduces the field equations (4)–(7) to Einstein's vacuum case.

Hence the only spherically symmetric static conformally flat solution of the new scalar–tensor theory in the presence of a source-free electromagnetic field is the empty flat space-time of Einstein's theory.

C. Disordered radiation

Here we consider the energy momentum tensor due to that of a perfect fluid distribution in the form

$$T_{ij} = (\rho + p)u_iu_j - pg_{ij}, \tag{15}$$

with equation of state

$$\rho = 3p, \tag{16}$$

where ρ is energy density and p is the pressure of the fluid. From (3), (15), and (16) the components of T^i_j in comoving coordinates are

$$T^i_j = \text{diag}(-p, -p, -p, 3p). \tag{17}$$

The conservation equation $T^i_{j;j} = 0$ leads to

$$\frac{dp}{dr} + (\rho + p)\frac{\alpha'}{2} = 0. \tag{18}$$

Using Eqs. (15)–(17) in field equations (4)–(7), one gets the field equations of a new scalar–tensor theory with disordered radiation.

Now using (8) in the difference of Eqs. (5) and (6) and using (17) in the sum of Eqs. (5) and (6) one easily gets $p = 0$ and therefore $\rho = 0$. This leads to the vacuum field equations in which case, as shown in Sec. III A, the only solution is the flat space-time of Einstein's theory.

Thus there are no spherically symmetric conformally flat solutions of the new scalar–tensor theory[1] representing disordered radiation in the presence of massless conformally invariant scalar field.

IV. CONCLUSION

In order to understand fully the scalar–tensor theories of gravitation it is useful to have a knowledge of some exact solutions of these equations. The search for an analytic solution is important due to the fact that once such a solution is obtained one can study all of its physical properties. Exact static spherically symmetric conformally flat solutions in vacuum, in the presence of an electromagnetic field and for disordered radiation, are considered in a new scalar–tensor theory proposed by Schmidt *et al.*[1] for the special case when the scalar field is massless and conformally invariant. It is observed that the only spherically symmetric conformally flat solution in this new scalar–tensor theory is the flat space-time of Einstein's theory.

ACKNOWLEDGMENTS

The author is thankful to Professor A. A. Kayande for his encouragement. The author is also thankful to Dr. T. Singh, Department of Applied Mathematics, Banaras Hindu University, Varanasi for his valuable suggestions.

[1] G. Schmidt, W. Greiner, U. Heinz, and B. Müller, Phys. Rev. D **24**, 1484 (1981).
[2] A. Banerjee and N. O. Santos, Phys. Rev. D **26**, 3747 (1982).
[3] T. Singh and T. Singh, Indian J. Phys. **58** B, 41 (1984).
[4] A. Banerjee, A. K. G. De Oliveira, and N. O. Santo, Gen. Relativ. Gravit. **17**, 371 (1985).
[5] A. J. Accioly, A. N. Vaidya, and M. M. Som, Phys. Rev. D **27**, 2282 (1983).
[6] T. Singh, Phys. Rev. D **34**, 646 (1986).
[7] F. Gürsey, Ann. Phys. (NY) **24**, 211 (1963).
[8] R. Penrose, Proc. R. Soc. London Ser. A **284**, 159 (1965).

Series representations for calculations in quantum statistics. II

William A. Barker
Department of Physics, Santa Clara University, Santa Clara, California 95053

(Received 12 August 1986; accepted for publication 30 January 1987)

The methods developed in the first paper of this series (I) [J. Math. Phys. **27**, 302 (1986)] are used to calculate the pressure in several cases of physical interest: conduction electrons inside and outside a heated tungsten cathode; helium atoms in liquid helium II and He vapor; electrons in a white dwarf star, and photons in a cavity. The latter two cases involve ultrarelativistic particles whose energy $\epsilon = pc$. The formulation in I is extended to include particle speeds comparable to c. The grand potential is used to establish two useful, exact relationships between pressure and energy density: $P = 2u/3$ and $P = u/3$, valid for $v \ll c$ and $v \simeq c$, respectively.

I. INTRODUCTION

In an earlier paper,[1] hereafter referred to as I, the three distribution functions of statistical physics are treated in a similar manner. In each case a single integral suffices. The quantity $C\epsilon^p$ is averaged over the Maxwell–Boltzmann (MB), Bose–Einstein (BE), and Fermi–Dirac (FD) distributions. Here C is a constant and ϵ is the particle energy raised to any power p. Reference 1 develops simple schemes for evaluating these integrals. The MB integral [Eq. (2)] is written in closed form in terms of the gamma function of $p+1$. The BE integral [Eq. (8)] is the product of the gamma function and the Riemann zeta function of $p+1$. The FD integral [Eq. (27)] is written as a rapidly converging series of derivatives for any value of p. In the literature, some calculations, using the quantum statistics of highly degenerate gases, are quite formidable. Now it is really quite easy. The physical properties of various systems of interest can be investigated with very little mathematical effort.

The purpose of this paper is to illustrate this point for seven cases of physical interest. We focus on a single thermodynamic coordinate, the pressure, and we calculate this quantity for the following: (1) the conduction electrons inside a heated tungsten cathode, (2) the conduction electrons in the evacuated region between a heated tungsten cathode and an anode, (3) the helium atoms in liquid helium II at 1 K, (4) the helium atoms in helium vapor at 1 K,(5) the ultrarelativistic electrons in a white dwarf star, (6) a photon gas in a cavity, and (7) a laser beam.

II. GENERAL RELATIONSHIPS BETWEEN PRESSURE AND ENERGY. THE GRAND POTENTIAL

The grand potential[2]

$$\Omega = \sum_k \Omega_k = \mp kT \sum_k \ln[1 \pm e^{(\mu - \epsilon_k)/kT}] \tag{1}$$

is applicable in determining various thermodynamic properties of a Fermi, Bose, or classical gas. Here Ω_k is the grand potential for the energy ϵ_k. The familiar distribution function of a particle in the k th state is obtained by

$$n_k = -\frac{\partial \Omega_k}{\partial \mu} = \frac{1}{e^{(\epsilon_k - \mu)/kT} \pm 1}. \tag{2}$$

If μ/kT is large and negative, as it is in the classical case,

$$n_k = 1/e^{(\epsilon_k - \mu)/kT}. \tag{3}$$

Formally the energy U for each of these three cases, in the nonrelativistic limit, is

$$U = C \int_0^\infty \frac{\epsilon^{3/2}\, d\epsilon}{e^{(\epsilon - \mu)/kt} \pm 1{,}0}, \tag{4}$$

where $C = (2S+1)(V/4\pi^2)(2m/\hbar^2)^{3/2}$. Here S = spin, V = volume, m = particle mass, and $\hbar$ = Planck's constant divided by 2π.

Now convert (1) from a sum to an integral, integrate by parts, and compare with (4). We find that

$$\Omega = -2U/3. \tag{5}$$

Using[3] $\Omega = -PV$, it follows that

$$PV = 2U/3 \tag{6}$$

for nonrelativistic fermions, bosons, and classical particles. If, on the other hand, the particles are ultrarelativistic, then $\epsilon = pc$ replaces $\epsilon = p^2/2m$. The constant C in Eqs. (2), (6), and (16) of I should be replaced by

$$C' = (2S+1)V/2\pi^2\hbar^3c^3. \tag{7}$$

This follows from the changes in the volume element $4\pi p^2\, dp$ in momentum space. There are corresponding changes in the integrands which are taken care of by the proper choice for the power p. The expression for the grand potential now becomes

$$\Omega = -\frac{C'}{3} \int_0^\infty \frac{\epsilon^3\, d\epsilon}{e^{(\epsilon - \mu)/kT} \pm 1{,}0} = -\frac{U}{3}. \tag{8}$$

Therefore,

$$PV = U/3 \tag{9}$$

for ultrarelativistic particles.

III. THE PRESSURE OF A GAS OF CONDUCTION ELECTRONS INSIDE AND OUTSIDE A HEATED TUNGSTEN CATHODE. $S = \frac{1}{2}$

Case 1 (Inside): The conduction electrons collide frequently with phonons. Consequently their ambient speed is very small indeed compared to c.

 0022-2488/87/061385-04$02.50

It is easy to show that the conduction electrons inside a tungsten cathode constitute a degenerate Fermi gas: Compare n_Q, the quantum concentration,[4] with n, the actual concentration of electrons in tungsten,

$$n_Q = (mkT/2\pi\hbar^2)^{3/2} = 2.16\times 10^{20}/\text{cm}^3. \tag{10}$$

Here m is the electron mass, k is Boltzmann's constant, and T is the temperature of the heated cathode, taken here to be 2000 K. On the other hand

$$n = fN_0\rho/M = 1.27\times 10^{23}/\text{cm}^3, \tag{11}$$

with f = the number of free electrons per atom = 2, ρ = density of tungsten = 19.35 g/cm^3, M = atomic weight = 183.85, and N_0 = Avogadro's number. We find, as expected, that $n/n_Q = 588 \gg 1$.

We seek, therefore, the pressure of a nonrelativistic, degenerate gas of electrons of concentration n. We use $PV = 2U/3$ from (6) and the FD expression Eq. (30) for U from I,

$$P = (3\pi^2)^{2/3}\frac{\hbar^2}{5m}n^{5/3}\left(1+\frac{5\pi^2}{12}\left(\frac{T}{T_F}\right)^2+\cdots\right). \tag{12}$$

The contribution of the first term in (16) is P_0, the pressure at absolute zero. Numerically $P_0 = 7.40\times 10^5$ atm, with n taken from (11). The increase, ΔP, in the pressure due to the second term in (12) is 1.05×10^3 atm. Here T_F = the Fermi temperature $= \mu_0/k = 1.072\times 10^5$ K for tungsten. The very large value for the pressure is due to the high concentration, n, of conduction electrons, which is raised to the 5/3 power in (12).

Case 2 [Outside (halfway between the cathode and the anode)]: The conduction electrons which are boiled off of the cathode are clearly nonrelativistic since $e\phi_b \ll mc^2$, where ϕ_b is the anode voltage.

It is easy to show that the conduction electrons in the evacuated region between the cathode and the anode constitute a classical MB gas. Again we need n_Q and n. The quantum concentration $n_Q = 1.25\times 10^{19}/\text{cm}^3$ where we have used a nominal $T = 300$ K in (10) instead of $T = 2000$ K. This is justified because a vacuum is a good thermal insulator.

For simplicity, we use parallel plate geometry and the assumption of space charge limitation. The solution is very well known.[5] It leads to the Child–Langmuir law[6] and enables us to calculate n.

The current density

$$J = nev = 2.33\times 10^{-6}\phi_b{}^{3/2}/d^2 = 2.16\times 10^{20}/\text{cm}^3 \tag{13}$$

if we choose ϕ_b = plate voltage = 50 V with d = cathode-plate separation $= 2\times 10^{-2}$ m.

The potential

$$K_1 x^{4/3} = 19.8\ \text{V} \tag{14}$$

for $K_1 = 9.18\times 10^3\ \text{V}/m^{4/3}$, if we take $x = d/2$. The electron speed follows from

$$mv^2/2 = e\phi. \tag{15}$$

We find $v = 2.64\times 10^6$ m/sec^{-1} and $n = 4.88\times 10^6$ cm^{-3}. We see that $v/c = 8.8\times 10^{-3} \ll 1$ and that $n/n_Q = 3.90\times 10^{-13} \ll 1$. The electron gas is nonrelativistic and nondegenerate as expected.

The pressure of this gas follows readily from $PV = 2U/3$ and $U = 3NkT/2$ [I, Eq. (5)]. We obtain the familiar equation of state for an ideal classical gas

$$P = nkT = 2\times 10^{-13}\ \text{atm}. \tag{16}$$

Comparing Cases 1 and 2, it is interesting that conduction electrons boiled off from the cathode represent a minute fraction, 3.8×10^{-17}, of the conduction electrons inside a hot cathode. The outside electron gas pressure is 2.7×10^{-19} times the inside gas pressure. It is even 16 orders of magnitude less than the pressure due to the small T-dependent term in (12).

IV. THE PRESSURE OF A GAS OF HELIUM ATOMS. $S=0$

Case 3 (Liquid helium at 1 K): We expect that this system will be a nonrelativistic Bose gas. The Einstein condensation temperature[7] in He^4 occurs at $T = 2.174$ K. This is the temperature at which the number of atoms in excited states equals the total number of atoms. As the temperature is lowered, there is a condensation of atoms into states of zero momentum. At 1 K only 20% of the normal component of He^4 remains,[8] 80% is superfluid with no velocity. The normal atoms have speeds[9] ~ 50 m sec$^{-1} \ll c$.

The density of liquid helium at 1 K is 0.145 g/cm^3. To find the concentration n, we use (11) with $f = 1$, $M = 4$. Thus $n = 2.18\times 10^{22}/\text{cm}^3$. To calculate the quantum concentration n_Q, we use (10) with m = helium atomic mass $= 6.64\times 10^{-24}$ g and $T = 1$ K. Thus $n_Q = 1.52\times 10^{21}/\text{cm}^3$ and $n/n_Q = 14.3 > 1$.

The helium gas is a nonrelativistic, degenerate boson gas, as expected. To find the pressure we use $PV = 2U/3$ with U taken from I, Eq. (14):

$$P = 1.341\, n_Q kT. \tag{17}$$

The numerical factor in (17) is the Riemann zeta function $\zeta(\frac{5}{2}) = \Sigma_{m=1}^{\infty} m^{-5/2}$. Numerically $P = 0.278$ atm. We note that the pressure of the degenerate He gas is independent of the concentration of helium atoms, varies as $T^{5/2}$ since $n_Q \sim T^{3/2}$, and is slightly greater than $\frac{1}{4}$ of one atmosphere. Inasmuch as we are using an ideal gas model to describe a liquid, the results can only be considered qualitatively correct. However, if we use Eq. (11) of I to calculate the Einstein temperature on this model it turns out to be 3.15 K as compared to the experimental value of 2.174 K. Accordingly, we might expect the pressure results to be correct within a factor of 2 or so.

Case 4 (The vapor pressure of helium atoms at 1 K): We expect that this gas will be nonrelativistic and nondegenerate. The vapor pressure of helium is important in measuring temperature. Consequently it has received a lot of attention and can be calculated in several ways. We chose the empirical equation of Clement, Logan, and Gaffney,[10]

$$\ln P = I - A/T + B\ln T + CT^2/2 - D((\alpha\beta/\beta^2 - 1) - T^{-1})\tan^{-1}(\alpha T - \beta). \tag{18}$$

Here P is measured in mm of Hg. The constants are $I = 4.6202$, $A = 6.399$, $B = 2.541$, $C = 0.006\,12$,

$D = 0.5197$, $\alpha = 7.00$, and $\beta = 14.14$. This expression, which lacks the sixth term given by Clement *et al.*, is of sufficient accuracy for our purposes. For a given T, we can compute the pressure which we use in the following statistical calculation.

At the lambda point, $T = 2.174$ K and $P = 38.47$ mm of Hg. At $T = 1$ K, $P = 0.116$ mm of Hg, a decrease by a factor of 332 from the vapor pressure at the condensation temperature.

At 1 K, using (18) and assuming the gas is classical,

$$P = nkT = 1.53\times 10^{-4} \text{ atm } (0.116 \text{ mm of Hg}). \quad (19)$$

From (19), $n = 1.20\times 10^{18}/\text{cm}^3$. We calculated the quantum concentration in Case 3 and accordingly we find that $n/n_Q = 7.73\times 10^{-4}$, well within the MB regime. Using

$$mv^2/2 = 3kT/2 \quad (20)$$

we calculate $v = 7.90\times 10^3$ cm sec$^{-1} \ll c$.

This justifies the assumption that the helium gas, in this case, is nonrelativistic and classical. It is interesting that the ratio of helium atoms in the vapor to helium atoms in the liquid at 1 K is 5.51×10^{-5}. The ratio of the corresponding pressures is 5.50×10^{-4}. Both these values are several orders of magnitude larger than the corresponding ratios of electrons outside and inside a hot metal.

V. ULTRARELATIVISTIC PARTICLES

Case 5 (Electrons. $S=\frac{1}{2}$): Consider a Fermi gas of ultrarelativistic electrons. This problem is important in the theory of white dwarfs.[11] Relativistic effects occur when a gas is compressed, the average energy of the electrons rises, and the Fermi energy becomes comparable to the rest energy.

It is no longer correct to take

$$\mu_0 = (\hbar^2/2m)(3\pi^2 n)^{2/3} \quad (21)$$

to solve for n, because (21) is obtained by considering particle energy nonrelativistically as $\epsilon = p^2/2m$. Instead by using $\epsilon = pc$, the correct expression for μ_0 is obtained from the integral

$$N = C' \int_0^\infty \frac{\epsilon^2\, d\epsilon}{e^{(\epsilon-\mu)/kT} + 1}, \quad (22)$$

where C' is given by (7). It follows from I, Eqs. (16) and (27), that

$$\mu_0 = (3n\pi^2)^{1/3}\hbar c. \quad (23)$$

Now take $\mu_0 = m_0c^2$ and solve for n,

$$n = (m_0c/\hbar)^3/3\pi^2 = 5.86\times 10^{29} \text{ cm}^{-3}. \quad (24)$$

This is six orders of magnitude larger than the concentration of conduction electrons in a metal. This suggests, but does not establish, high degeneracy. The reason is that it is not correct to compare this n with the quantum concentration n_Q given by (10). We note that there $n_Q \sim m^{3/2}$ and that for the ultrarelativistic case $m_0 = 0$. Equation (10) is derived for $\epsilon = p^2/2m$ and like the Fermi energy must be rederived for $\epsilon = pc$. With the aid of Eq. (2) from I, $p = 2$, $C \to C'$, it follows that

$$\mu = kT \ln(n/n_Q)(2S+1)^{-1}. \quad (25)$$

This is the same form as Eq. (3) in I, but now

$$n_Q = (kT/\hbar c)^3/\pi^2. \quad (26)$$

The temperature in the interior of a white dwarf[12] is $\sim 10^7$ K. Thus $n/n_Q = (m_0c^2/kT)^3/3 = 6.98\times 10^7$, which proves this gas to be highly degenerate.

In the ultrarelativistic case $PV = U/3$ [see (9)], where U can be found from Eqs. (16) and (27) in I with $C \to C'$ and $p = 3$. It follows directly that

$$P = (3\pi^2)^{1/3}(n^{4/3})\hbar c/4 = 1.19\times 10^{17} \text{ atm}, \quad (27)$$

where n is given by (24).

Case 6 [Photons: $v=c$, $m_0=0$, $S=1$ (the quantum degeneracy factor is 2 not 3)]: Consider a gas of photons in a cavity. The fact that this gas is degenerate is very well known historically. The correct explanation of cavity radiation by Planck started quantum theory.[13] We establish degeneracy here by using the formalism of I,

$$N = C'(kT)^3 \int_0^\infty \frac{\epsilon^2\, d\epsilon}{e^{(\epsilon-\mu)/kT} - 1}, \quad (28)$$

where C' is given by (7). The chemical potential μ for a Bose (or Fermi) gas has a finite nonzero value if the number of particles N of the system is constant. But photons may be emitted or absorbed by the walls of the cavity. Thus N is not conserved. Therefore $\mu = 0$. Using Eqs. (6) and (8) in I,

$$n = 0.244(kT/\hbar c)^3. \quad (29)$$

From (26), $n/n_Q = 2.408$, which establishes the degeneracy for any T. To obtain (29), we use $\zeta(3) = \Sigma_{m=1}^\infty m^{-3} = 1.202$.

In the ultrarelativistic regime, $PV = U/3$. The average energy U may be calculated from Eqs. (6) and (8) in I with $C \to C'$ and $p = 3$. The pressure follows immediately,

$$P = 2(kT)^4\zeta(4)/\pi^2\hbar^3c^3, \quad (30)$$

where the Riemann zeta function $\zeta(4) = 1.082$. The photon gas pressure may be written in an equivalent, but more familiar form,[13]

$$P = 4\sigma T^4/3c, \quad (31)$$

where $\sigma = \pi^2k^4/60\hbar^3c^2$ W/m^2 (K)^4 is the Stefan–Boltzmann constant. Numerically $P = 3.99\times 10^{-8}$ atm for $T = 2000$ K and $\sigma = 5.67\times 10^{-8}$ in SI units.

Case 7 (The pressure due to a laser beam): In the previous example, the cavity is maintained at a temperature = 2000 K. The radiation emitted, through a small hole in the cavity, is characteristic of a thermal equilibrium distribution. Photons of all wavelengths leave and enter this hole.

In the case of a laser beam, photons of a single wavelength emerge from a "hole" at one end of a Fabry–Perot cavity. Here, we ask, what steady state power in the He–Ne laser at 632.8 nm is required to generate a beam pressure which is equal to the black-body radiation pressure of Case 6? We note that this physical situation does not constitute a statistical problem. But it does provide an interesting comparison.

Suppose the steady state power L of this laser is measured in watts. The beam intensity is the power divided by the cross sectional area of the beam,

$$I = L/\pi r^2. \quad (32)$$

TABLE I. The pressure of various physical systems.

Physical system	Spin	Statistics	Temperature (K)	Pressure (atm)
Free electrons in tungsten	$\frac{1}{2}$	Nonrelativistic FD	2000	7.4×10^{5}
Space charge limited electron flow	$\frac{1}{2}$	Nonrelativistic MB	300	2×10^{-13}
Liquid helium atoms	0	Nonrelativistic BE	1	2.78×10^{-1}
Helium vapor atoms	0	Nonrelativistic MB	1	1.53×10^{-4}
Electrons in a white dwarf star	$\frac{1}{2}$	Ultrarelativistic FD	10^{7}	1.19×10^{17}
Cavity radiation	1	Ultrarelativistic BE	2000	3.98×10^{-8}

The radiation pressure is

$$P = I/c = L/\pi r^2 c. \tag{33}$$

We take $P = 4.03\times10^{-3}\,N/m^2$ from the black-body radiation problem and solve for L, with $r = 1$ mm.

Here $L = 3.80$ W. This corresponds to $N = L\lambda/hc = 1.21\times10^{19}$ red He–Ne laser photons/sec.

The beam pressure is independent of wavelength, but of course the associated number of photons/sec depends on the energy of monochromatic photons.

VI. CONCLUSION

The results of this paper are summarized in Table I.

The methods developed in I make it quite easy to investigate the physical properties of diverse systems of interest. This paper focuses on the pressure of each system. This thermodynamic coordinate ranges from 10^{17} atm in a white dwarf to 10^{-13} atm in the evacuated region of a diode.

Each physical system in Table I is investigated to determine whether the particle speed is nonrelativistic or ultrarelativistic and whether the concentration is smaller or larger than the quantum concentration. This must be established, along with the particle's spin, in order to identify the appropriate statistical formulation, as given in Table I, column 3.

Ultrarelativistic particles, characterized by $\epsilon = pc$, lead to an extension of the formalism of I. There the nonrelativistic expression $\epsilon = p^2/2m$ is used. The essential change occurs in the volume element in momentum space. In the nonrelativistic case $4\pi p^2\,dp = 4\pi 2^{1/2} m^{3/2}\epsilon^{1/2}\,d\epsilon$. In the ultrarelativistic case this becomes $4\pi\epsilon^2\,d\epsilon/c^3$. All of the familiar quantities like chemical potential, quantum concentration, and internal energy are quite different in the ultrarelativistic regime.

[1] W. A. Barker, J. Math. Phys. **27**, 302 (1986).
[2] C. Kittel, *Elementary Statistical Physics* (Wiley, New York, 1958), p. 63.
[3] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Addison–Wesley, Reading, MA, 1958), p. 70.
[4] See, for example, C. Kittel and H. Kroemer, *Thermal Physics* (Freeman, San Francisco, 1980), p. 73.
[5] See, for example, *Applied Electronics*, edited by T. S. Gray (Wiley, New York, 1954), pp. 125–132.
[6] C. D. Child, Phys. Rev. **1**, 32, 498 (1911); I. Langmuir, *ibid.* **2**, 450 (1913).
[7] *Handbook of Physics and Chemistry* (CRC, Cleveland, 1977), 57th ed., p. B-25.
[8] Reference 4, p. 206.
[9] Reference 4, p. 215.
[10] J. R. Clement, J. K. Logan, and J. Gaffney, Bull. Inst. Int. Froid Suppl. Annexe **3**, 601 (1955). See also *Handbuch der Physik, Low Temperature Physics II* (Springer, Berlin, 1956), p. 405.
[11] M. Harwitt, *Astrophysical Concepts* (Wiley, New York, 1973), p. 163.
[12] See Ref. 7, p. 198.
[13] See Ref. 3, p. 175.

Series representations for calculations in quantum statistics. III

William A. Barker
Department of Physics, Santa Clara University, Santa Clara, California 95053

(Received 12 August 1986; accepted for publication 30 January 1987)

In two previous papers [J. Math. Phys. **27**, 1 (1986) (I); **28**, 1385 (1987) (II)] simple mathematical procedures are developed for averaging the particle energy ϵ, raised to any power $p \geqslant 1$ over Maxwell–Boltzmann, Bose–Einstein, and Fermi–Dirac distributions. In I, the particle energy is treated nonrelativistically: $\epsilon = p^2/2m$. In II, the particle energy is treated ultrarelativistically: $\epsilon = pc$. Here, the particle energy is treated relativistically $\epsilon^2 = p^2c^2 + m_0^2c^4$. For each distribution function, expressions are obtained for the chemical potential, internal energy, and heat capacity for two cases: $(kT/m_0c^2)^2 \ll 1$ and $(m_0c^2/kT)^2 \ll 1$.

I. INTRODUCTION

In the literature of classical and quantum statistics, most calculations are made for nonrelativistic particles.[1] There are some which use the ultrarelativistic case,[2] but there are very few which treat the particle energy exactly.[3] From a mathematical point of view, this is understandable. To average the relativistic density of states expression over any of the distributions

$$f(\epsilon) = 1/(e^{(\epsilon-\mu)/kT} \pm 1{,}0) \tag{1}$$

is indeed formidable. Furthermore, most of the cases of physical interest are nonrelativistic. However, the mathematical simplifications introduced in I make it quite easy to use two approximate forms: either $(kT/m_0c^2)^2 \ll 1$ or $(m_0c^2/kT)^2 \ll 1$. The purpose of this paper is to use these two approximations to the relativistic density of states expression and extend I and II to derive the corresponding expressions for the chemical potential, energy, and heat capacity for Maxwell–Boltzmann (MB), Bose–Einstein (BE), and Fermi–Dirac (FD) statistics.

II. DENSITY OF STATES

The density of states takes three different forms in calculations using MB, BE, and FD statistics.

In the nonrelativistic approximation,[4] $\epsilon = p^2/2m$ and

$$D(\epsilon)d\epsilon = (2S+1)2^{1/2}Vm^{3/2}\epsilon^{1/2}\,d\epsilon/(2\pi^2\hbar^3)\,. \tag{2}$$

In the ultrarelativistic limit,[5] $\epsilon = pc$ and

$$D(\epsilon)d\epsilon = (2S+1)V\epsilon^2\,d\epsilon/(2\pi^2\hbar^3c^3)\,. \tag{3}$$

In the relativistic case, which includes the foregoing approximations,

$$\epsilon^2 = p^2c^2 + m_0^2c^4 \tag{4}$$

and

$$D(\epsilon)d\epsilon = (2S+1)V\epsilon(\epsilon^2 - m_0^2c^4)^{1/2}\,d\epsilon/(2\pi^2\hbar^3c^3)\,. \tag{5}$$

Each of these expressions is easily found from

$$D(\mathbf{p},\mathbf{r})d\tau = (2S+1)dp_x\,dp_y\,dp_z\,dx\,dy\,dz/(2\pi\hbar)^3\,. \tag{6}$$

First, transform to polar coordinates in momentum space, $dp_x\,dp_y\,dp_z = 4\pi p^2\,dp$, and then use each of the three relationships connecting particle momentum and energy.

III. CHEMICAL POTENTIALS; THE FERMI, BOSE, AND CLASSICAL ENERGY

The range of energy in the three types of integrals in I and II is from 0 to ∞. This makes the mathematics tractable. However, in the relativistic expression (4), the lower limit on $\epsilon = m_0c^2$, corresponding to $p = 0$. This suggests a change in variables to $x = (\epsilon - m_0c^2)/kT$. The general form of the integral is now

$$I = \int_0^\infty \frac{f(x)dx}{e^x e^{-(\mu - m_0c^2)/kT} \pm 1{,}0}\,. \tag{7}$$

The quantity $(\mu - m_0c^2)$ in the denominator of (7) is a crucial parameter in this theory. In the FD case, the Fermi energy $E_F = \mu - m_0c^2$. By analogy, in the MB and BE cases, we define the Bose energy E_B and the classical energy E_C to be $\mu - m_0c^2$. The chemical potential μ is different for each of the distribution functions. It depends on particle concentration or number, rest energy, and temperature. The corresponding energy parameter reflects this dependence.

IV. EXACT AND APPROXIMATE SOLUTIONS

At this point, we make two exact calculations using (5), with a limiting value for the FD distribution. This will be used to provide an internal consistency check on the approximate FD calculations and as a guide for interpreting results obtained from the other statistical distributions.

Define $W \equiv \epsilon - m_0c^2$ and consider the FD distribution function

$$f(W) = 1/(e^{(W-E_F)/kT} + 1) \tag{8}$$

in the limit as $T \to 0$. For $0 \leqslant W \leqslant E_F$, $f(W) = 1$; W has an upper limit, namely, $E_F(0,m_0)$, the Fermi energy at $T = 0$.

There is a very simple argument which leads to an exact expression for $E_F(0,m_0)$. Start with the density of states expression (6), take $S = \frac{1}{2}$, write out the integral for the total number of particles N, using the Fermi function $f(p) = 1$,

$$N = \frac{V}{\pi^2\hbar^3}\int_0^{p_0} p^2\,dp\,, \tag{9}$$

where p_0 is the limiting value of the momentum corresponding to $E_F(0,m_0)$. Use the relativistic formula (4) and $W = \epsilon - m_0c^2$ to rewrite N as an integral over W,

 0022-2488/87/061389-05$02.50

$$N=\frac{V}{\pi^2\hbar^3c^3}\int_0^{E_F(0,m_0)}(W+m_0c^2)$$
$$\times(W^2+m_0c^2)(W^2+2Wm_0c^2)^{1/2}\,dW. \quad (10)$$

This yields

$$n=N/V=(E_F^2+2E_Fm_0c^2)^{3/2}/3\pi^2\hbar^3c^3. \quad (11)$$

In the limit as $m_0\to 0$,

$$E_F(0,0)=(3\pi^2n)^{1/3}\hbar c=p_0c, \quad (12)$$

a value which may be used in other expressions for $E_F(T,m_0)$. Using (11), the exact relation for

$$E_F(0,m_0)=(m_0^2c^4+E_F^2(0,0))^{1/2}-m_0c^2. \quad (13)$$

If $E_F/m_0c^2\ll 1$, it follows from (13) that

$$E_F(0,\infty)=E_F^2(0,0)/2m_0c^2=(3\pi^2n)^{2/3}\hbar^2/2m_0. \quad (14)$$

If $m_0c^2/E_F\ll 1$, it follows from (13) that

$$E_F(0,m_0)=E_F(0,0)(1-m_0c^2/E_F(0,0)). \quad (15)$$

The physical significance of these four expressions for the Fermi energy will be discussed later in the paper.

To find the internal energy $U=U(0,m_0)$, multiply the integrand in (9) by ϵ from (4)

$$U=\frac{Vc}{\pi^2\hbar^3}\int_0^{p_0}p^2(p^2+m_0^2c^2)^{1/2}\,dp. \quad (16)$$

The result, expressed in terms of $E_F=E_F(0,0)$, is

$$U=V/8\pi^2\hbar^3c^3[E_F(2E_F^2+m_0^2c^4)(E_F^2+m_0^2c^4)^{1/2}$$
$$-m_0^4c^8\sinh^{-1}(E_F/m_0c^2)). \quad (17)$$

Equation (17) is equivalent to the formula for the energy given by Landau and Lifschitz.[6] Its form is more convenient than an alternate expression for U written in terms of $E_F(0,m_0)$, which may be obtained by multiplying the integrand of (10) by $\epsilon=W+m_0c^2$. The foregoing formula (17) will be used later in the paper to compare with various approximate forms for $U(T,m_0)$.

V. THE MAXWELL–BOLTZMANN DISTRIBUTION

Case A [$(kT/m_0c^2)^2\ll 1$]: In this approximation, the density of states (5) is

$$D(x)dx\cong[2^{1/2}(2S+1)V(m_0kT)^{3/2}/2\pi^2\hbar^3]$$
$$\times(x^{1/2}+5ax^{3/2}/4+7a^2x^{5/2}/32+\cdots), \quad (18)$$

where $a=(kT/m_0c^2)$. The total number of particles

$$N=e^{E_C/kT}\int_0^\infty D(x)e^{-x}\,dx$$
$$=A(kT)^{3/2}e^{E_C/kT}(1+1.88a+0.82a^2), \quad (19)$$

where $A=(2S+1)Vm_0^{3/2}/(2\pi\hbar^2)^{3/2}$. The integral (19) is the sum of three gamma functions of the form $\Gamma(p+1)$ where p is the exponent of x. Solving for the chemical energy yields

$$E_C=kT\ln(n/n_Q(2S+1))(1-1.88a+2.70a^2), \quad (20)$$

where

$$n_Q=(m_0kT/2\pi\hbar^2)^{3/2}. \quad (21)$$

Compare with Eqs. (3) and (4) of I, which apply to the nonrelativistic approximation. Here, $E_C=\mu-m_0c^2$ replaces the chemical potential. It has the standard form, with two correction terms. The quantum concentration is identical with the nonrelativistic value in I.

The energy and the heat capacity follow directly. Multiply dN by $\epsilon=m_0c^2+kTx$ and integrate,

$$U=Nm_0c^2(1+1.25a-1.25a^2), \quad (22)$$

$$C_V=\tfrac{3}{2}Nk(1+2.50a-3.75a^2). \quad (23)$$

Certainly the rest energy term in (22) was to be expected. The second term in (22) and the first term in (23) agree with Eqs. (5) and (6) in I. The remaining terms are thermal corrections to the nonrelativistic approximation.

Case B [$(m_0c^2/kT)^2\ll 1$]: In this approximation, the density of states is given by

$$D(x)dx=[(2S+1)V(kT)^3/2\pi^2\hbar^3c^3]$$
$$\times(x^2+2bx+7b^2/8)dx, \quad (24)$$

where $b=m_0c^2/kT$.

The total number of particles

$$N=e^{E_C/kT}\int_0^\infty D(x)e^{-x}\,dx$$
$$=B(kT)^3e^{E_C/kT}(1+b+0.438b^2), \quad (25)$$

where $B=(2S+1)V/2\pi^2\hbar^3c^3$. Solving for the classical energy

$$E_C=kT\ln(n/n_Q(2S+1))(1-b+0.562b^2), \quad (26)$$

where

$$n_Q=(kT/\hbar c)^3\pi^{-2} \quad (27)$$

is the ultrarelativistic expression for the quantum concentration. [See Eq. (26) of II.] E_C has the expected form plus both linear and quadratic rest energy correction terms.

To obtain the energy U, multiply the integrand in (25) by $kTx+m_0c^2$ and integrate,

$$U=e^{E_C/kT}\int_0^\infty(kTx+m_0c^2)D(x)e^{-x}\,dx. \quad (28)$$

With the aid of (25), the energy and heat capacity are written

$$U=3NkT(1+1.92b^2), \quad (29)$$

$$C_V=\frac{dU}{dT}=3Nk(1-1.92b^2). \quad (30)$$

The ultrarelativistic results[7] follow from (29) and (30) in the limit as $m_0\to 0$. Here U and C_V both have a quadratic rest energy correction term. It is interesting to note that (22), in the nonrelativistic limit, and (29) in the ultrarelativistic limit, lead to precisely the same familiar equation of state $PV=NkT$. This follows because $PV=2U/3$ and $U/3$ in these two limits, respectively. However, neither of these relationships are consequences of the relativistic expression connecting particle energy and momentum.

VI. THE BOSE–EINSTEIN DISTRIBUTION

The BE integrals have the form

$$I_{\rm BE} \propto \int_0^\infty \frac{f(x)dx}{e^x e^{-E_B/kT} - 1}, \tag{31}$$

where $E_B = (\mu - m_0c^2)$ is the Bose energy. Let $z = E_B/kT$. The quantity e^{-z} can be shown to be ≈ 1 below the Einstein condensation temperature. The argument parallels the development in I [See Eqs. (10)–(12)] inclusive. If N_0 is the occupation number of the ground state when $\epsilon = m_0c^2$ and $x = 0$, then

$$N_0 = (e^{-z} - 1)^{-1}, \tag{32}$$

$$e^{-z} = 1 + 1/N_0, \tag{33}$$

$$E_B \cong -kT/N_0. \tag{34}$$

Case A [$(kT/m_0c^2)^2 \ll 1$]: The density of states is given by (18). The total number of bosons in excited energy states

$$N_e = \int_0^\infty \frac{D(x)dx}{e^x - 1}. \tag{35}$$

The integral involves the sum of three gamma-Riemann zeta function products. As in I, this integrates immediately to

$$N_e = 2.612(2S+1)n_Q V(1 + 0.962a + 0.355a^2). \tag{36}$$

This agrees with the nonrelativistic expression (10) in I, with two thermal correction terms. To find U, multiply the integrand in (35) by $(m_0c^2 + kTx)$,

$$U = m_0c^2 \int_0^\infty \frac{(1+ax)D(x)dx}{e^x - 1}. \tag{37}$$

This integrates to

$$U = N_e m_0c^2 + 1.341(2S+1)(3NkT/2) \times (1 + 2.62a + 1.51a^2)n_Q/n. \tag{38}$$

There is an important distinction between N_e and N in Eq. (38). N_e is the number of excited bosons whose $p > 0$. Here N is the total number of bosons, excited and condensed. It enters (38) via $V = N/n$. This distinction can be kept in mind by realizing that the Einstein condensation is a condensation in momentum space, but not in coordinate space.

The heat capacity C_V follows by differentiating U with respect to T,

$$C_V = 3.35(2S+1)(3NkT/2) \times (1 + 9.19a + 6.77a^2)n_Q/n. \tag{39}$$

Equations (38) and (39) agree with the nonrelativistic Eqs. (14) and (15) in I in the appropriate limit. Both U and C_V have linear and quadratic correction terms.

Case B [$(m_0c^2/kT)^2 \ll 1$]: In this approximation, the density of states is given by (24). The total number of bosons in excited states

$$N_e = \int_0^\infty \frac{D(x)dx}{e^x - 1}. \tag{40}$$

This integrates to

$$N_e = 1.202(2S+1)N(1 + 1.31b)n_Q/n, \tag{41}$$

where the quantum concentration is the ultrarelativistic value given in (27). The third term in the density of states expression (24) is not carried in this calculation, as it leads to a divergent term in (40). The energy

$$U = \int_0^\infty \frac{(kTx + m_0c^2)D(x)dx}{e^x - 1} \tag{42}$$

integrates to

$$U = N_e mc^2 + 1.082(2S+1)(3NkT)(1 + 0.741b)n_Q/n. \tag{43}$$

The corresponding heat capacity is

$$C_V = 13(2S+1)Nk(1 + .556b)n_Q/n. \tag{44}$$

The correction terms in (43) and (44) are linear terms in $b = m_0c^2/kT$.

The Einstein condensation temperature, in the ultrarelativistic case, is an additional quantity of interest in BE statistics. The Einstein temperature T_E is that temperature for which the number of bosons N_e in excited states equals the total number of bosons N. Using (41) with $m_0 = 0$

$$T_E = 2.017(n/2S+1)^{1/3}\hbar c/k. \tag{45}$$

The fraction of particles in excited states is

$$N_e/N = (T/T_E)^3, \tag{46}$$

and the number of particles in the ground state is

$$N_0 = N - N_e = N(1 - (T/T_E)^3). \tag{47}$$

Consider a photon gas at $T = 2000$ K with a concentration of $n = 10n_Q = 6.74\times 10^{11}/\text{cm}^3$. The Einstein condensation temperature from (45) is 3.22×10^3 K. The fraction in excited states is about 0.24. Compare this with the Einstein condensation for liquid helium. The experimental value for $T_E = 2.174$ K. Using $N_e/N = (T/T_E)^{3/2}$, the fraction in excited states at 1 K is about 0.31. Actually, there is no photon condensation because the number of photons is not a constant.[8]

VII. THE FERMI–DIRAC DISTRIBUTION

Fermi–Dirac integrals, treated relativistically, are represented by a superposition of integrals of the form

$$I_{\rm FD} \propto \int_0^\infty \frac{x^p\,dx}{e^x e^{-E_F/kT} + 1}, \tag{48}$$

where $E_F = \mu - m_0c^2$ and $p = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \ldots$. With $z = E_F/kT$, (48) has the same structure as Eq. (16) in I. The Blankenbecler[9] method applies and we can write

$$I_{\rm FD} \propto \left(1 + \frac{\pi^2}{6}\frac{\partial^2}{\partial z^2} + \cdots\right)\frac{z^{p+1}}{p+1}. \tag{49}$$

Case A [$(kT/m_0c^2)^2 \ll 1$]: The approximate form for the density of states in this case is given by (18). The total number of particles

$$N = \int_0^\infty \frac{D(x)dx}{e^x e^{-z} + 1}. \tag{50}$$

Using (49) and solving for N with $S = \frac{1}{2}$,

$$N = \tfrac{4}{3}AE_F^{3/2}(1 + (3E_F/4m_0c^2) + (\pi^2/8)(kT/E_F)^2), \tag{51}$$

where $A = Vm_0^{3/2}/(2\pi\hbar^2)^{3/2}$. In the limit as $T \to 0$, $m_0 \to \infty$, we find that

$$E_F(0,\infty) = \frac{E_F^2(0,0)}{2m_0c^2} = (3\pi^2 n)^{1/3}/2m_0, \quad (52)$$

in agreement with (14). Retaining $T=0$, we can solve for

$$E_F(0,m_0) = \frac{E_F^2(0,0)}{2m_0c^2}\left(1 - \frac{E_F^2(0,0)}{4m_0^2c^4}\right), \quad (53)$$

in agreement with an expansion of the exact expression (13). Finally

$$E_F(T,m_0) = E_F(0,m_0)(1 - (\pi^2/12)(kT/E_F)^2), \quad (54)$$

which is the standard result,[10] subject to the new meaning for $E_F(0,m_0)$, as given by Eq. (53).

To solve for U multiply dN in (50) by $\epsilon = m_0c^2 + kTx$,

$$U = \int_0^\infty \frac{(m_0c^2 + kTx)D(x)dx}{e^x e^{-z} + 1}. \quad (55)$$

The result is

$$U = Nm_0c^2 + (\tfrac{3}{5})NE_F(0,m_0)(1 + (5\pi^2/12)(kT/E_F)^2). \quad (56)$$

The algebra is tedious, but the result is of the expected form, subject to a new meaning for $E_F(0,m_0)$.

It is a simple matter to demonstrate that $\lim_{T\to 0} U$ is an approximation for U given in (17). To show this, note that $V = N/n$, $n\hbar^3c^3 = E_F^3(0,0)/3\pi^2$, and $E_F(0,m_0) \simeq E_F^2(0,0)/2m_0c^2$. Expand (17) to the second power in $E_F(0,0)$. The values for $U(0,m_0)$ from (56) and (17) agree,

$$U(0,m_0) = Nm_0c^2(1 + (3E_F^2(0,0)/10m_0^2c^4)). \quad (57)$$

The heat capacity

$$C_V = \pi^2 Nk^2 k^2 T/2E_F(0,m_0) \quad (58)$$

is of standard form, but there is a correction for $E_F(0,m_0)$, as shown in (53).

Case B [$(m_0c^2/kT)^2 \ll 1$]: The density of states is given by (24). The total number of particles

$$N = \int_0^\infty \frac{D(x)dx}{e^x e^{-z}+1} = \frac{VE_F^3}{3\pi^2\hbar^3c^3}(1 + (3m_0c^2/E_F) \times (1 + m_0c^2/2E_F) + \pi^2(kT/E_F)^2). \quad (59)$$

From (59)

$$E_F(0,m_0) = E_F(0,0)(1 - m_0c^2/E_F + m_0^2c^4/2E_F^2) \quad (60)$$

in agreement with an expansion of the exact expression (13). Again from (59) and (60),

$$E_F(T,m_0) = E_F(0,0)(1 - m_0c^2/E_F + m_0^2c^4/2E_F^2 - \pi^2(kT/E_F)^2). \quad (61)$$

To find U, multiply the integrand in (59) by $\epsilon = kTx + m_0c^2$ and integrate. After some tedious algebra, the result is

$$U(T,m_0) = \tfrac{3}{4}NE_F(0,0) \times (1 + \tfrac{4}{3}(m_0^2c^4/E_F^2) + \pi^2(kT)^2/E_F^2). \quad (62)$$

This expression, at $T=0$, agrees with an expansion of the exact formula (17). It is easy to see that the second term in (17) does not contribute to this order. Write

$$0.5(m_0c^2/E_F)^4 \sinh^{-1}(E_F/m_0c^2) = -0.5(m_0c^2/E_F)^4 \times \log[m_0c^2/((E_F^2 + m_0^2c^4)^{1/2} + E_F) + 1]. \quad (63)$$

An expansion generates terms $O(m_0c^2/E_F)^5$ and smaller.

The heat capacity

$$C_V = (3\pi^2/4)Nk(kT/E_F(0,0)). \quad (64)$$

This has the same structure as the Case A result [Eq. (58)]. However, the heat capacity in Case A is larger than in Case B because $E_F(0,m_0)$ is smaller than $E_F(0,0)$. See Eq. (52).

VIII. DISCUSSION

A. Three tests distinguish among the 12 statistical formulations

For the individual who is about to make some calculations in statistical mechanics, there may appear to be a bewildering array of choices. For each of the three distributions, MB, BE, and FD, there are four possible expressions for the density of states: nonrelativistic, relativistic A, relativistic B, and ultrarelativistic.

The choice is clear once three tests are made.

In the first test, a comparison of kT with m_0c^2 distinguishes amongst the density of states formulas. If $kT/m_0c^2 \leqslant 10^{-2}$, use the nonrelativistic formula [Eq. (2)]. If $kT/m_0c^2 > 10^{-2}$ and $(kT/m_0c^2)^2 \leqslant 10^{-2}$, use the relativistic A formula [Eq. (18)]. If $m_0c^2/kT > 10^{-2}$ and $(m_0c^2/kT)^2 \leqslant 10^{-2}$, use the relativistic B formula [Eq. (24)]. If $m_0c^2/kT \ll 10^{-2}$, use the ultrarelativistic formula [Eq. (3)].

In the second test, the particle density is compared with the quantum concentration n_Q. This determines whether the problem is classical or degenerate.

If the first test establishes the density of states to be either nonrelativistic or relativistic A, then if $n \ll n_Q = (mkT/2\pi\hbar^2)^{3/2}$, the problem is nondegenerate. Use MB statistics. If $n \gtrsim n_Q$, the problem is degenerate. Use BE or FD statistics. On the other hand, if the first test establishes the density of states to be either relativistic B or ultrarelativistic, then if $n \ll n_Q = (kT/\hbar c)^3/\pi^2$, the problem is nondegenerate and calls for MB statistics. If $n \gtrsim n_Q$, the problem is degenerate, requiring either BE or FD statistics.

In the third test, the particle spin is used to distinguish bosons from fermions. Let $S_0 = n\hbar/2$, where $n = 0,1,2,\ldots$. Then for n odd (even), the particle is a fermion (boson).

If tests 1 and 2 establish the problem as degenerate and either nonrelativistic, relativistic A, relativistic B, or ultrarelativistic, then n even (odd) requires BE (FD) statistics.

B. The classical, Bose, and Fermi energies

In MB statistics, the leading term for the classical energy E_C has the same form for both the relativistic A and B regimes [Eqs. (20) and (26)]. However, the quantum concentration n_Q, as described in test 2, has two quite different meanings [Eqs. (21) and (27)]. The first (second) is appropriate for relativistic A (B).

Suppose that $S = \frac{1}{2}$ and that $n/n_Q \leqslant 10^{-2}$, then $E_C/kT = \ln(n/n_Q) \leqslant -4.6$. The classical energy E_C will always be negative, as required by $n \ll n_Q$. In this situation, MB statistics are used.

The Bose energy E_B has a small negative value $\sim -kT/N$, if the particle number N is conserved. However, $E_B = 0$ if the particle number is not conserved. The former value is valid if the actual temperature is less than the Einstein condensation temperature.

The Fermi energy takes several forms. In the case of a problem which is nonrelativistic or relativisitic A, the leading term is $E_F(0,m_0) = (3\pi^2 n)^{2/3}\hbar^2/2m$. However, if the problem is relativistic B or ultrarelativistic, the leading term is $E_F(0,0) = (3\pi^2 n)^{1/3}\hbar c$. These two expressions differ substantially in structure and magnitude. Consider, for example, conduction electrons of concentration $n = 10^{22}/\text{cm}^3$. Then $E_F(0,m_0) = 1.70$ eV and $E_F(0,0) = 1.32\times 10^3$ eV. All the other contributions to E_F are due to rest energy and thermal effects and are less than these leading terms.

C. The third law of thermodynamics

The third law of thermodynamics requires that $\lim_{T\to 0} C_V = 0$. This law is violated by the MB heat capacity expressions (23) and (30), as is well known. However, the BE and FD expressions for C_V, as given in Eqs. (39), (44), (58), and (64), are in agreement with the third law. In verifying this for BE statistics, note that $n_Q \sim T^{3/2}$ or T^3. This temperature dependence of the quantum concentration guarantees that the correction terms, as well as the leading terms, go to zero as $T\to 0$.

IX. CONCLUSION

The methods developed in I are used, in this paper, to give a systematic discussion of MB, BE, and FD distributions when the density of states expression is based on the exact relativistic energy-momentum relationship. Two approximations are used. When $(kT/m_0c^2)^2 \ll 1$, the results are the same as for the nonrelativistic approximation, but with correction terms in (kT/m_0c^2). When $(m_0c^2/kT)^2 \ll 1$, the results agree with the ultrarelativistic approximation but with correction terms in m_0c^2/kT. The principal focus of the paper is on the quantity $\mu - m_0c^2$, where μ is the chemical potential. This is the Fermi energy in FD statistics, and it is given a comparable definition for BE and MB statistics. As in I, calculations are made for the chemical potential, internal energy, and heat capacity for each of the three distributions. A feature of the paper are two exact solutions. The relativistic density of states formula is used, without approximation, with the FD distribution function at $T = 0$, $\epsilon \leqslant E_F$ to obtain exact expressions for the Fermi energy and the internal energy. When these formulas are expanded in a series, they agree to the same order with the results obtained using the series representation in I.

[1]C. Kittel, *Elementary Statistical Physics* (Wiley, New York, 1958), pp. 86–96; see also C. Kittel and H. Kroemer, *Thermal Physics* (Freeman, San Francisco, 1980), pp. 183–198.
[2]L. D. Landau and E.M. Lifschitz, *Statistical Physics* (Addison–Wesley, Reading, MA, 1968), pp. 165–167.
[3]See Ref. 2, pp. 167 and 168.
[4]William A. Barker, J. Math. Phys. **27**, 1 (1986).
[5]William A. Barker, J. Math. Phys. **28**, 1385 (1987).
[6]See Ref. 2, p. 168.
[7]See C. Kittell, Ref. 1, p. 60.
[8]See Ref. 1, Kittel and Kroemer, p. 202.
[9]R. Blankenbecler, Am. J. Phys. **25**, 279 (1957).
[10]See Ref. 1, Kittel, p. 94. [Eq. (20.24)].

Stable thermodynamic states

J. S. Cohen
Philips Research Laboratories, Eindhoven, The Netherlands

M. Winnink
Institute for Theoretical Physics, University of Groningen, The Netherlands

(Received 2 December 1986; accepted for publication 11 February 1987)

A general class of perturbations of the dynamics for thermodynamic quantum systems is discussed. Without making use of weak asymptotic Abelianess, stability of a state for these perturbations is shown to lead to the ϕ-KMS condition and to the KMS condition in particular cases. Conversely, ϕ-KMS states satisfy the stability property introduced here.

I. INTRODUCTION

The derivation of equilibrium properties for states of thermodynamic systems has already been of interest for some time. It is well known that so-called KMS states may be obtained from stability for perturbation of the dynamics. This has been discussed initially by Haag, Kastler, and Trych-Pohlmeyer, by Kastler and Bratteli, and by Hoekman; for a review cf. Ref. 1, Chap. 5.4.2. However, owing to the assumed rapid decay of the time correlation functions these KMS states can only describe pure thermodynamic phases. A further restriction of the method is that it can be applied only to states of dynamical systems that are weakly asymptotically Abelian; viz. $\int dt\, \omega([A,\alpha_t B]) = 0$. Here A, B denote elements of the C^* algebra $\mathfrak{U}$; $\omega \in E_{\mathfrak{U}}$ is a state over $\mathfrak{U}$, and $\alpha_t \in \mathrm{aut}\,\mathfrak{U}$ describes the time evolution.

In this paper we shall discuss a stability property which leads to states that satisfy the ϕ-KMS condition introduced recently.[2] The main advantage of the stability criterion put forward here is that neither are assumptions made on the decay of the correlation functions nor is the dynamics assumed to act weakly asymptotically Abelian.

Depending on the details of the perturbation for which stability is imposed, the ϕ-KMS states in some cases are KMS states. For an infinite quantum lattice system the ϕ-KMS condition and KMS condition are equivalent.[2] Consequently, in this instance either our stability condition leads to a KMS state or the system does not admit states that are stable for the particular perturbation. For any finite system or for continuous quantum systems, however, the only states that fulfill the stability criterium are ϕ-KMS states.

For a finite system the presently proposed stability property is stronger than the condition imposed by Lebowitz *et al.*[3] For thermodynamic systems our conditions are weaker than those introduced by Kastler[4] (cf. Ref. 5).

II. A GENERALIZED PERTURBED DYNAMICS

In the Heisenberg picture the equation of motion for the unperturbed evolution reads

$$\frac{d}{dt}\alpha_t(A) = i\alpha_t(\delta(A)), \quad A \in D(\delta) \subset \mathfrak{U}, \tag{2.1}$$

where the derivation δ is the infinitesimal generator of the group of * automorphisms $\{\alpha_t\}$. A perturbed dynamics can be considered as the solution of the differential equation

$$\frac{d}{dt}\tilde{\alpha}_t^h(A) = i\,\tilde{\alpha}_t^h(\delta(A)) + i\,\tilde{\alpha}_t^h([h_t, A]), \tag{2.2}$$

cf., e.g., Ref. 6, with $h_t = h_t^* \in \mathfrak{U}$. One may choose the particular form $h_t = f(t)h$ with $h = h^* \in \mathfrak{U}$ and $f: \mathscr{R} \to \mathscr{R}$, so that the perturbation becomes localized in time if, e.g., $\mathrm{supp} f$ is compact or $f \in L_1(\mathscr{R})$. The family $\{h_t\}$ can be interpreted as the action of some external agent on the system. Owing to the time dependence of h_t the mappings $\{\tilde{\alpha}_t^h\}$ do not form a group.

A further generalization of the perturbed dynamics is obtained from the following equation of motion:

$$\frac{d}{dt}\tilde{\alpha}_t^h(A) = i\,\tilde{\alpha}_t^h(\delta(A)) + if_1(t)\tilde{\alpha}_t^h(hA) - if_2(t)\tilde{\alpha}_t^h(Ah). \tag{2.3}$$

The solution to this equation is the family of mappings $\tilde{\alpha}_t^h$: $\mathfrak{U} \to \mathfrak{U}$ given by

$$\tilde{\alpha}_t^h(A) = \tilde{\gamma}_t^h(\alpha_t(A)), \tag{2.4a}$$

$$\tilde{\gamma}_t^h(A) = \tilde{u}_1^h(t) A \tilde{u}_2^h(t)^*, \tag{2.4b}$$

$$\tilde{u}_j^h(t) = \sum_{n=0}^{\infty} \left\{ i^n \int_0^t ds_1 \cdots \int_0^{s_{n-1}} ds_n \prod_{k=1}^{n} [f_j(s_k)\alpha_{s_k}(h)] \right\}, \tag{2.4c}$$

with $f_j \in L_1(\mathscr{R})$ and $h = h^* \in \mathfrak{U}_0$. The unperturbed dynamics α_t is assumed to be strongly continuous on a $\sigma(\mathfrak{U},N)$-dense subalgebra $\mathfrak{U}_0 \subset \mathfrak{U}$. Here N denotes the set of locally normal states on $\mathfrak{U}$ (cf. Ref. 7). In general, the mappings $\tilde{\alpha}_t^h$ and $\tilde{\gamma}_t^h$ will not be positivity preserving. The operators $\tilde{u}_j^h(t)$ are easily seen to be unitary. The integrals in (2.4c) exist as Bochner integrals. We now give some useful properties of the generalized perturbed dynamics in the following.

Proposition 2.1: For $A \in \mathfrak{U}$ and $h = h^* \in \mathfrak{U}_0$,

$$\tilde{u}_j^h(t) = \mathbb{1} + i\int_0^t ds f_j(s)\alpha_s(h)\tilde{u}_j^h(s), \tag{2.5a}$$

$$\frac{d}{dt}\tilde{u}_j^h(t) = if_j(t)\alpha_t(h)\tilde{u}_j^h(t); \tag{2.5b}$$

$$\tilde{\gamma}_t^h(A) = A + i\int_0^t ds[f_1(s)\tilde{\gamma}_s^h(\alpha_s(h)A) - f_2(s)\tilde{\gamma}_s^h(A\alpha_s(h))], \tag{2.5c}$$

 0022-2488/87/061394-04$02.50

$$\tilde{\alpha}_t^h(A) = \alpha_t(A) + i\int_0^t ds[\, f_1(s)\alpha_s(h)\alpha_t(A) - f_2(s)\alpha_t(A)\alpha_s(h)\,] + \cdots . \tag{2.5d}$$

The omitted terms in (2.5d) are $O(h^2)$.

Proof: By iteration of (2.5a) we obtain (2.4c). The equivalence of (2.5c) and (2.5b) then follows from the initial condition $\tilde{u}_j^h(0) = 1$. Finally, (2.5c) and (2.5d) are obtained with the use of (2.4b) and (2.4c) along similar lines as in the discussion of the cocycle property (cf., e.g., Ref. 1).

We now turn to the introduction of the notion of stability for perturbations from the unperturbed dynamics α_t as described by (2.5d). Succintly, one assumes that close to the original state $\omega\in\mathfrak{U}^*$ there exists a bounded linear functional $\omega^h\in\mathfrak{U}^*$ that is *almost* invariant for the perturbed evolution $\tilde{\alpha}_t^h$. At this point we shall impose some restrictions on the functions f_j.

Definition 2.2: For a pair of functions f_1 and f_2 such that

(1) $f_j\in L_1(\mathscr{R})\cap C^1(\mathscr{R})$;

(2) $\hat{f}_j\in C^\infty(\mathscr{R})$ and invertible on sp α, i.e.,

$$\hat{f}_j(\lambda)\neq 0 \quad \forall\lambda\in \mathrm{sp}\,\alpha = \{\lambda\in\mathscr{R}\,|\,\hat{g}(\lambda)\neq 0 \quad \forall g:\ \int dt\, g(t)\alpha_t(A) = 0 \quad \forall A\in\mathfrak{U}_0\};$$

(3) $\hat{f}_1(\lambda) = \hat{f}_2(\lambda)$ iff $\lambda = 0$;

[$\hat{g}(\lambda) = \int dt\, e^{-i\lambda t}g(t)$ is the Fourier transform] we say that a state ω is (f_1,f_2)-stable if there exists a bounded linear functional $\omega^{\mu h}\in\mathfrak{U}^*$ such that for μ in a neighborhood of the origin

$$\lim_{t\to\pm\infty}\omega^{\mu h}(\tilde{\gamma}_t^{\mu h}A) = \omega_\pm(A); \tag{2.6}$$

$$\omega_+^{\mu h}(A) - \omega_-^{\mu h}(A) = o(\mu); \tag{2.7}$$

and

$$\lim_{\mu\to 0}\omega^{\mu h}(\alpha_t A) = \omega(\alpha_t A), \text{ uniformly in } t \tag{2.8}$$

for all $A\in\mathfrak{U}$. With the use of (2.5c) a simple estimate shows that $\omega_+^{\mu h}(A) - \omega_-^{\mu h}(A) = O(\mu)$ so that (2.7) does not seem to be a very severe assumption.

The conditions (2.6) and (2.7) are in fact the same as the ones introduced by Kastler[4] and Hoekman[5] because there $\omega^{\mu h}$ is a perturbed state which is invariant for the perturbed dynamics. In Ref. 5 a perturbed dynamics α_t^h is considered that is an Abelian group of transformations. As a consequence, the perturbed state could be explicitly constructed, viz. $\omega^{\mu h}(A) = \mathfrak{M}_t\omega(\alpha_t^{\mu h}A)$, where $\mathfrak{M}$ is an invariant mean over the additive group of the real numbers and t is a dummy variable.[8] If, in addition, one has that $(\mathfrak{U}_0,\alpha_t)$ is L_1-asymptotically Abelian, then the convergence (2.8) can be derived.

We shall now proceed with the demonstration that without loss of generality the perturbed state $\omega^{\mu h}$ may be assumed to be *approximately invariant* for the perturbed dynamics.

Lemma 2.3: Let $\mathfrak{M}$ be an invariant mean over the additive group of the real numbers.[8] Then the time-averaged perturbed state $\mathfrak{M}_{t'}\omega^{\mu h}(\alpha_{t'}A)\equiv\mathfrak{M}\omega^{\mu h}(A)$, where t' is to be considered as a dummy variable, satisfies

$$\lim_{\mu\to 0}\mathfrak{M}\omega^{\mu h}(\alpha_t A) = \mathfrak{M}\omega(\alpha_t A), \text{ uniformly in } t \tag{2.9a}$$

and

$$\lim_{\mu\to 0}[\,\mathfrak{M}\omega^{\mu h}(\tilde{\alpha}_t^h A) - \mathfrak{M}\omega^{\mu h}(A)\,] = 0, \quad \text{uniformly in } t. \tag{2.9b}$$

Proof: Because $\mathfrak{M}$ is an invariant mean we have[8]

$$|\mathfrak{M}[\,\omega^{\mu h}(\alpha_t A) - \omega(\alpha_t A)\,]| \leqslant \sup_{t\in\mathscr{R}}|\omega^{\mu h}(\alpha_t A) - \omega(\alpha_t A)| < \epsilon,$$

for $\mu < \mu_0(\epsilon)$. This establishes the continuity property of the time averaging. Similarly, with the use of (2.5c) we obtain

$$\begin{aligned}|\mathfrak{M}[\,\omega^{\mu h}(\tilde{\alpha}_t^{\mu h}A) - \omega^{\mu h}(A)\,]| &= |\mathfrak{M}[\,\omega^{\mu h}(\tilde{\alpha}_t^{\mu h}A) - \omega^{\mu h}(\alpha_t A)\,]| \\ &\leqslant \|\omega^{\mu h}\|\,\|\tilde{\alpha}_t^{\mu h}(A) - \alpha_t(A)\| \\ &\leqslant \|\omega^{\mu h}\|\,|\mu|\,\|h\|\,\|A\|(\|f_1\|_1 + \|f_2\|_1),\end{aligned}$$

so that (2.9b) follows from this estimate.

We shall denote the set of $(\check{f}_1,\check{f}_2)$-stable states by $I_{1,2}$. In order to study the consequences of $(\check{f}_1,\check{f}_2)$ stability we shall derive a condition which involves only the unperturbed entities ω and α_t and the functions f_1, f_2. Here $\check{f}(x)$ denotes $f(-x)$.

Proposition 2.4: Let $\omega\in I_{1,2}$ be continuous in the $\sigma(\mathfrak{U},N)$ topology. Then

$$\int_{-\infty}^{\infty} dt\, f_1(t)\omega(A\alpha_t B) = \int_{-\infty}^{\infty} dt\, f_2(t)\omega(\alpha_t(B)A), \quad \text{for } A,B\in\mathfrak{U}. \tag{2.10}$$

Proof: From Lemma 2.3 it follows that without loss of generality one may assume $\omega^{\mu h}$ to be approximately invariant for $\tilde{\alpha}_t^{\mu h}$, in the sense of (2.9b). For $h = h^*\in\mathfrak{U}$, $A\in\mathfrak{U}$ we write

$$\int_{T_1}^{T_2} dt\,\frac{d}{dt}[\,\omega^{\mu h}(\tilde{\gamma}_t^{\mu h}A)\,] = \omega^{\mu h}(\tilde{\gamma}_{T_2}^{\mu h}A) - \omega^{\mu h}(\tilde{\gamma}_{T_1}^{\mu h}A).$$

With the use of (2.5c) and (2.6) we find

$$\int_{-\infty}^{\infty} dt\,[\,\check{f}_1(t)\omega^{\mu h}(\tilde{\alpha}_t^{\mu h}(h\alpha_{-t}A)) - \check{f}_2(t)\omega^{\mu h}(\tilde{\alpha}_t^{\mu h}(\alpha_{-t}(A)h))\,] = (i/\mu)[\,\omega_+^{\mu h}(A) - \omega_-^{\mu h}(A)\,].$$

The right-hand side vanishes as $\mu\to 0$ due to (2.7). Because $f_i\in L_1(\mathscr{R})$ the Lebesgue dominated convergence theorem yields for $\mu\to 0$,

$$\int_{\infty}^{\infty} dt\, f_1(t)\omega(h\alpha_t A) = \int_{\infty}^{\infty} dt\, f_2(t)\omega(\alpha_t(A)h), \tag{*}$$

for $h = h^*\in\mathfrak{U}$ and $A\in\mathfrak{U}$. Now consider the GNS representation $(\mathfrak{h}_\omega,\pi_\omega,\Omega_\omega)$ associated with the state ω. Owing to Kaplansky's density theorem $\pi_\omega(h) = \pi_\omega(h)^*$ can be approximated strongly by a net $h_\alpha\in\pi_\omega(\mathfrak{U}_0)$ of self-adjoint elements.

With the help of a three-ϵ argument and the polarization method we can now extend (*) to all $h\in\mathfrak{U}$.

Throughout this paper we shall adopt the $\sigma(\mathfrak{U},N)$ continuity which we assumed in the preceding proposition.

III. INVARIANCE, SEPARATING CHARACTER, AND THE MODULAR GROUP

From the stability condition (2.11) we now explore the ensuing properties of a state $\omega\in I_{1,2}$. For a state to be stable it should at least be time invariant; i.e., invariant for α_t. To deal with this problem we formulate the following.

Lemma 3.1: (See Ref. 1.) Let F be a bounded function of two variables and $h\in L_1(\mathscr{R}_2)$. If $F(h)=\int ds\,dt\,F(s,t)h(s,t)$ vanishes for all h with $\hat{h}(p,q)$ having compact support not containing $q=0$ and $h(s,t)$ is differentiable with respect to t, with $\partial h(s,t)/\partial t\in L_1(\mathscr{R}_2)$, then

$$F(s,t)=G(s),$$

for some bounded function G. Now we are able to prove the desired invariance.

Proposition 3.2: If $\omega\in I_{1,2}$ then ω is invariant for the unperturbed dynamics α_t.

Proof: Let $A=\mathbb{1}$ and $B=C_g=\int dt\,g(t)\alpha_t(C)$, then (2.11) yields

$$\omega(C_{h_1}-C_{h_2})=\omega(C_{h_1-h_2})=0, \tag{3.1}$$

for $C\in\mathfrak{U}_0$, $\hat{g}\in D$, and $h_j=f_j*g$. From Lemma 3.1 we now conclude that $\omega(\alpha_t(C))$ is a constant for all $C\in\mathfrak{U}_0$. Invoking the continuity of ω yields invariance, viz. $\omega\circ\alpha_t=\omega$.

To proceed further it is now convenient to write the stability condition (2.11) in the GNS representation. Let $(\mathfrak{h},\pi,\Omega)$ be the GNS triple associated with $\omega\in I_{1,2}$. Since ω is invariant, the group of * automorphisms $\{\alpha_t\}$ can be implemented by a strongly continuous group of unitaries on $\mathfrak{h}$. To this end we must also assume that the correlation functions $t\to\omega(A\alpha_t B)$ are continuous. Explicitly, we then have $\pi(\alpha_t(A))=U_t\pi(A)U_{-t}$ and $U_t\Omega=\Omega$. The stability criterion (2.11) can now be written as

$$\int dt f_1(t)(\Omega,AU_tB\Omega)=\int dt f_2(t)(\Omega,BU_{-t}A\Omega), \tag{3.2}$$

for $A,B\in\pi(\mathfrak{U})''$.

The infinitesimal generator of U_t, i.e., the Liouville operator, will be denoted by L, with the spectral representation $L=\int dE_\lambda\lambda$.

Proposition 3.3: If $\omega\in I_{1,2}$ then Ω is separating.

Proof: From (3.2) we have

$$(\Omega,A\check{\hat{f}}_1(L)B\Omega)=(\Omega,B\hat{f}_2(L)A\Omega), \tag{3.3}$$

where $\check{f}(\lambda)=f(-\lambda)$. Let $A\Omega=0$ then

$$(\Omega,A\check{\hat{f}}_1(L)B\Omega)=0,$$

so that

$$(\hat{f}_1(L)A^*\Omega,B\Omega)=0, \tag{3.4}$$

for all $B\in\pi(\mathfrak{U})''$. Because $\hat{f}_1$ is invertible op sp $\alpha=\{\lambda\,|\hat{g}(\lambda)=0\ \forall g\colon \int g(t)\alpha_t(A)dt=0\ \forall A\in\mathfrak{U}_0\}\supset\{\lambda\,|\hat{g}(\lambda)=0\ \forall g\colon \int g(t)U_tA\Omega\,dt=0\ \forall A\in\pi(\mathfrak{U}_0)\}=\text{sp}\,L$, f_1 is also invertible on sp L. Since Ω is cyclic it follows now from (3.4) that $A^*\Omega=0$ and therefore $A=0$ by standard arguments.[9]

As a consequence we have that the state $(\Omega,\cdot\,\Omega)$ is a Tomita state on $\pi(\mathfrak{U})''$ with the modular automorphism $\sigma_t(\cdot)=\Delta^{it}\cdot\Delta^{-it}$, where Δ is the modular operator.

The preceding results already show great similarity of the (f_1,f_2)-stable states with thermodynamic equilibrium states. We shall make this connection more explicit in the following.

Theorem 3.4: If $\omega\in I_{1,2}$ with $f_{1,2}$ such that $\phi(\lambda)=\hat{f}_1(\lambda)/\hat{f}_2(\lambda)$ satisfies

(i) $\phi(\lambda)\check{\phi}(\lambda)=1$, with $\check{\phi}(\lambda)=\phi(-\lambda)$;

(ii) $\phi(\lambda)>0$ for $\lambda\in\text{sp}\,L$;

then $\omega\in K_\phi$, i.e., ω is a ϕ-KMS state.

Whenever (i) and (ii) are not fulfilled $I_{12}=\varnothing$. Conversely, if $\omega\in K_\phi$ then $\omega\in I_{1,2}$ for some nonunique $f_{1,2}$ such that $\hat{f}_1/\hat{f}_2=\phi$.

Proof: Suppose $\omega\in I_{1,2}$, then from (3.3) it follows that

$$(A^*\Omega,\check{\hat{f}}_1(L)B\Omega)=(B^*\Omega,\hat{f}_2(L)A\Omega).$$

Now choose $A=A_g$ with $\hat{g}=\bar{\check{\hat{f}}}_1$ then

$$\begin{aligned}&(\check{\hat{f}}_1(L)A^*\Omega,\check{\hat{f}}_1(L)B\Omega)\\&\quad=(B^*\Omega,\hat{f}_2(L)\bar{\hat{f}}_1(L)A\Omega)\\&\quad=(\bar{\hat{f}}_1(L)B^*\Omega,[\hat{f}_2(L)]/[\hat{f}_1(L)]\bar{\hat{f}}_1(L)A\Omega).\end{aligned}$$

Furthermore, we may let $B=A^*$ and since we assumed $\phi=\hat{f}_1/\hat{f}_2>0$ on sp L we have

$$\|\bar{\hat{f}}_1(L)A^*\Omega\|=\|\check{\phi}(L)^{1/2}\bar{\hat{f}}_1(L)A\Omega\|. \tag{3.5}$$

Since $\hat{f}_1(L)$ is invertible we can use the same reasoning as in Ref. 5 to conclude from (3.5) that the modular operator Δ can be written

$$\Delta^2=\check{\phi}(L)=\left[\frac{\hat{f}_1(L)}{\hat{f}_2(L)}\right]^{\check{}}=\frac{\hat{f}_2(L)}{\hat{f}_1(L)}. \tag{3.6}$$

It was shown in Ref. 2 that (3.6) is equivalent with $\omega\in K_\phi$.

The proof of the converse is quite easy. If $\omega\in K_\phi$ then

$$\int dt f_\phi(t)\omega(A\alpha_tB)=\int dt f(t)\omega(\alpha_t(B)A), \tag{3.7}$$

for all $A,B\in\mathfrak{U}$, $\hat{f}\in D$, and $\hat{f}_\phi=\phi\hat{f}$, with $\phi\in C^\infty(\mathscr{R})$. Now choose $\hat{F}\neq0$ on sp L and a sequence $(\hat{g}_n)_{n=1}^\infty$ in D, such that $\hat{g}_n\to\hat{F}$ and $\phi\hat{g}_n\to\hat{F}_\phi$ in S. As ϕ may have an essential singularity at infinity, owing to a theorem of Weierstrass, we can write $\phi=\exp(g)$. Here g is odd and finite in the finite complex plane. Now choose a function h with a Laurent expansion such that $\psi=\exp(h)\in S$ and $\phi\psi\in S$. Then we have for any $\hat{G}\in S$ that $\hat{F}=\psi\hat{G}\in S$ and $\phi\hat{F}=(\phi\psi)\hat{G}\in S$. Then it follows that ω satisfies (2.11), with $f_1=F_\phi$ and $f_2=F$, where $F,F_\phi\in L_1(\mathscr{R})\cap C^1(\mathscr{R})$. Obviously we have $\hat{f}_1/\hat{f}_2=\phi$; and since $\check{\phi}(L)=\Delta^2$, (3.7) can only be satisfied if on sp L $\phi>0$ and $\phi\check{\phi}=1$ (Ref. 2, Lemma 4).

We conclude with a further remark which can now be made regarding the set $I_{1,2}$.

Remark 3.5: From (3.6) it follows that the modular operator Δ commutes with the Liouville operator L. Then one may follow the line of reasoning given in Ref. 10 to establish that $I_{1,2}$ is a lattice in its own order. In general $I_{1,2}$ will not be closed and hence *a fortiori* not compact. If one assumes in addition the compactness of $I_{1,2}$ in the w^* topology, then it follows that $I_{1,2}$ is a Chôquet simplex.

ACKNOWLEDGMENT

Part of this work was done while one of us (J.S.C.) was at the Institute for Theoretical Physics of the University of Amsterdam as a member of the scientific staff of the "Stichting Fundamenteel Onderzoek der Materie" (FOM), which is financially supported by the "Nederlandse Organisatie voor Zuiver Wetenschappelijk Onderzoek" (ZWO).

[1]O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics* (Springer, Berlin, 1979, 1981).

[2]J. S. Cohen, H. A. M. Daniëls, and M. Winnink, Commun. Math. Phys. **84**, 449 (1982).

[3]J. L. Lebowitz, M. Aizenman, and S. Goldstein, J. Math. Phys. **16**, 1284 (1975).

[4]D. Kastler, Symp. Math. **20**, 49 (1976).

[5]F. Hoekman, Ph.D. thesis, University of Groningen, 1977.

[6]W. Pusz and S. L. Woronowicz, Commun. Math. Phys. **58**, 273 (1978).

[7]G. ten Brinke and M. Winnink, Commun. Math. Phys. **51**, 135 (1976).

[8]F. P. Greenleaf, *Invariant Means on Topological Groups* (Van Nostrand, New York, 1969).

[9]M. Winnink, "Some general properties of thermodynamic states in an algebraic approach," in *Statistical Mechanics and Field Theory*, edited by R. N. Sen and C. Weil (Keter, Jerusalem, 1971).

[10]M. Takesaki and M. Winnink, Commun. Math. Phys. **30**, 129 (1973).

Nonequilibrium entropy in classical and quantum field theory

Henry E. Kandrup[a)]
Center for Theoretical Physics, University of Maryland, College Park, Maryland 20742

(Received 31 July 1986; accepted for publication 18 February 1987)

This paper proposes a definition of nonequilibrium entropy appropriate for a bosonic classical or quantum field, viewed as a collection of oscillators with equations of motion which satisfy a Liouville theorem (as is guaranteed for a Hamiltonian system). This entropy S is constructed explicitly to provide a measure of correlations and, as such, is conserved absolutely in the absence of couplings between degrees of freedom. This means, e.g., that there can be no entropy generation for a source-free linear field in flat space, but that S need no longer be conserved in the presence of couplings induced by nonlinearities, material sources, or a nontrivial dynamical background space-time. Moreover, through the introduction of a "subdynamics," it is proved that, in the presence of such couplings, the entropy will satisfy an H-theorem inequality, at least in one particular limit. Specifically, if at some initial time t_0 the field is free of any correlations, it then follows rigorously that, at time $t_0 + \Delta t$, the entropy will be increasing: $dS/dt > 0$. Similar arguments demonstrate that this S is the only measure of "entropy" consistent mathematically with the subdynamics. It is argued that this entropy possesses an intrinsic physical meaning, this meaning being especially clear in the context of a quantum theory, where a direct connection exists between entropy generation and particle creation. Reasonable conjectures regarding the more general time dependence of the entropy, which parallel closely the conventional wisdom of particle mechanics, lead to an interpretation of S which corroborates one's naive intuition as to the behavior of an "entropy."

I. INTRODUCTION

Conventional wisdom holds that the "entropy" associated with some system should be interpreted probabilistically as a measure of how generic its state really is. A state which is comparatively random, and which could be realized in many different ways, has associated with it a large entropy; a state which is somehow improbable, requiring, e.g., a special preparation, is considered to have low entropy. "Equilibrium" is, in this context, interpreted as a state of "maximum randomness" and "maximum entropy." The content of Boltzmann's[1] classic H-theorem is that a system will evolve towards this state of maximum entropy.

There is, however, a well-known difficulty in implementing this general picture. Consider, for example, a collection of N classical point particles. Suppose in the usual way that this system is characterized by an N-particle distribution function μ and that the evolution of this μ is governed by an N-particle Liouville equation which expresses probability conservation. The standard paradigm then implies (i) that the entropy $S = -\mathrm{Tr}\,\mu \log \mu$, where Tr denotes a trace over the degrees of freedom of the N particles, and (ii) that the unique equilibrium corresponds to a state $\mu \propto \exp(-\beta H)$, where β is a constant and H the N-particle Hamiltonian. The problem, however, is that the Liouville equation guarantees that $dS/dt \equiv 0$! This S does not change with time and, consequently, there can be no systematic evolution towards an equilibrium state of maximum entropy.

This well-known difficulty led historically to the idea of a "coarse-grained averaging," namely the notion that an H-theorem does not hold on a truly microscopic level, but, instead, holds only on a quasimacroscopic level for an appropriately averaged μ. This is, at least superficially, an extremely attractive idea, but it suffers from two related drawbacks: (1) it seems difficult to generate a general algorithm to effect the desired coarse graining; and (2) even if one were to construct a working algorithm, one would be faced with the additional problem of demonstrating that it is "canonical" in some natural sense. There would remain, e.g., the task of either ascertaining the scale on which the averaging is to be implemented, or, alternatively, of demonstrating the scale invariance of the averaging.

To the extent that no canonical prescription exists, one seems forced ultimately to the viewpoint adopted by Jaynes, namely that "Entropy is a property, not of the physical system, but of the particular experiments you or I choose to perform on it."[2] It is, therefore, natural to ask what it is that one typically measures when one probes the state of the system. And the answer to that would appear quite clear. One seeks typically to measure the one-particle distribution function f (i.e., the probability density for finding a particle at a given point $\mathbf{x}$ with momentum $\mathbf{p}$ at time t), or perhaps the pair or three-body correlation functions. But one does not even try to measure the detailed correlations amongst the particles buried in the full N-particle μ. One might, therefore, argue that, as a practical matter, the physically relevant notion of coarse graining does not involve a macroscopic averaging but, instead, entails a loss of information about higher-order interparticle correlations.

From this point of view, it would seem natural to conjecture that the entropy of the system should be defined in terms of the reduced one-particle f, rather than the full N-particle μ. And, as such, it would be natural to propose an entropy

a) Present address: Department of Physics, Syracuse University, Syracuse, New York 13244.

 0022-2488/87/061398-18$02.50

$$S(t) \equiv -\sum_i \int d\mathbf{x}_i\, d\mathbf{p}_i\, f(i) \log f(i), \tag{1.1}$$

where

$$f(i) \equiv f(\mathbf{x}_i, \mathbf{p}_i, t) = \int \prod_{j \neq i} d\mathbf{x}_j\, d\mathbf{p}_j\, \mu(\mathbf{x}_1, \mathbf{p}_1, \ldots, \mathbf{x}_N, \mathbf{p}_N; t). \tag{1.2}$$

There are solid reasons to believe that this S provides a reasonable notion of entropy. Most obvious is the fact that this S will in general be time dependent. Thus, for example, allowing for particle interactions derived from two-body potentials V_{ij}, one concludes that, for a system of N identical particles,

$$\frac{dS}{dt} = N(N-1) \int d\mathbf{x}_i\, d\mathbf{p}_i \int d\mathbf{x}_j\, d\mathbf{p}_j\, f_2(i,j) \times \frac{\partial V_{ij}}{\partial \mathbf{x}_i} \cdot \frac{\partial}{\partial \mathbf{p}_i} \log f(i), \tag{1.3}$$

where $f_2(i,j)$ is the reduced two-particle distribution function for particles i and j.

Also significant is the fact that this S is truly canonical, being constructed in a systematic and unambiguous fashion from the one-particle f, an object of obvious physical significance. Of particular relevance in this regard is the fact that this f satisfies a "subdynamics," [3] decoupled from the higher-order correlations. Specifically, by means of projection operator techniques, one can derive for the evolution of f an exact, closed (albeit nonlocal and nonlinear) equation which contains no explicit reference to the higher-order correlations buried in such quantities as $f_2(i,j)$. It is the nonlinearity of this equation which leads to a nonconserved entropy; it is the linearity of the fundamental Liouville equation that guarantees that $\mathrm{Tr}\, \mu \log \mu$ is a constant of the motion.

This S coincides, moreover, with the entropy entering into the standard H-theorem; and, consequently, one anticipates that this S really will increase monotonically at least in some approximate limit. It is in fact well known that the exact equation satisfied by f reduces to the standard Landau equation[3] if one assumes (i) that the interactions are weak and comparatively short range (dilute gas approximation), (ii) that initial conditions were specified at a time in the past long compared with the duration of a typical interaction, and (iii) that one can neglect the effects of nontrivial initial conditions and suppose that, at some initial time t_0, $\mu = \Pi_i f(i)$.

The first two of these requirements seem reasonable physically; and indeed, they can be relaxed, at least in principle, in the context of a systematic perturbation expansion. The third requirement needs some further justification. For "ordinary," reasonably well behaved interactions, one can argue convincingly, and in certain cases prove,[3] that generic nontrivial initial conditions will in fact decay as time goes by. The physical content of this statement is that any correlations present at the outset will eventually become irrelevant compared with the systematically evolving correlations generated by the subsequent dynamics. For more perverse, long-range interactions, like Newtonian gravity, this argument is most likely invalid: numerical simulations suggest that self-gravitating systems really do "remember" their initial conditions for a very long time.[4] For these perverse interactions, a systematic statistical analysis really requires the ansatz that no significant initial correlations be present. Whether this is reasonable is of course exceedingly difficult to say. Such a viewpoint is, however, if nothing else, consistent with the speculation that the Universe originated from a state of "maximum simplicity." After all, it is only on very large scales that self-gravity becomes important.

It should also be stressed that, even neglecting "special" initial conditions, not every physical system of interacting entities will satisfy an H-theorem for all times. Consider, for example, a pair of harmonic oscillators with natural frequencies ω_1 and ω_2, connected by a linear coupling derived from a pair potential $V_{12} = \lambda x_1 x_2$, where λ measures the strength of the interaction. In this simple case, one can solve explicitly for the evolution of the system in terms of arbitrary initial conditions; and, for appropriate choices of ω_1, ω_2, and λ, it is easy to see that the motion will be periodic. After some time τ, the system will return to its initial state: there can be no progression towards equilibrium. This means that, if the entropy increases at one point of time, it must at some other time decrease.

This does not, however, imply that the entropy (1.1) is devoid of physical meaning. Even if this S does not satisfy an H-theorem, it can provide a useful measure of the degree of correlations in the system. Thus, e.g., if one supposes that, at some initial time t_0, the system was free of correlations, the interactions between the oscillators will result in an initial generation of correlations and a concomitant increase in entropy!

This initial entropy increase is in fact a very general result. Consider a collection of N objects which interact via arbitrary two- and higher-body forces characterized by a coupling constant λ. Assume then that, at some initial time t_0, the system was completely free of correlations, so that $\mu = \Pi_i f(i)$. It follows that, an instant Δt later,

$$\frac{dS(t_0 + \Delta t)}{dt} = \lambda^2 |a|^2 \Delta t > 0, \tag{1.4}$$

where $|a|^2$, which is intrinsically positive, reflects the form of the initial state. A proof of this claim for a specific model interaction was provided in Ref. 5. A more general proof will be provided below. The important point to note here is that, even if an H-theorem does not hold for all times, the entropy (2.1) can still provide a useful measure of correlations. To show that S reflects correlations, and to show that S always increases, are two distinct and separate issues.

To the extent that a system is truly periodic and no systematic evolution towards a more "random" state is obtained, one might perhaps argue that S does not warrant the appellation entropy. What does, however, appear to be an empirical fact is that, for realistic complicated systems, couplings between degrees of freedom lead to a progression towards a more random sort of state.

The object of this paper is to formulate a notion of nonequilibrium entropy for a classical or quantum field which parallels as closely as possible the particle entropy described in this Introduction. This notion of entropy is such that it is conserved absolutely for a source-free linear field in Min-

kowski space, but assumes a time dependence in the presence of couplings induced, e.g., by nonlinearities, sources, or a nontrivial background space-time.

The program of this paper is as follows: Section II sketches out the basic picture for a classical or quantum field in Minkowski space, emphasizing in particular (i) the question of physical interpretation and (ii) the formal similarities between the classical and quantum theories. Section III then addresses the additional complications, both conceptual and practical, which arise for fields propagating in a nontrivial dynamical space-time. Section IV introduces the notion of a subdynamics to show explicitly that the field entropy S defined in Secs. II and III does in fact provide a measure of correlations, demonstrating that, in the presence of couplings between degrees of freedom, an initially uncorrelated state leads necessarily to an initial increase in entropy. Section V turns to the special question of physical interpretation in the context of a quantum theory, focusing upon a fundamental connection between changes in the field entropy and the phenomenon of particle creation. The principal conclusion here is that the mechanism which gives rise to an initial increase in entropy will also cause an initial enhancement in the rate of particle creation. Section VI demonstrates that, in a precise and well-defined mathematical sense, the S defined in Secs. II and III is the only measure of entropy consistent with the notion of subdynamics defined in Sec. IV. Finally, Sec. VII summarizes the principal results, speculates upon the more generic time dependence of S, and concludes by reflecting upon the connection between the entropy defined here and the "geometric" entropy associated, e.g., with the event horizon of a black hole in general relativity.

It should, perhaps, be noted that, although the analysis presented here is comparatively abstract, the key ideas were motivated originally by a desire to understand nonequilibrium processes in the early Universe. The sense in which these ideas are relevant there is considered in a companion paper by Hu and Kandrup.[6]

Units are chosen throughout such that Planck's constant $h/2\pi$ and the speed of light c are equal to unity.

II. NONEQUILIBRIUM ENTROPY IN MINKOWSKI SPACE

A. Classical field theory

As noted already, the object of this paper is to construct a notion of entropy for a classical or quantum field which parallels as closely as possible the notion of particle entropy posited in Sec. I. This implies, in particular, three specific requirements.

(1) This entropy S must provide a measure of the degree of correlations in the system. If there are no couplings between degrees of freedom, so that no correlations can be generated, S must be conserved.

(2) This S must satisfy an H-theorem inequality, at least in some appropriate limit. One demands that an "initially uncorrelated" system leads to an initial increase in S; and one anticipates (or at least hopes) that S will increase monotonically for all times, this corresponding to an approach towards equilibrium.

(3) This S must in some sense be physically meaningful.

The key idea underlying the analysis is that the field entropy S should be constructed from an appropriate analog of the one-particle distribution functions $f(i)$. One knows that, at least in flat space, a classical field theory is equivalent mathematically to an infinite set of oscillators. The statistical description of such a field takes, therefore, as its fundamental object a many-oscillator distribution function μ, the evolution of which is governed by a Liouville equation which expresses conservation of probability. Given this fundamental μ, one can then define reduced distribution functions $g(q_A,\pi_A,t)\equiv g(A)$ in the obvious way for each oscillator by integrating over the "position" and "momentum," q_B and π_B, of each of the remaining oscillators:

$$g(A)\equiv g(q_A,\pi_B,t)=\int \prod_{B\neq A} dq_B\, d\pi_B\, \mu(\{q_C,\pi_C\};t). \tag{2.1}$$

Given these one-oscillator $g(A)$'s, one is then instructed to define an entropy

$$S(t)=-\sum_A \int dq_A\, d\pi_A\, g(A)\log g(A). \tag{2.2}$$

Because the field is equivalent to an infinite set of oscillators, the sum in this expression is an infinite one, so that the entropy so defined might well prove infinite, at least formally. This, however, is not especially relevant for the present discussion. What is relevant is whether this S can be shown to increase in the presence of evolving correlations.

The first obvious point to note is that, although, in the absence of material sources, $\mathrm{Tr}\,\mu\log\mu$ is a constant of the motion, the entropy of Eq. (2.2) will in fact be time dependent if there exist couplings between degrees of freedom. It follows trivially that, for a linear free field, $dS/dt\equiv 0$, but, in the presence of nonlinearities such as those arising in a $\lambda\Phi^p$ field theory, one concludes instead that, in general, $dS/dt\neq 0$.[5] This is no different from the statement that a system of N noninteracting particles must conserve its entropy, but that, once one allows for particle interactions, the entropy will in fact change with time.

One can, moreover, demonstrate that, in the presence of such ccouplings induced by nonlinearities, S will satisfy for short times the same sort of H-theorem as did the particle entropy of Sec. I. Specifically, if one supposes that, at some initial time t_0, the system was free of correlations, so that $\mu=\Pi_A g(A)$, it follows rigorously that $dS(t_0+\Delta t)/dt>0$. The proof of this statement, provided in Sec. IV, is the same for particle and field theories. The crucial point is simply that, for early times, dS/dt will be quadratic in the interaction Liouvillian which generates the evolving correlations! For a system of particles interacting via two-body forces, the interaction Liouvillian for a pair of particles i and j takes the form

$$L^I_{ij}=\lambda\frac{\partial V_{ij}}{\partial \mathbf{x}_i}\cdot\frac{\partial}{\partial \mathbf{p}_i}, \tag{2.3}$$

where λ is a coupling constant. For a simple $\lambda\Phi^3$ field theory, the analogous object is

$$\mathcal{L}^I_{AB} = \lambda q_B q_{A-B} \frac{\partial}{\partial \pi_A}. \tag{2.4}$$

In each case, one concludes that, to first order in Δt,

$$\frac{dS(t_0+\Delta t)}{dt} = \lambda^2 |a|^2 \Delta t > 0, \tag{2.5}$$

where $|a|^2 > 0$ depends upon the form of the initial state.

In the presence of material sources (e.g., when considering a collection of charged particles interacting via an electromagnetic field) the analysis becomes more complicated mathematically, but the physical picture remains unchanged. The fundamental object for the composite system of particles and fields is now an enlarged distribution function ν depending upon both particle and field variables which satisfies a Liouville equation expressing probability conservation in an enlarged particle-plus-field phase space. It follows trivially from the linearity of the Liouville equation that Tr $\nu \log \nu$ is conserved absolutely, but that the total entropy

$$S \equiv -\sum_i \int d\mathbf{x}_i\, d\mathbf{p}_i\, f(i) \log f(i) - \sum_A \int dq_A\, d\pi_A\, g(A) \log g(A) \tag{2.6}$$

will in general exhibit a nontrivial time dependence. Local equations for $\partial f(i)/\partial t$ or $\partial g(A)/\partial t$ analogous to Eq. (1.3) now involve the particle-oscillator correlation function

$$h(i,A) = \int \prod_{j\neq i} d\mathbf{x}_j\, d\mathbf{p}_j \int \prod_{B\neq A} dq_B\, d\pi_B \times \nu(\mathbf{x}_1,\mathbf{p}_1,\ldots,\mathbf{x}_N,\mathbf{p}_N,\{q_C,\pi_C\};t). \tag{2.7}$$

And the couplings buried in this $h(i,A)$ induce a nontrivial dS/dt just as surely as do the couplings associated with the direct particle–particle interactions of ordinary Newtonian dynamics or the nonlinearities discussed above.

Since the physical picture here is not different from that arising in a nonlinear field theory, but significantly messier mathematically, this situation will not be considered in any detail in this paper. One may, however, note that it is completely straightforward to use the techniques developed in Ref. 7 to parallel the discussion here and obtain a theory of nonequilibrium entropy for an interacting system of particles and fields.

It remains here to at least address the issue of whether the field entropy (2.2) is physically meaningful. After all, the oscillators that one is considering are only mathematical constructs, and one would not expect to measure any $g(A)$ in a realistic experiment. One reason to believe that the $g(A)$'s, and hence S, are meaningful is that the $g(A)$'s serve at least to define physically measurable average values. Thus, for example, if one considers a scalar field

$$\Phi(\mathbf{x},t) \equiv \sum_k q_k(t) \exp(-i\mathbf{k}\cdot\mathbf{x}), \tag{2.8}$$

it follows immediately that the statistical average value

$$\langle \Phi(\mathbf{x},t) \rangle = \sum_k \langle q_k(t) \rangle \exp(-i\mathbf{k}\cdot\mathbf{x}), \tag{2.9}$$

where

$$\langle q_k(t) \rangle = \mathrm{Tr}\, \mu q_k(t) = \int dq_k\, d\pi_k\, g(k) q_k(t). \tag{2.10}$$

This is completely analogous to the situation in a particle theory. Here, for example, the true mass density

$$\rho(\mathbf{x},t) = \sum_i m_i \delta_D[\mathbf{x} - \mathbf{x}_i(t)], \tag{2.11}$$

whereas the average density

$$\langle \rho(\mathbf{x},t) \rangle = \sum_i m_i \langle n_i(\mathbf{x},t) \rangle, \tag{2.12}$$

with

$$\langle n_i(\mathbf{x},t) \rangle = \mathrm{Tr}\, \mu \delta_D(\mathbf{x} - \mathbf{x}_i) = \int d\mathbf{x}_i\, d\mathbf{p}_i\, f(i) \delta_D[\mathbf{x} - \mathbf{x}_i(t)]. \tag{2.13}$$

One should note, moreover, that if an H-theorem holds, S can provide a useful diagnostic for the evolution of physical observables. For a free, linear field, each mode will of course evolve independently, so that the entropy, which is after all conserved, will not be of particular significance. Thus a scalar field satisfying the Klein–Gordon equation

$$-\partial_t^2 \Phi + \Delta \Phi = 0 \tag{2.14}$$

leads trivially to the averaged equation

$$-\partial_t^2 \langle \Phi \rangle + \Delta \langle \Phi \rangle = 0. \tag{2.15}$$

If, however, there exist correlations generated by couplings between the modes, one acquires an effective "source" Σ, and the evolving S can provide useful information about the effects of this Σ. Thus, for example, the true nonlinear equation

$$-\partial_t^2 + \Delta \Phi + \lambda \Phi^2 = 0 \tag{2.16}$$

leads to a statistically averaged equation

$$-\partial_t^2 \langle \Phi \rangle + \Delta \langle \Phi \rangle + \lambda \langle \Phi \rangle^2 = \Sigma, \tag{2.17}$$

where the source Σ involves couplings between each mode A and the modes B and $A - B$.[5]

B. Quantum field theory

The discussion hitherto has focused exclusively upon a classical field theory. It is, however, completely straightforward to formulate a corresponding quantum theory for a bosonic system with integral spin, provided only that the fundamental dynamics can be cast into a Hamiltonian form[8] (the generalization to a spin-$\frac{1}{2}$ fermionic system is currently under investigation[9]). All that one need do is implement the standard formalism of canonical quantization, assuming that the Poisson bracket is to be replaced by a commutator. Indeed, one could equally well construct a quantum statistical theory for a non-Hamiltonian system provided only that one is willing to accept *some* (perhaps quite *ad hoc*) quantization prescription.

It is, however, important to emphasize that the interpretation of the field entropy (2.2) as providing a measure of correlations *does* rely upon one important feature guaranteed for Hamiltonian systems which will not, however, hold in general, namely the notion of conservation of phase expressed by the Liouville theorem. Consider, e.g., a collection

of N identical classical, noninteracting particles which feel the influence of an external force F_a which depends upon the coordinates, momenta, and time in a perverse fashion which cannot be formulated in a Hamiltonian framework. In this case, one still has a well-defined notion of probability conservation in an appropriate $6N$-dimensional $(\mathbf{x},\mathbf{p})$ space, and it follows at once that the N-particle μ will satisfy the relation

$$\frac{\partial\mu}{\partial t}+\sum_{i=1}^{N}\frac{\partial}{\partial x_i{}^a}\left(\frac{dx_i{}^a}{dt}\mu\right)+\sum_{i=1}^{N}\frac{\partial}{\partial p^i{}_a}\left(\frac{dp^i{}_a}{dt}\mu\right)=0, \tag{2.18}$$

so that, since the particles are noninteracting,

$$\frac{\partial f}{\partial t}+\frac{p^a}{m}\frac{\partial f}{\partial x^a}+\frac{\partial}{\partial p_a}(F_a f)=0. \tag{2.19}$$

It does not, however, follow in the usual way that the Boltzmann entropy will be conserved! Rather, one calculates explicitly that

$$\frac{dS}{dt}=N\int d\mathbf{x}\,d\mathbf{p}\,\frac{\partial F_a}{\partial p_a}f(1+\log f), \tag{2.20}$$

which certainly need not vanish. In the absence of conservation of phase, which would be guaranteed if

$$\frac{\partial}{\partial x^a}\left(\frac{dx^a}{dt}\right)+\frac{\partial}{\partial p_a}\left(\frac{dp_a}{dt}\right)=0, \tag{2.21}$$

even a noninteracting theory entails a time-dependent entropy.

It may also be emphasized that this is not a purely technical observation with no physical relevance. A similar problem arises in general relativity if one wishes to reexpress geodesic flows in a space-time with metric $g'_{\alpha\beta}$ in terms of a slightly different metric $g_{\alpha\beta}$ (as would, e.g., be required for a phase space description of linearized perturbations away from some static background). In this case, the prescription of Israel and Kandrup[10] leads to a four-force

$$F_\alpha=\delta\Gamma^\lambda_{\mu\nu}\Delta_{\alpha\lambda}p^\mu p^\nu/m,$$

where $\Delta_{\alpha\lambda}$ is the spatial projection tensor constructed from p_α and $g_{\alpha\lambda}$, and $\delta\Gamma^\lambda_{\mu\nu}$ is the difference between the Christoffel symbols associated with $g'_{\alpha\beta}$ and $g_{\alpha\beta}$. It then follows that $\partial F_\alpha/\partial p_\alpha\neq 0$, and this implies that the entropy flux s^μ, as defined, e.g., by Israel,[11] will not be divergence-free.

In any case, at least for Hamiltonian systems, the only really new feature of a quantum statistical description is that the distribution function μ must be reinterpreted as a density matrix with an evolution governed by the usual quantum Liouville equation. Reduced distribution functions, such as $g(A)$, realized as reduced density matrices, are obtained by partial traces over the degrees of freedom of some subset of the oscillators. The field entropy S is then defined in terms of the reduced $g(A)$'s by the obvious prescription

$$S=-\sum_A \mathrm{Tr}_A\, g(A)\log g(A), \tag{2.22}$$

where Tr_A denotes a trace over the degrees of freedom of the A th oscillator, realized in an arbitrary (e.g., coordinate or momentum) representation.

As illustrated in Sec. IV, one can, either classically or in the framework of a quantum description, formulate a subdynamics for the evolution of the $g(A)$'s which contains no explicit reference to the higher-order correlations. And, by using this subdynamics, one can again prove an initial entropy increase (2.5) for a system which evidences no initial correlations. There remains the well-known problem of obtaining a true probabilistic interpretation of the quantum μ or $g(A)$'s, but this has no immediate bearing upon the interpretation of S as a measure of correlations.

Given the quantum or classical equations of motion, and the existence of a density matrix or distribution function μ which satisfies a linear Liouville equation, the quantum and classical statistical theories are essentially identical!

The quantum theory does, however, admit to one enrichment of interpretation. Specifically, in the context of a quantum description, it is conventional to interpret a change in the field as representing the creation or destruction of particles. Thus, in particular, if the A th oscillator is characterized by a natural frequency ω_A, it is customary to interpret the statistical average

$$\langle N_A\rangle\equiv\langle\tfrac{1}{2}\pi_A{}^2+\tfrac{1}{2}\omega_A{}^2q_A{}^2\rangle/\omega_A-\tfrac{1}{2} \tag{2.23}$$

as representing the "number of quanta in the A th mode." One might, therefore, seek to establish a connection between changes in the entropy (2.22) and changes in the average $\langle N_A\rangle$. The gratifying fact, discussed in great detail by Hu and Kandrup,[6] and considered briefly in Sec. V, is that such a connection does exist. Evolving correlations lead not only to an increase in the entropy, but to an increase in particle number.

That this is the case is illustrated by the following statement, a proof of which is presented in Sec. V. Consider a source-free nonlinear field theory in flat space realized as a collection of oscillators, and suppose that the $\langle N_A\rangle$ of Eq. (2.23) represents the number of particles in the A th mode. Write the total density matrix μ in the form

$$\mu=\prod_A g(A)+\mu_I, \tag{2.24}$$

where $\mu_I\equiv\mu-\Pi_A g(A)$ reflects the "piece" of the total μ which contains information about correlations amongst the degrees of freedom. Suppose then that, at some initial time t_0, μ_I vanishes identically. It follows rigorously that, at time $t_0+\Delta t$, the contribution to $d\langle N_A\rangle/dt$ which involves the generated $\mu_I(t_0+\Delta t)$ is intrinsically positive.

III. NONEQUILIBRIUM ENTROPY IN CURVED SPACE-TIMES

A. Conceptual issues

The further generalization of this basic picture to a quantum field theory in a fixed curved background space-time is again comparatively straightforward, at least formally, provided only that one can implement a preferred $3+1$ splitting into space and time, and that one specifies a preferred set of spatial functions at each instant of time to generalize the standard decomposition into plane waves. Provided that the space-time is not too perverse, the Cauchy problem will be well defined, and, given a consistent notion of dynamics, a statistical description should be possible. The only new significant technical complications arise from the fact (i)

that, in general, one cannot expand in spatial plane waves, and (ii) that, in a dynamical space-time, the three-Hamiltonian H derived from the fundamental action will manifest an explicit time dependence.

The fact that H is time dependent does not, in itself, generate any serious conceptual or practical problem.[12] A time-dependent H most likely precludes the possibility of an equilibrium,[11] but a nonequilibrium statistical description is still quite possible. At a conceptual level, field theory in a classical time-dependent gravitational field is not very different, in many respects, from quantum electrodynamics in a time-dependent, externally imposed, classical electromagnetic field.

Strictly speaking, the introduction of a specific $3+1$ splitting, or, equivalently, the specification of a family of preferred observers, arises already in Minkowski space. Thus ordinary quantum field theory is usually formulated from the point of view of inertial observers; and it is well known that, even for the simplest case of a source-free linear field theory, an accelerated observer will interpret the physics very differently. Thus, for example, the state which, to an inertial observer, corresponds to a true vacuum will, to a uniformly accelerated observer, correspond instead to a thermal state with a temperature proportional to its acceleration.[13]

One might, therefore, conjecture that in curved, as in flat, space one ought simply to restrict attention to freely falling observers. This, however, does not suffice to solve the problem. The symmetries of Minkowski space imply that any inertial observer will see "space" as homogeneous and isotropic, free of event horizons and other global complications, so that there is a natural decomposition of any field into plane waves $\exp(-i\mathbf{k}\cdot\mathbf{x})$. In curved spaces, however, there is in general no obvious analog of these plane waves, and two different freely falling observers might find it convenient to expand in two very different sets of "spatial" functions.

Given that the natural spatial functions are no longer necessarily plane waves, it is not especially convenient to introduce the concept of a Wigner function except in a type of "quasilocal" approximation. Indeed, the conventional noncovariant construction of such Wigner functions by means of the Weyl prescription exploits, in a deep and fundamental way, the space translational symmetry of Minkowski space.[14] And, similarly, the covariant Wigner functions of flat space exploit the time translational symmetry as well.[15]

This does not, however, imply that Wigner functions are totally useless in curved space-times. By assuming that the space-time is comparatively smooth, i.e., that the gravitational "field" does not change too quickly in space and time, one can use a Riemann normal coordinate construction to obtain a notion of the Wigner function which does provide useful insights into the overall dynamics. Thus, for example, one can show that a free scalar field may be characterized by a Wigner function $f_W(x^\alpha,p_\alpha)$ which, in a first approximation, satisfies the "collisionless Boltzmann" equation[16]

$$\frac{df_W}{d\tau}=p^\alpha\frac{\partial f}{\partial x^\alpha}+\Gamma^\lambda_{\alpha\mu}p_\lambda p^\mu\frac{\partial f}{\partial p_\alpha}=0, \tag{3.1}$$

the same equation which would be satisfied by a collection of noninteracting classical point masses. This is a very intuitive sort of relation, implying simply that f_W is conserved under Lie transport along the classical geodesics

$$\frac{dx^\alpha}{d\tau}=p^\alpha \quad\text{and}\quad \frac{dp_\alpha}{d\tau}=\Gamma^\lambda_{\alpha\mu}p_\lambda p^\mu. \tag{3.2}$$

The important point to observe, however, is that, whereas this equation is exact for classical point masses, it is only an approximation for the quantum field. A more careful analysis shows that[16]

$$\frac{df_W}{d\tau}=C[f_W], \tag{3.3}$$

where $C[f_W]$, which can be calculated perturbatively, involves the Riemann curvature tensor and derivatives thereof.

This is qualitatively similar to what happens to a charged particle in flat space. Here, classically, the one-particle $f(x^\alpha,p_\alpha)$ will satisfy exactly the equation

$$\frac{df}{d\tau}=p^\alpha\frac{\partial f}{\partial x^\alpha}+eF^\alpha{}_\mu p^\mu\frac{\partial f}{\partial p^\alpha}=0, \tag{3.4}$$

where e is the charge and $F_{\alpha\beta}$ the Maxwell tensor. However, the Wigner function $f_W(x^\alpha,p_\alpha)$ appropriate for a Dirac field will instead satisfy an equation of the form

$$\frac{df_W}{d\tau}=\mathscr{C}[f_W], \tag{3.5}$$

where $\mathscr{C}[f_W]$ involves the derivative $\nabla_\alpha F_{\mu\nu}$. In each case, one acquires corrections proportional to the "tidal forces" acting between nearby points which may be interpreted as reflecting the effects of "virtual particles."

For better or ill, the basic point of view adopted in this paper circumvents completely the notion of the Wigner function by working directly with the density matrix μ. One advantage therein is the fact that one has not built into the basic formalism any underlying assumptions that reflect the special symmetries of flat space. The density matrix μ has a natural meaning in and of itself in terms of the $3+1$ splitting and an arbitrary choice of spatial functions.

Another related feature of this description is that it is intrinsically nonlocal. The $g(A)$'s are defined in an abstract (q,π) space which has no direct connection with the space-time manifold or the associated cotangent bundle. This means that there is no natural sense in which some piece of the total entropy S can be associated with some piece of the space-time. The $g(A)$'s are dependent upon the choice of spatial functions which contain important information about the global properties of the space-time. As will be discussed in Sec. VII, this seems a desirable feature if one wishes to establish a connection between the ideas presented here and the geometric entropy associated, e.g., with a black hole in a static space-time. Here, following Bekenstein and Hawking,[17] one is wont to associate with the space-time an entropy proportional to the area of the event horizon, but it is clear that the existence of such an event horizon is manifest only in a global description of the physics.

This nonlocality does, however, have one perhaps undesirable implication, namely that it would seem difficult, if

not impossible, to construct a natural covariant generalization. The situation is, e.g., very different in relativistic kinetic theory.[12] There one is wont to associate with the distribution of matter in the space-time an entropy flux $s^\mu(x^\alpha)$, a vector field defined in the space-time manifold. The content of an H-theorem is then encapsulated in the covariant statement that the covariant divergence of s^μ is intrinsically non-negative: $\nabla_\mu s^\mu \geqslant 0$. Only by breaking manifest covariance and integrating over some arbitrary spacelike hypersurface does one reach the noncovariant (but observer-independent!) conclusion that $dS/dt \geqslant 0$.

In the absence of a covariant theory, it is not at all clear that the monotonic increase of the field entropy (2.22) is a statement with which all observers would agree. Indeed, one is confronted with two even more fundamental questions: (1) Will all observers agree that correlations do, or do not, exist between degrees of freedom for the field?; and (2) will they even agree whether couplings exist between these degrees of freedom? These, however, are issues which arise already in ordinary quantum field theory in curved space-time, and, especially, in analyses of particle creation, so that it is probably fair to say: In this regard, at least, the notion of entropy presented here is as reasonable—or unreasonable—as the standard discussions of particle creation.

In any case, to recapitulate: If one is willing to implement a preferred 3 + 1 splitting and a preferred decomposition into spatial functions, the formal structure of the statistical description is not very different from that arising in Minkowski space. What is different are the new sorts of physical implications induced by a nontrivial dynamical background space-time.

B. Three examples

1. A static space-time

In this case, there exists a natural 3 + 1 splitting, t being the coordinate associated with the time translational invariance. The metric can then be written in the form

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu$$
$$= g_{tt}(dt)^2 + g_{ab}\, dx^a dx^b \quad (a,b = 1,2,3), \tag{3.6}$$

where the metric functions g_{tt} and g_{ab} depend only upon the spatial coordinates x^a. As a simple illustration, consider the minimally coupled massive Klein–Gordon field satisfying

$$\nabla_\mu \nabla^\mu \Phi - m^2 \Phi = 0, \tag{3.7}$$

where ∇_μ denotes a covariant derivative. For the metric of Eq. (3.6), Φ satisfies the equation

$$g^{tt} \frac{\partial^2 \Phi}{\partial t^2} + (-g)^{-1/2} \frac{\partial}{\partial x^a} (-g)^{1/2} g^{ab} \frac{\partial}{\partial x^b} \Phi - m^2 \Phi = 0, \tag{3.8}$$

and, consequently, it is natural to expand in terms of the eigenfunctions of the operator

$$\Delta \equiv (-g^{tt})^{-1} \left\{ (-g)^{-1/2} \frac{\partial}{\partial x^a} (-g)^{1/2} g^{ab} \frac{\partial}{\partial x^b} - m^2 \right\}, \tag{3.9}$$

the natural generalization of the flat space Laplacian. One discovers thereby that the modes effectively decouple, each q_A satisfying a distinct equation; and, consequently, it follows that the entropy is conserved. For a source-free linear field in a static space-time, $dS/dt \equiv 0$.

This result manifests an obvious connection with the phenomenon of particle creation, it being well known that, in the absence of perverse global effects like event horizons, a source-free linear field in a static space-time admits no particle creation.[18] If, however, one allows for nonlinearities in the basic field equation, as would be induced, e.g., by a $\lambda\Phi^p$ coupling, and if, as would seem natural, one still expands in terms of the eigenfunctions of Δ, the q_A's will be coupled with one another. And this coupling guarantees that dS/dt will no longer be conserved. Correlations will evolve and these will lead to changes in the entropy. This change in S is again related to particle creation. As will be shown in Sec. V, if the system is free of correlation at some initial time t_0, the evolving correlations induced by the dynamics will lead both to an initial increase in entropy and to an enhancement in the rate of particle creation.

2. A conformally static space-time

A net overall expansion or contraction has comparatively little effect upon the evolution of the entropy. It is natural now to write the metric in the form

$$ds^2 = \Omega^2(t) \left[g_{tt}(dt)^2 + g_{ab}\, dx^a dx^b \right], \tag{3.10}$$

which differs from Eq. (3.6) only by the presence of the conformal factor $\Omega(t)$. And, with such a metric, it is again natural to expand the Klein–Gordon field (3.7) in terms of the eigenfunctions of the Δ of Eq. (3.9). The equation satisfied by each q_A now becomes more complicated,[19] but, nevertheless, one discovers thereby that, in the absence of nonlinearities, the modes remain decoupled so that $dS/dt \equiv 0$. If, alternatively, one allows for nonlinearities, one again induces couplings between the modes which lead to a nonconserved entropy and a simple H-theorem inequality.

For the case of a conformally static space-time, the connection between entropy generation and particle creation becomes more subtle, it being well known that even a source-free linear field theory can lead to the creation of particles. The point to observe, however, is that particle creation can be triggered by two quite distinct mechanisms.

(1) The fact that the space-time is dynamic leads to a mixing of positive and negative frequency states, and hence a change in the number of particles in any given mode.[20] This may be interpreted as a manifestation of the type of "parametric amplification" well known in quantum optics.[6]

(2) Couplings between modes will lead to the creation or destruction of particles whether or not the space-time is dynamic.

The entropy S defined in this paper says absolutely nothing about the effects of "parametric amplification." The particle creation associated with "mode–mode coupling" does, however, correspond to an increase in the entropy. If one starts with an initially uncorrelated state at some time t_0, the generated correlations reflected in the μ_I of Eq. (2.24) necessarily induce a net creation of particles at time $t_0 + \Delta t$.

3. A Mixmaster Universe (Ref. 21)

These models correspond to a class of cosmological solutions to the Einstein equation which admit a foliation into a family of spacelike hypersurfaces, each of which is homogeneous but characterized by a nontrivial and dynamic spatial curvature. The existence of the foliation leads to a natural 3 + 1 splitting, and the theory of harmonic analysis leads to a natural selection of spatial eigenfunctions in terms of which to expand, so that the choice of mode decomposition is very nearly canonical.

The crucial new effect in such a universe is the fact that even a linear field theory leads to a collection of *coupled* oscillators! The ordinary flat space equations

$$\frac{d^2q_A}{dt^2} + \omega_A{}^2 q_A = 0, \quad \omega_A{}^2 = k_A{}^2 + m^2, \tag{3.11}$$

for a Klein–Gordon field are now replaced by more complicated equations[22]

$$\frac{d^2q_A}{dt^2} + \sum_B b_{AB}\frac{dq_B}{dt} + \sum_B c_{AB}q_B = 0, \tag{3.12}$$

where b_{AB} and c_{AB} are time-dependent coefficients which reflect the effects of the dynamical curvature. For the special case of a conformally static Friedmann Universe, the matrices b_{AB} and c_{AB} are diagonal and the modes decouple,[19] but, in general, one infers a linear coupling between the modes which induces a time dependent $S(t)$. Formally, this may be understood by observing that, in the presence of nontrivial dynamics, the appropriate choice of spatial eigenfunctions will in fact exhibit a parametric time dependence.

This illustrates the important fact that, in a realistic dynamical space-time, even the simplest source-free linear field theory can lead to a nontrivial generation of entropy. The notion of entropy becomes, if anything, even more important in a dynamical space-time manifold.

IV. THE DERIVATION OF A SUBDYNAMICS AND AN *H*-THEOREM INEQUALITY

A. The derivation of a subdynamics (Ref. 23)

The object of this section is to derive an exact closed equation for the evolution of $S(t)$ involving only the reduced one-oscillator $g(A)$'s, and to use that exact equation to prove a short time H-theorem for a state evidencing no initial correlations. Most of the work entails the construction of a subdynamics for the $g(A)$'s and a derivation of an exact equation for their evolution which involves no explicit reference to the higher-order correlations. Once such an equation has been obtained, it is straightforward to evaluate dS/dt, and it is completely trivial to show that an initially uncorrelated state leads to an initial entropy generation.

The starting point for the analysis is a collection of "objects" that interact in a fashion described by the classical equations of motion

$$\frac{dq_A}{dt} = \alpha_A(q_A,\pi_A;t) + \xi_A(\{q_B,\pi_B\};t)$$

and

$$\frac{d\pi_A}{dt} = \gamma_A(q_A,\pi_A;t) + \eta_A(\{q_B,\pi_B\};t), \quad A = 1,2,3,..., \tag{4.1}$$

where α, γ, ξ, and η are arbitrary time-dependent functions of the q's and π's, required only to satisfy the conditions

$$\sum_A\left(\frac{\partial\alpha_A}{\partial q_A} + \frac{\partial\gamma_A}{\partial\pi_A}\right) = 0$$

and

$$\sum_A\left(\frac{\partial\xi_A}{\partial q_A} + \frac{\partial\eta_A}{\partial\pi_A}\right) = 0. \tag{4.2}$$

These equations could have been derived from a Hamiltonian—as would be expected for most realistic systems (recall, however, the caveat in Sec. II B!)—but that is by no means necessary. The conditions (4.2) ensure an analog of the standard Liouville theorem, i.e., conservation of phase, so that the fundamental Liouville equation is truly linear. The construction of a subdynamics *per se* does not require the imposition of these conditions, but, if they are not imposed, one cannot infer a one-to-one correspondence between changes in the entropy (2.22) and the evolution of correlations in the system.

Note also that the equations of motion (4.1) are significantly more general than one might reasonably expect for any realistic field theory. This generality should, however, serve to emphasize that the proof presented here is of very broad generality, including as a special case a collection of N point masses interacting via two-body forces.

The fundamental object for the statistical description of such a classical system is a many-object distribution function μ defined in an infinite-dimensional phase space. The details of this phase space are not particularly relevant here; and, as discussed by Kandrup,[19] one can assume simply that it is constructed in the obvious way as the direct product of an infinite number of two-dimensional, flat, one-object phase spaces. The only important point is that there exist a notion of probability conservation, so that

$$\frac{\partial\mu}{\partial t} + \sum_A\left(\frac{dq_A}{dt}\mu\right) + \sum_A\left(\frac{d\pi_A}{dt}\mu\right) = 0. \tag{4.3}$$

Given this expression of conservation of probability, it follows immediately from Eqs. (4.1) and (4.2) that the evolution of μ is governed by the linear Liouville equation

$$\frac{\partial\mu}{\partial t} + \sum_A(\alpha_A + \xi_A)\frac{\partial\mu}{\partial q_A} + \sum_A(\gamma_A + \eta_A)\frac{\partial\mu}{\partial\pi_A} \equiv \frac{\partial\mu}{\partial t} + Lu = 0. \tag{4.4}$$

It is useful to view this L as a sum of free and interaction Liouvillians, denoted, respectively, L^0 and L^I,

$$L^0 \equiv \sum_A \alpha_A\frac{\partial}{\partial q_A} + \sum_A\gamma_A\frac{\partial}{\partial\pi_A} \equiv \sum_A L^0{}_A \tag{4.5}$$

and

$$L^I \equiv \sum_A \xi_A\frac{\partial}{\partial q_A} + \sum_A\eta_A\frac{\partial}{\partial\pi_A} \equiv \sum_A L^I{}_A. \tag{4.6}$$

The obvious point is that it is the $L^I{}_A$'s which induce evolv-

ing correlations amongst the q_A's. If $L^I{}_A \equiv 0$ for all the A's, the system is noninteracting and $dS/dt \equiv 0$.

Two important identities should be noted.

(1) Let Tr_A denote a trace over the degrees of freedom of the A th object, i.e., an integration $\int dQ_A\, d\pi_A$. It then follows trivially that the operators $L^0{}_A$ and $L^I{}_A$ are antisymmetric in the sense that, for any two functions ψ and χ,

$$\mathrm{Tr}_A\, \psi L^0{}_A \chi = -\mathrm{Tr}_A\, \chi L^0{}_A \psi$$

and (4.7)

$$\mathrm{Tr}_A\, \psi L^I{}_A \chi = -\mathrm{Tr}_A\, \chi L^I{}_A \psi.$$

This implies, in particular, that

$$\mathrm{Tr}_A\, L^0{}_A \chi \equiv 0 \equiv \mathrm{Tr}_A\, L^I{}_A \chi. \tag{4.8}$$

(2) Similarly, it follows that, for any function ψ of the $g(A)$'s,

$$L^0{}_A \psi[g(A)] = \frac{d\psi}{dg(A)} L^0{}_A g(A)$$

and (4.9)

$$L^I{}_A \psi[g(A)] = \frac{d\psi}{dg(A)} L^I{}_A g(A).$$

The operators $L^0{}_A$ and $L^I{}_A$ satisfy the Leibnitz rule. It is this identity which would fail in the absence of conservation of phase.

The generalization to a quantum description introduces no serious mathematical complications. If the equations of motion (4.1) derive from a Hamiltonian, one can simply implement the standard prescription of canonical quantization. In this event, the classical Liouvillian L is nothing other than the Poisson bracket $\{H,...\}$, and the corresponding quantum L is a commutator $[H,...]_-$. If a Hamiltonian H does not exist, one requires some other rule to generate the quantum analog of the equations of motion. In either case, the distribution function μ is simply reinterpreted as a density matrix and the classical equation (4.4) as a quantum Liouville equation

$$\frac{\partial \mu}{\partial t} + L\mu = 0, \tag{4.10}$$

where L is an abstract operator. The only important requirements are (i) that L remain linear, (ii) that L still admit a decomposition into linear contributions $L^0{}_A$ and $L^I{}_A$, and (iii) that the quantum $L^0{}_A$ and $L^I{}_A$ still satisfy the identities (4.7) and (4.9), where Tr_A is interpreted now as an abstract trace. These requirements are, e.g., guaranteed in the framework of canonical quantization. The analysis henceforth in this section will be formulated abstractly in a fashion applicable to either a classical or a quantum description.

The idea underlying a subdynamics is that one can view the total distribution function or density matrix μ as being a sum of "relevant" and "irrelevant" contributions, μ_R and μ_I, and that one can extract from the full Liouville equation an exact equation for the evolution of μ_R which contains no explicit reference to μ_I. The desired entropy $S(t)$ is to be constructed in terms of the reduced one-object $g(A)$'s defined by the prescription

$$g(A) \equiv \prod_{B \neq A} \mathrm{Tr}_B\, \mu, \tag{4.11}$$

and, consequently, it is natural to introduce the decomposition (2.24), setting

$$\mu = \mu_R + \mu_I,$$

where

$$\mu_R \equiv \prod_A g(A). \tag{4.12}$$

The object now is to implement this decomposition in a canonical fashion by means of projection operators.[24] Specifically, what is required is a linear operator $P(t)$ defined to satisfy the three requirements

$$P(t)\mu(t) = \mu_R(t), \tag{4.13}$$

$$P(t_2)P(t_1) = P(t_2) \quad \text{for} \quad t_2 \geqslant t_1, \tag{4.14}$$

and

$$\left[P(t), \frac{\partial}{\partial t}\right]\mu(t) = 0. \tag{4.15}$$

The first of these ensures that P projects out the desired μ_R. The second implies that P is in fact idempotent. The third guarantees that the operations of projection and time evolution commute, at least when restricted to the fundamental $\mu(t)$.[25]

Given such a P, it is straightforward to obtain the desired closed equation for μ_R. By acting upon the fundamental Liouville equation with the operators P and $(1-P)$, one is led to the coupled system

$$\frac{\partial \mu_R}{\partial t} + PL\mu_R = -PL\mu_I \tag{4.16}$$

and

$$\frac{\partial \mu_I}{\partial t} + (1-P)L\mu_I = -(1-P)L\mu_R. \tag{4.17}$$

It is then trivial to write down a formal solution to Eq. (4.17), yielding $\mu_I(t)$ in terms of the retarded $\mu_R(t-\tau)$; and, by substituting that solution back into Eq. (4.16), one obtains the desired closed equation for μ_R. Thus, in terms of an initial condition $\mu_I(t_0)$, one concludes that

$$\begin{aligned}\frac{\partial \mu_R(t)}{\partial t} &+ P(t)L(t)\mu_R(t) \\ &= -P(t)L(t)\mathcal{G}(t,t_0)\mu_I(t_0) \\ &\quad + \int_0^{t-t_0} d\tau\, P(t)L(t)\mathcal{G}(t,t-\tau) \\ &\quad \times [1-P(t-\tau)]L(t-\tau)\mu_R(t-\tau),\end{aligned} \tag{4.18}$$

where

$$\mathcal{G}(t_2,t_1) = T\exp\left\{-\int_{t_1}^{t_2} d\tau [1-P(\tau)]L(\tau)\right\} \tag{4.19}$$

and T denotes a time ordering operator.

By integrating over the degrees of freedom of all but one of the objects in the system, one then obtains equations for each $\partial g(A)/\partial t$ in terms of $g(A)$ and the remaining $g(B)$'s.

And, given these equations, it is straightforward to evaluate dS/dt. Alternatively, one could evaluate that quantity directly by observing that the field entropy (2.22) can also be written in the form

$$S = -\mathrm{Tr}\,\mu_R \log \mu_R, \tag{4.20}$$

where Tr denotes a trace over all the degrees of freedom. It remains, however, to prove that the desired P actually exists and to consider in greater detail its action on such quantities of interest as $L\mu_R$. These issues could be addressed at purely formal level by endowing P with additional properties and then constructing a rigorous existence proof.[26]

It is, however, instructive instead to proceed constructively by exhibiting a specific P that satisfies Eqs. (4.13)–(4.15) and then evaluating directly its effects upon the relevant quantities. One convenient choice, considered in the past,[5,7] is obtained by generalizing the approach of Willis and Picard[2] for a system of N interacting point masses. Specifically, one is instructed to view the infinite collection of objects as the $N\to\infty$ limit of a finite system, and then define

$$P(t) = \lim_{N\to\infty} \sum_{A=1}^{N} \prod_{B\neq A} g(B) \operatorname*{Tr}_B - \lim_{N\to\infty} (N-1) \prod_{B=1}^{N} g(B) \operatorname*{Tr}_B . \tag{4.21}$$

It is simple to verify that this P satisfies Eq. (4.13), and a proof of Eq. (4.15) is also not hard. The proof that P is in fact indempotent is somewhat more difficult, but can be constructed straightforwardly by observing that any function $\psi(q_A,\pi_A,\ldots;t)$ can be approximated with arbitrary precision as a sum of terms

$$\psi_i \equiv \prod_A \psi_i{}^A(q_A,\pi_A). \tag{4.22}$$

By exploiting the explicit form of P and the first identity (4.8), it is easy to see that, when acting upon μ_R, the operators P and L^0 commute, so that

$$PL^0{}_A\mu_R = L^0{}_A P\mu_R = L^0{}_A\mu_R. \tag{4.23}$$

Similarly, one can prove that, when acting on μ_R, PL^I serves to define an "average" value. Thus, one verifies explicitly that

$$PL^I{}_A\mu_R = \{L^I{}_A\}\mu_R, \tag{4.24}$$

where $\{L^I{}_A\}$ denotes an "average interaction Liouvillian" involving only the variables q_A and π_A. The form of this average will of course depend upon the form of $L^I{}_A$, and, particularly, whether the fundamental interactions are two-, three-, or higher-body. If, for example, $L^I{}_A$ can be written as a sum of contributions $L_A^{B_1\cdots B_n}$ involving the interaction of A with objects $B_1,\ldots,B_n$, each of which satisfies Eq. (4.7), one can conclude that

$$\{L^I{}_A\}\mu_R = \sum_{B_1}\cdots\sum_{B_n} \{L_A^{B_1\cdots B_n}\}\mu_R, \tag{4.25}$$

where, explicitly,

$$\{L_A^{B_1\cdots B_n}\} = \operatorname*{Tr}_{B_1}\cdots\operatorname*{Tr}_{B_n} L_A^{B_1\cdots B_n} g(B_1)\cdots g(B_n). \tag{4.26}$$

Note in particular that $\{L^I{}_A\}$, like $L^I{}_A$, satisfies the identities (4.7) and (4.9).

Now introduce the operator

$$\Delta_A \equiv L^I{}_A - \{L^I{}_A\}. \tag{4.27}$$

One concludes then that, in the absence of initial correlations, so that $\mu_I(t_0)\equiv 0$,

$$\frac{\partial\mu_R(t)}{\partial t} + \sum_A L^0{}_A(t)\mu_R(t) + \sum_A \{L^I{}_A(t)\}\mu_R(t) = \int_0^{t-\tau} d\tau\, P(t)L(t)\mathcal{G}(t,t-\tau) \times \sum_B \Delta_B(t-\tau)\mu_R(t-\tau). \tag{4.28}$$

This implies that each $g(A)$ will satisfy an equation of the form

$$\frac{\partial g(A,t)}{\partial t} + L^0{}_A(t)g(A,t) + \{L^I{}_A(t)\}g(A,t) = \mathcal{S}[g(A,t)] \equiv \prod_{C\neq A} \operatorname*{Tr}_C \int_0^{t-\tau} d\tau\, P(t)L(t)\mathcal{G}(t,t-\tau) \times \sum_B \Delta_B(t-\tau)\prod_D g(D,t-\tau). \tag{4.29}$$

If $\mathcal{S}[g]$ were assumed to vanish identically, one would be reduced to a type of "mean field" description appropriate in the limit that $\mu \simeq \Pi_A g(A)$. It is the nontrivial $\mathcal{S}$ which leads to a nonconserved entropy.

Equation (4.29) simplifies further. Thus, by virtue of Eq. (4.7) and the definition of P, it follows that, for any function ψ,

$$\prod_{B\neq A} \operatorname*{Tr}_B (1-P)\psi = 0, \tag{4.30}$$

where the trace extends over the degrees of freedom of all but one of the objects in the system. Given, moreover, that $\mathcal{G}(t,t-\tau)$ is constructed as a sum of terms involving $(1-P)$, it follows that

$$\prod_{B\neq A} \operatorname*{Tr}_B \mathcal{G}(t,t-\tau)\psi = 0. \tag{4.31}$$

This means that, in Eq. (4.29), the operator $P(t)$ can be omitted. Equation (4.7) implies further that $L(t)$ may be replaced by $L_A(t)$, and it is also found easy to see that, since $L^0{}_A$ is independent of the variables $B\neq A$, its contribution vanishes identically. Finally, note that Eq. (4.31) implies also that $L^I{}_A$ can be replaced by Δ_A. One concludes, therefore, that

$$\mathcal{S}[g] = \prod_{C\neq A} \operatorname*{Tr}_C \int_0^{t-\tau} dt\, \Delta_A(t)\mathcal{G}(t,t-\tau) \times \sum_B \Delta_B(t-\tau)\mu_R(t-\tau). \tag{4.32}$$

B. The time dependence of the entropy

Given Eq. (4.32), it becomes straightforward to evaluate dS/dt. It follows at once that

$$\frac{dS}{dt} = -\sum_A \operatorname*{Tr}_A [1+\log g(A)]\frac{\partial g(A)}{\partial t}. \tag{4.33}$$

Equations (4.9) and (4.29) then guarantee that, in the absence of initial correlations,

$$\frac{dS}{dt} = \sum_A \mathrm{Tr}[L^0{}_A g(A)\log g(A) + \{L^I{}_A\} g(A)\log g(A)]$$

$$- \sum_A \mathrm{Tr}[1 + \log g(A)]\mathscr{S}[g(A)]. \tag{4.34}$$

Equation (4.7), and the analog satisfied by $\{L^I{}_A\}$, imply further that the first two terms in Eq. (4.34) vanish identically, so that

$$\frac{dS(t)}{dt}$$

$$= -\mathrm{Tr}\int_0^{t-\tau} d\tau \sum_A [1 + \log g(A)]\Delta_A(t)\mathscr{G}(t,t-\tau)$$

$$\times \sum_B \Delta_B(t-\tau)\mu_R(t-\tau). \tag{4.35}$$

The antisymmetry of Δ_A means that

$$\frac{dS(t)}{dt} = \mathrm{Tr}\int_0^{t-\tau} dt \left[\sum_A \Delta_A(t)\log g(A,t)\right]\mathscr{G}(t,t-\tau)$$

$$\times \sum_B \Delta_B(t-\tau)\mu_R(t-\tau), \tag{4.36}$$

and the fact that it satisfies the Leibnitz rule guarantees further that

$$\Delta_A \log g(A) = g^{-1}(A)\Delta_A g(A) = \mu_R{}^{-1}\Delta_A \mu_R, \tag{4.37}$$

so that, finally, one concludes that

$$\frac{dS(t)}{dt} = \mathrm{Tr}\int_0^{t-\tau} d\tau\, \mu_R{}^{-1}(t)\zeta(t)\mathscr{G}(t,t-\tau)\zeta(t-\tau), \tag{4.38}$$

where

$$\zeta = \sum_A \Delta_A \mu_R. \tag{4.39}$$

Given Eq. (4.38), the "short-time" H-theorem follows immediately. In the limit that $\tau \to 0$, $\mathscr{G}(t,t-\tau) \to 1$, so that for small intervals Δt,

$$\frac{dS(t_0+\Delta t)}{dt} = \mathrm{Tr}\,\mu_R{}^{-1}(t)|\zeta(t)|^2 \Delta t > 0. \tag{4.40}$$

An absence of initial correlations guarantees an initial increase in entropy. Note further that this quantity is quadratic in the interaction Liouvillian, and, as such, if the typical interaction is characterized by a coupling constant λ, the spontaneous entropy generation induced by μ_I will, at least for short times, scale as λ^2.

For the case of a classical system, where μ is realized as a distribution function, the final inequality in (4.40) follows immediately from the fact that μ can never be negative. In the quantum case, however, the density matrix μ need not be positive definite, so that the positivity of $dS(t_0+\Delta t)/dt$ is less transparent. The basic conclusion does, however, follow immediately by working in a representation in which μ_R is diagonal. That μ_R may be diagonalized follows in turn from the fact that the density matrix must be symmetric (i.e., Hermitian), and, once μ_R has been diagonalized, the inequality obtains trivially from the fact that the diagonal elements of μ_R, and hence $\mu_R{}^{-1}$, must be non-negative.[27]

If, initially, $\mu_I(t)$ does not vanish identically, the expression (4.38) for dS/dt will contain an additional contribution of the form

$$\mathrm{Tr}[1 + \log \mu_R(t)]P(t)L(t)\mathscr{G}(t,t_0)\mu_I(t_0). \tag{4.41}$$

By exploiting the identity (4.30) and the fact that L is antisymmetric and satisfies the Leibnitz rule, one concludes then that, quite generally,

$$\frac{dS}{dt} = \mathrm{Tr}\int_0^{t-\tau} d\tau\, \mu_R{}^{-1}(t)\zeta(t)\mathscr{G}(t,t-\tau)\zeta(t-\tau)$$

$$- \mathrm{Tr}\,\mu_R{}^{-1}(t)\sigma(t)\mathscr{G}(t,t_0)\mu_I(t_0), \tag{4.42}$$

where

$$\sigma = L\mu_R. \tag{4.43}$$

This means that, in the limit that $t = t_0$, the derivative dS/dt need not vanish identically. Rather, one sees that

$$\frac{dS(t_0)}{dt} = -\mathrm{Tr}\,\mu_R{}^{-1}(t_0)\sigma(t_0)\mu_I(t_0), \tag{4.44}$$

which, depending upon the form of the initial μ_I and μ_R, could be either positive or negative! By judicious choice of initial conditions, one can, at least in principle, induce an initial decrease in the entropy.

This might, naively, seem a disturbing result, but further reflection demonstrates that such a possibility should have been expected. If one is allowed to choose the initial conditions arbitrarily, one *should* be able to stimulate either an initial increase or an initial decrease in S. As discussed in Ref. 6, the situation is qualitatively similar to that arising in the analysis of particle creation in a time-dependent electromagnetic or gravitational field. Here again the choice of an especially simple state, such as the vacuum or an eigenstate of the number operator, leads to a spontaneous generation of particles; but an allowance for more generic initial conditions, in this case reflecting nontrivial "phase" information, leads to the possibility of stimulated creation or destruction of quanta.

Given these observations, it becomes clear that, as was intimated in the Introduction, an H-theorem could only be expected to hold in full generality at late times, after which any nontrivial initial correlations have either "decayed" or else have been completely "dwarfed" by the systematic correlations induced spontaneously by the evolving dynamics.

V. ENTROPY GENERATION AND PARTICLE CREATION

A. A simple example

The purpose of this section is to provide some general insights into the phenomenon of particle creation induced by mode–mode coupling and by parametric amplification, and, specifically, to demonstrate that the former mechanism is connected intimately with the generation of entropy. Specifically, it is also possible to prove a short-term "H-theorem" for particle creation which states that correlations induced from an initially uncorrelated state lead to an overall enhancement in the rate at which quanta are created. This theorem, and other related phenomena, will be discussed first for a system described by a specific model Hamiltonian

$H(t)$, and thereafter, a general proof of the particle creation H-theorem will be presented for a more general system described by an arbitrary interaction Hamiltonian H^I constructed from the "coordinates" q_A.

The specific model Hamiltonian is chosen to take the form

$$H=\sum_A \frac{1}{2}(\pi_A{}^2+\omega_A{}^2 q_A{}^2)+\sum_{B\neq A}\frac{1}{2}c_{AB}q_A q_B$$

$$\equiv\sum_A H^0{}_A+\sum_{B\neq A}H^I_{AB}, \tag{5.1}$$

where $\omega_A{}^2>0$ and c_{AB} are arbitrary real functions of time. This leads to classical equations of motion

$$\frac{dq_A}{dt}=\pi_A$$

and

$$\frac{d\pi_A}{dt}=-\omega_A{}^2 q_A-\sum_B c_{AB}q_B, \tag{5.2}$$

or, equivalently,

$$\frac{d^2 q_A}{dt^2}+\omega_A{}^2 q_A+\sum_B c_{AB}q_B=0. \tag{5.3}$$

This H is reasonable for several reasons.

(1) If $c_{AB}\equiv 0$ and the ω_A's are assumed time-independent constants, one recovers the massless Klein–Gordon equation in flat space. Thus, if one interprets A as labeling a $\mathbf{k}$ vector and sets $\omega_A{}^2=k_A{}^2$, the decomposition (2.8) leads immediately to Eq. (2.14).

(2) If $c_{AB}\equiv 0$ but the ω_A's are allowed to be functions of time, Eq. (5.1) includes as a special example the Klein–Gordon equation in a $k=0$ Friedmann cosmology. Thus the identification

$$\omega_k{}^2=k^2-\ddot{\Omega}/\Omega, \tag{5.4}$$

where Ω is the conformal factor and an overdot is a conformal time derivative, together with the decomposition

$$\Phi(\mathbf{x},t)=\Omega^{-1}\sum_k q_k(t)\exp(-i\mathbf{k}\cdot\mathbf{x}), \tag{5.5}$$

leads to a field equation

$$-\frac{\partial^2\Phi}{\partial t^2}-\frac{2\dot{\Omega}}{\Omega}\frac{\partial\Phi}{\partial t}+\delta^{ab}\frac{\partial}{\partial x^a}\frac{\partial}{\partial x^b}\Phi=0, \tag{5.6}$$

which is nothing other than the Klein–Gordon equation appropriate for a conformally flat space-time with $ds^2=\Omega^2(t)\eta_{\mu\nu}\,dx^\mu dx^\nu$.

(3) In the case that c_{AB} and ω_A are both nonvanishing and exhibit a nontrivial time dependence, the Hamiltonian (5.1) mocks quite reasonably the behavior of a scalar field in a Mixmaster Universe. One might like also to allow for more complicated nonlinear couplings, as induced, e.g., by a $\lambda\Phi^p$ field theory, but, as will be seen eventually, more complicated contributions involving c_{ABC}, c_{ABCD}, etc. do not alter the existence of the basic H-theorem for particle creation.

The passage to a quantum theory can now be effected trivially by canonical quantization, the Poisson bracket structure being replaced by a canonical commutation relation

$$[q_A,\pi_B]_-=i\delta_{AB}. \tag{5.7}$$

The quantum system is then characterized by a many-oscillator density matrix μ, the evolution of which is governed by the quantum Liouville equation

$$\frac{\partial\mu}{\partial t}=-L\mu=-i[H,\mu]_-. \tag{5.8}$$

Note for future reference that the free Hamiltonian $L^0{}_A$ satisfies

$$L^0{}_A\xi=i[H^0{}_A,\xi]_-, \tag{5.9}$$

and that the average interaction Liouvillian defined in Sec. IV takes the form

$$\{L^I{}_A\}\xi=i[\{H^I{}_A\},\xi]_-, \tag{5.10}$$

where

$$\{H^I{}_A\}=\sum_{B\neq A}\operatorname{Tr}_B H^I_{AB}g(B)=\sum_B\frac{1}{2}c_{AB}\langle q_B\rangle q_A, \tag{5.11}$$

the angular brackets $\langle\ \rangle$ denoting an expectation value with respect to the full many-oscillator μ. This implies further that the Δ_A of Sec. IV can be realized in the form

$$\sum_A\Delta_A\,\xi=\sum_A\sum_{B\neq A}i[h_{AB},\xi]_-, \tag{5.12}$$

where

$$h_{AB}\equiv H^I_{AB}-\operatorname{Tr}_B H^I_{AB}g(B)-\operatorname{Tr}_A H^I_{AB}g(A). \tag{5.13}$$

At this stage, it is customary to define creation and annihilation operators

$$a_A=(\sqrt{2\omega_A})^{-1}(\omega_A q_A+i\pi_A)$$

and (5.14)

$$a^\dagger{}_A=(\sqrt{2\omega_A})^{-1}(\omega_A q_A-i\pi_A),$$

in terms of which the free Hamiltonian

$$H^0=\sum_A(\omega_A{}^2 q_A{}^2+\pi_A{}^2)$$

$$=\sum_A\frac{1}{2}\omega_A(a^\dagger{}_A a_A+a_A a^\dagger{}_A)$$

$$=\sum_A\omega_A\left(a^\dagger{}_A a_A+\frac{1}{2}\right). \tag{5.15}$$

In the absence of any interaction H^I, the operator

$$N_A\equiv a^\dagger{}_A a_A \tag{5.16}$$

would admit to an unambiguous interpretation as "the number of quanta in mode A." In the presence of nontrivial interactions, this interpretation is less clear-cut, but, at least in the limit of weak couplings, where, typically, $|c|\ll\omega^2$, it is still conventional to think of N_A as reflecting some measure of "particle number." Whether this generalization is reasonable is, perhaps, unclear, but for the remainder of this section, changes in the normal-ordered statistical expectation value

$$\langle N_A\rangle\equiv\operatorname{Tr}a^\dagger{}_A a_A\mu=\operatorname{Tr}(2\omega_A)^{-1}(\pi_A{}^2+\omega_A{}^2 q_A{}^2)\mu$$

$$=(\omega_A)^{-1}\langle H^0{}_A\rangle \tag{5.17}$$

will be spoken of as corresponding to "changes in particle number." The statements in the Introduction and in the remainder of this section regarding the net creation of quanta refer specifically to the time dependence of $\langle N_A \rangle$.

The object now is to evaluate $\partial \langle N_A \rangle / \partial t$, allowing for contributions arising both from a changing μ and from the time dependence of the ω_A's. Thus, one sees immediately that

$$\frac{\partial \langle N_A \rangle}{\partial t} = \mathrm{Tr} \frac{1}{\omega_A} H^0{}_A \frac{\partial \mu}{\partial t} + \mathrm{Tr}\, \mu \frac{\partial}{\partial t}\left(\frac{1}{\omega_A} H^0{}_A\right). \tag{5.18}$$

It is then trivial to verify that

$$\begin{aligned} \mathrm{Tr}\, H^0{}_A \frac{\partial \mu}{\partial t} &= -i\, \mathrm{Tr}\, H^0{}_A [H,\mu]_- \\ &= -i\, \mathrm{Tr}\, \mu [H^0{}_A, H]_- \\ &= -i\, \mathrm{Tr}\, \mu [H^0{}_A, H^I]_-. \end{aligned} \tag{5.19}$$

And, therefore, one concludes that

$$\begin{aligned} \frac{\partial \langle N_A \rangle}{\partial t} = &-\frac{\dot{\omega}_A}{2\omega_A{}^2} \mathrm{Tr}\, \mu (\pi_A{}^2 - \omega_A{}^2 q_A{}^2) \\ &- \sum_{B \neq A} \frac{c_{AB}}{\omega_A} \mathrm{Tr}\, \mu \pi_A q_B. \end{aligned} \tag{5.20}$$

The first term in Eq. (5.20) reduces at once to a partial trace Tr_A weighted by the reduced $g(A)$. The second entails couplings between modes A and B, so that one obtains instead a double trace $\mathrm{Tr}_A\, \mathrm{Tr}_B$ weighted by the two-oscillator $g_2(A,B)$. Thus, if one writes

$$g_2(A,B) = g(A)g(B) + \nu(A,B), \tag{5.21}$$

where $\nu(A,B)$ denotes the pair correlation function, one concludes that

$$\begin{aligned} \frac{\partial \langle N_A \rangle}{\partial t} = &-\frac{\dot{\omega}_A}{2\omega_A{}^2} \langle \pi_A{}^2 - \omega_A{}^2 q_A{}^2 \rangle \\ &- \sum_{B \neq A} \frac{c_{AB}}{\omega_A} \langle \pi_A \rangle \langle q_B \rangle + \frac{\partial \langle N_A \rangle^c}{\partial t}, \end{aligned} \tag{5.22}$$

where

$$\frac{\partial \langle N_A \rangle^c}{\partial t} \equiv - \sum_{B \neq A} \mathrm{Tr}_A\, \mathrm{Tr}_B\, \nu(A,B) \pi_A q_B \frac{c_{AB}}{\omega_A}. \tag{5.23}$$

Alternatively, in terms of the μ_I defined in Sec. IV,

$$\frac{\partial \langle N_A \rangle^c}{\partial t} = - \sum_{B \neq A} \mathrm{Tr}\, \mu_I \pi_A q_B \frac{c_{AB}}{\omega_A}. \tag{5.24}$$

The first two terms in Eq. (5.22) can be interpreted as representing a mean field particle creation $\partial \langle N_A \rangle^m / \partial t$ which will be present even if, in the spirit of the collisionless Boltzmann equation, one were to pretend that $\mu_I \simeq 0$. The final term represents instead a "correlational" particle creation induced specifically by the correlations amongst the modes. The field entropy (5.22) has no direct connection with the mean field contributions, but it does connect intimately with $\partial \langle N_A \rangle^c / \partial t$. Specifically, the same arguments which led to the short-time H-theorem (5.40) imply also that $\partial \langle N_A(t_0 + \Delta t) \rangle^c / \partial t > 0$.

B. Correlational particle creation for the simple example

Turn, therefore, to an explicit evaluation of this quantity for an initially uncorrelated state. The analysis of Sec. IV indicates that, in the absence of any initial correlations,

$$\mu_I(t) = \int_0^{t-t_0} d\tau\, \mathcal{G}(t,t-\tau) \sum_A \Delta_A(t-\tau) \mu_R(t-\tau). \tag{5.25}$$

And, thus, in terms of the Δ_A of Eq. (4.27), one concludes immediately that

$$\begin{aligned} \frac{\partial \langle N_A(t) \rangle^c}{\partial t} = &\, i \sum_B \mathrm{Tr} \frac{c_{AB}}{\omega_A} \pi_A q_B \int_0^{t-t_0} d\tau\, \mathcal{G}(t,t-\tau) \\ &\times \sum_{C \neq D} [h_{CD}(t-\tau), \mu_R(t-\tau)]_-. \end{aligned} \tag{5.26}$$

A simple application of the cyclic trace rule

$$\mathrm{Tr}\, ABC = \mathrm{Tr}\, CAB \tag{5.27}$$

then implies that

$$\begin{aligned} \frac{\partial \langle N_A(t) \rangle^c}{\partial t} = &-i \sum_B \mathrm{Tr} \frac{c_{AB}}{\omega_A} \int_0^{t-t_0} d\tau \\ &\times \sum_{C \neq D} [h_{CD}(t-\tau), q_B \pi_A \mathcal{G}(t,t-\tau)]_- \\ &\times \mu_R(t-\tau). \end{aligned} \tag{5.28}$$

For small τ, $\mathcal{G}(t,t-\tau) \to 1$ and all the quantities in Eq. (5.28) can be approximated by their values at time t_0, so that

$$\begin{aligned} &\frac{\partial \langle N_A(t_0 + \Delta t) \rangle^c}{\partial t} \\ &\quad = -i \sum_B \mathrm{Tr} \frac{c_{AB}}{\omega_A} \sum_{C \neq D} [h_{CD}, q_B \pi_A]_- \mu_R(t_0) \Delta t. \end{aligned} \tag{5.29}$$

One then verifies immediately that

$$\begin{aligned} [h_{CD}, q_B \pi_A]_- = &\, (i/2) q_B (c_{AC}(q_C - \langle q_C \rangle)\delta_{AD} \\ &+ c_{AD}(q_D - \langle q_D \rangle)\delta_{AC}), \end{aligned} \tag{5.30}$$

so that

$$\begin{aligned} &\frac{\partial \langle N_A(t_0 + \Delta t) \rangle^c}{\partial t} \\ &\quad = \sum_B \mathrm{Tr} \frac{c_{AB}}{\omega_A} \sum_C q_B c_{AC} (q_C - \langle q_C \rangle) \mu_R(t_0) \Delta t. \end{aligned} \tag{5.31}$$

The trace in Eq. (5.31) implies that the only nonvanishing contributions arise when $C = B$, so that

$$\frac{\partial \langle N_A(t_0 + \Delta t) \rangle^c}{\partial t} = \sum_B \frac{(c_{AB})^2}{\omega_A} (\langle q_B{}^2 \rangle - \langle q_B \rangle^2) \Delta t > 0! \tag{5.32}$$

One obtains a positive particle creation at a rate proportional to the square of the coupling constant c.

C. General proof of spontaneous particle creation induced by correlations

The object now is to demonstrate explicitly that an analog of Eq. (5.32) will hold for an arbitrary interaction Ham-

iltonian H^I constructed from the q's. For such a more general H^I, Eq. (5.24) will be replaced by the relation

$$\frac{\partial \langle N_A \rangle^c}{\partial t} = -i \operatorname{Tr} \frac{1}{\omega_A} [H^0{}_A, H^I]_- \mu_I, \tag{5.33}$$

which can be written in the form

$$\frac{\partial \langle N_A \rangle^c}{\partial t} = -\frac{1}{2\omega_A} \operatorname{Tr} \left(\pi_A \frac{\partial H^I}{\partial q_A} + \frac{\partial H^I}{\partial q_A} \pi_A \right) \mu_I, \tag{5.34}$$

where $\partial / \partial q_A$ is interpreted as an operator derivative.

The irrelevant contribution μ_I now may be written as

$$\mu_I(t) = i \int_0^{t-t_0} d\tau \, \mathcal{G}(t, t-\tau) [h^I(t-\tau), \mu_R(t-\tau)]_-, \tag{5.35}$$

where h^I denotes an appropriate "fluctuating" interaction Hamiltonian. For short times Δt, one then infers that

$$\mu_I(t_0 + \Delta t) = i [h^I, \mu_R(t_0)]_- \Delta t, \tag{5.36}$$

so that

$$\frac{\partial \langle N_A(t_0 + \Delta t) \rangle^c}{\partial t} = \frac{1}{\omega_A} \operatorname{Tr} \frac{\partial H^I}{\partial q_A} \frac{\partial h^I}{\partial q_A} \mu_R(t_0) \Delta t. \tag{5.37}$$

Quite generally, the fluctuating h^I can be written as a difference $H^I - \{H^I\}$, where $\{H^I\}$ denotes an average interaction Hamiltonian. And thus, one sees that the right hand side of Eq. (5.37) reduces to

$$\frac{1}{\omega_A} \left(\left\{ \left(\frac{\partial H^I}{\partial q_A} \right)^2 \right\} - \left\{ \frac{\partial H^I}{\partial q_A} \right\}^2 \right) \Delta t > 0, \tag{5.38}$$

which establishes the inequality.

As a concrete example, one can consider an interaction H^I again involving only pairs of oscillators, but now quadratic in the q's,

$$H^I{}_A = \sum_{B \neq A} \frac{1}{2} c_{AB} q_A{}^2 q_B{}^2. \tag{5.39}$$

In this case,

$$h^I{}_A = \sum_{B \neq A} \frac{1}{2} c_{AB} q_A{}^2 (q_B{}^2 - \langle q_B{}^2 \rangle), \tag{5.40}$$

and one concludes that

$$\frac{\partial \langle N_A(t_0 + \Delta t) \rangle^c}{\partial t} = \sum_B \frac{(c_{AB})^2}{\omega_A} \langle q_A{}^2 \rangle (\langle q_B{}^4 \rangle - \langle q_B{}^2 \rangle^2). \tag{5.41}$$

Similarly, one can suppose instead that each oscillator A interacts linearly with n other oscillators $B_1, \ldots, B_n$ via a coupling $c_{AB_1 \cdots B_n}$. The net result then is that

$$\frac{\partial \langle N_A(t_0 + \Delta t) \rangle^c}{\partial t} = \sum_{B_1 \neq \cdots \neq B_n} \cdots \sum \frac{(c_{AB_1 \cdots B_n})^2}{\omega_A} \left\{ \prod_{B_1}^{B_n} \langle q_{B_i}{}^2 \rangle - \prod_{B_1}^{B_n} \langle q_{B_i} \rangle^2 \right\}. \tag{5.42}$$

VI. THE UNIQUENESS OF THE ENTROPY

Another important question remains to be asked. To what extent does the $S(t)$ defined by Eq. (2.22) constitute the unique field entropy? Even if one accepts the notion that any measure of entropy should be constructed solely from the $g(A)$'s, and even if one could prove rigorously a very general H-theorem for $S(t)$, there would remain the issue of whether some other entropy $Z(t)$ could be shown to satisfy the same properties.

Broadly speaking, given the general philosophy developed in this paper, there are only three mathematical requirements which must be imposed upon such a candidate $Z(t)$.

(1) This $Z(t)$ must be constructed solely from the $g(A)$'s and contain no information about the interoscillator correlations.

(2) In the absence of couplings between degrees of freedom, it must follow that $dZ/dt \equiv 0$.

(3) At least in some appropriate limit, one must be able to show that $dZ/dt \geqslant 0$.

To the extent, however, that one takes very seriously the notion of a subdynamics introduced in Sec. IV, these demands can be refined somewhat. Thus, in particular, it would be reasonable to demand not simply that $Z(t)$ be constructed only from the $g(A)$'s, but that it be realizable as a functional of the relevant μ_R at the same instant of time t, i.e., that

$$Z(t) = \operatorname{Tr} \Psi[\mu_R(t)], \tag{6.1}$$

where Ψ is some arbitrary real function of μ_R. Similarly, in the same spirit, one might demand further that, in the absence of initial correlations, dZ/dt be realizable quite generally as the simple trace of some object $\mathcal{A}$ involving only the relevant $\mu_R(t-\tau)$, the Greenian $\mathcal{G}(t, t-\tau)$, and the fluctuating Liouvillian $\Delta \equiv \Sigma_A \Delta_A$,

$$\frac{dZ}{dt} = \operatorname{Tr} \mathcal{A}[\mu_R, \Delta, \mathcal{G}]. \tag{6.2}$$

An explicit dependence on the individual $g(A)$'s or any functionals thereof would be considered inappropriate. As illustrated by Eq. (4.38), this requirement is satisfied by $S(t)$. And finally, it would be reasonable to demand at least that, if $\mu_I(t_0) = 0$,

$$\frac{dZ(t_0 + \Delta t)}{dt} > 0. \tag{6.3}$$

As will be demonstrated below, it is in fact easy to construct Z's which will satisfy the criteria (6.1) and (6.3), but it is far more difficult to satisfy the condition (6.2). Thus, for example, Eqs. (6.1) and (6.3) will hold for any

$$Z(t) = -\operatorname{Tr} \mu_R{}^p(t), \tag{6.4}$$

where p is a real number greater than unity, whereas Eq. (6.2) will not be satisfied. Indeed, it is possible to prove that, modulo an overall multiplicative factor and the addition of conserved quantities, the field entropy $S(t)$ is the unique object to satisfy the conditions (6.1)–(6.3).

Consider first the proof that the $Z(t)$ defined by Eq. (6.4) will satisfy the short-time H-theorem (6.3). A simple application of the chain rule shows that

$$\frac{dZ}{dt} = -p\sum_A \prod_{B\neq A} \sigma(B) \operatorname*{Tr}_A g^{p-1}(A)\frac{\partial g(A)}{\partial t}, \tag{6.5}$$

where

$$\sigma(A) \equiv \operatorname*{Tr}_A g^p(A). \tag{6.6}$$

One then verifies immediately that the mean field contributions to $\partial g(A)/\partial t$ have a vanishing contribution to dZ/dt by virtue of Eqs. (4.7) and (4.9), so that

$$\frac{dZ}{dt} = -p\sum_A \prod_{B\neq A} \sigma(B) \operatorname*{Tr}_A g^{p-1}(A)\mathscr{S}[g(A)]. \tag{6.7}$$

By inserting into Eq. (6.7) the explicit form of $\mathscr{S}[g(A)]$ given by Eq. (4.32), one sees that

$$\begin{aligned}\frac{dZ}{dt} = & -p\sum_A \prod_{B\neq A} \sigma(B,t) \\ & \times \operatorname{Tr} \int_0^{t-\tau} d\tau\, g^{p-1}(A,t)\Delta_A(t)\mathscr{G}(t,t-\tau) \\ & \times \sum_C \Delta_C(t-\tau)\mu_R(t-\tau).\end{aligned} \tag{6.8}$$

The facts that $\Delta_A(t)$ is antisymmetric and that it satisfies the Leibnitz rule then imply that

$$\begin{aligned}\frac{dZ}{dt} = & p(p-1)\sum_A \prod_{B\neq A} \sigma(B,t) \\ & \times \operatorname{Tr} \int_0^{t-\tau} d\tau\, g^{p-2}(A,t)[\Delta_A(t)g_A(t)]\mathscr{G}(t,t-\tau) \\ & \times \sum_C \Delta_C(t-\tau)\mu_R(t-\tau),\end{aligned} \tag{6.9}$$

or, equivalently, in terms of the quantity

$$\theta_A = \Delta_A \mu_R, \tag{6.10}$$

that

$$\begin{aligned}\frac{dZ}{dt} = & p(p-1)\sum_A \prod_{B\neq A} \sigma(B,t) \\ & \times \operatorname{Tr} \int_0^{t-\tau} d\tau\, g^{p-1}(A,t)\mu_R^{-1}(t)\theta_A(t)\mathscr{G}(t,t-\tau) \\ & \times \sum_C \theta_C(t-\tau).\end{aligned} \tag{6.11}$$

The presence of the contributions involving $\sigma(B)$ and $g^{p-1}(A)$ demonstrate explicitly that this candidate entropy cannot satisfy the criterion (6.2). At this point, however, it is easy to see that Eq. (6.3) will in fact hold. For small Δt, it follows immediately that

$$\begin{aligned}\frac{dZ(t_0+\Delta t)}{dt} = & p(p-1)\sum_A \prod_{B\neq A} \sigma(B) \\ & \times \operatorname{Tr} g^{p-1}(A)\mu_R^{-1}\theta_A \sum_C \theta_C \Delta t.\end{aligned} \tag{6.12}$$

And thus, since

$$\operatorname*{Tr}_A \theta_A(t) \equiv 0, \tag{6.13}$$

one can conclude that

$$\begin{aligned}\frac{dZ(t_0+\Delta t)}{dt} = & p(p-1)\sum_A \prod_{B\neq A} \sigma(B,t_0) \\ & \times \operatorname{Tr} g^{p-1}(A,t_0)\mu_R^{-1}(t_0)|\theta_A(t_0)|^2\Delta t.\end{aligned} \tag{6.14}$$

For the special cases $p=0$ and $p=1$, dZ/dt vanishes identically, as must of course be the case. However, for $p>1$, one concludes instead that $dZ(t_0+\Delta t)/dt > 0$.

If the system under consideration were a collection of N identical point masses, rather than an infinite set of different oscillators, Eqs. (6.11) and (6.14) would assume a more palatable form. Thus, if one considers N particles characterized by a distribution function or density matrix symmetric under particle interchange, one would conclude that

$$Z = \sigma^N, \tag{6.15}$$

where, now,

$$\sigma(t) \equiv \operatorname*{Tr}_i f^p(i;t) \tag{6.16}$$

is the same for each particle. This means that Eq. (6.11) will be replaced by a relation of the form

$$\begin{aligned}\frac{d\sigma}{dt} = & p(p-1)\operatorname{Tr} f^{p-1}(i,t)\mu_R^{-1}(t)\theta_i(t) \\ & \times \mathscr{G}(t,t-\tau)\sum_{j=1}^N \theta_j(t-\tau),\end{aligned} \tag{6.17}$$

so that, for short times,

$$\begin{aligned}\frac{d\sigma(t_0+\Delta t)}{dt} = & p(p-1)\operatorname{Tr} f^{p-1}(i,t_0) \\ & \times \mu_R^{-1}(t_0)|\theta_i(t_0)|^2\Delta t,\end{aligned} \tag{6.18}$$

where $\mu_R(t_0)$ is now a product of N identical quantities.

Turn finally to the general uniqueness proof. For an arbitrary function of the form (6.1), one concludes again that a nontrivial dZ/dt is induced only by the $\mathscr{S}[g(A)]$'s, so that, in the absence of initial correlations,

$$\begin{aligned}\frac{dZ}{dt} = & -\operatorname{Tr}\sum_A \frac{d\Psi}{d\mu_R} g^{-1}(A)\mu_R \prod_{C\neq A} \operatorname*{Tr}_C \Delta_A^{(C)} \\ & \times \int_0^{t-\tau} d\tau\, \mathscr{G}(t,t-\tau)\sum_D \Delta_D(t-\tau)\mu_R(t-\tau),\end{aligned} \tag{6.19}$$

where the superscript (C) is a reminder that $\Delta_A^{(C)}$ depends also upon the variables $C\neq A$. The antisymmetry of Δ_A and the fact that it satisfies the Leibnitz rule then imply that

$$\begin{aligned}\frac{dZ}{dt} = & \operatorname*{Tr}_A \prod_{B\neq A} \operatorname*{Tr}_B \frac{d^2\Psi}{d\mu_R^2} g^{-1}(A)\mu_R \\ & \times \prod_{C\neq A} \operatorname*{Tr}_C [\Delta_A^{(C)}\mu_R(A,B)] \int_0^{t-\tau} d\tau\, \mathscr{G}(t,t-\tau) \\ & \times \sum_D \Delta_D(t-\tau)\mu_R(t-\tau).\end{aligned} \tag{6.20}$$

Here $\Delta_A^{(C)}\mu_R(A,B)$ means that $\Delta_A^{(C)}$, an operator involving A and the dummy variables $C\neq A$, acts on μ_R, viewed as a function of A and different dummy variables $B\neq A$.

If dZ/dt is to depend only on μ_R, rather than upon the individual $g(B)$'s, one must demand that

$$\prod_{B\neq A}\mathrm{Tr}_B \frac{d^2\Psi}{d\mu_R{}^2}\mu_R(A,B)\Delta_A^{(C)}\mu_R(A,B) \tag{6.21}$$

be independent of the $g(B)$'s. Since, however, the only relevant universal property of $g(B)$ is that it have unit trace, this implies further that $(d^2\Psi/d\mu_R{}^2)\mu_R$ must itself be independent of $g(B)$. Given, moreover, that $d^2\Psi/d\mu_R{}^2$ must be a function only of μ_R, this can only be true if

$$\frac{d^2\Psi}{d\mu_R{}^2} = a\mu_R{}^{-1}, \tag{6.22}$$

where a is an arbitrary constant. In this case,

$$\begin{aligned}\frac{dZ(t)}{dt} &= a\,\mathrm{Tr}\sum_A \Delta_A\mu_R\int_0^{t-\tau} d\tau\,\mathscr{G}(t,t-\tau)\\ &\quad\times\sum_B \Delta_B(t-\tau)\mu_R(t-\tau)\\ &= a\,\mathrm{Tr}\int_0^{t-\tau} d\tau\,\zeta(t)\mathscr{G}(t,t-\tau)\zeta(t-\tau),\end{aligned} \tag{6.23}$$

where ζ was defined in Eq. (4.39).

The right-hand side of Eq. (6.23) is, modulo the constant factor a, of precisely the form satisfied by the field entropy $S(t)$ of Eq. (2.22). Thus, one concludes (i) that $Z(t)$ is consistent with the demand (6.2) if and only if $\Psi[\mu_R]$ satisfies Eq. (6.22), and (ii) that $Z(t)$ will evidence a monotonic evolution if and only if $S(t)$ increases monotonically. The most general solution to Eq. (6.22) is of course

$$\Psi[\mu_R] = a\mu_R\log\mu_R + b\mu_R + c, \tag{6.24}$$

where b and c are arbitrary constants. And, since

$$\frac{d}{dt}\mathrm{Tr}(b\mu_R + c)\equiv 0, \tag{6.25}$$

one can set $b=c=0$ without any loss of generality. The most general entropy $Z(t)$ consistent with Eq. (6.2) is, therefore,

$$Z(t) = -a\,\mathrm{Tr}\,\mu_R\log\mu_R, \tag{6.26}$$

where a is a constant, and the demand that a be real and positive guarantees that $Z(t)$ will satisfy Eq. (6.3).

Net conclusion: Modulo the addition of conserved quantities and multiplication by a positive constant, the $S(t)$ defined by Eq. (2.22) is the unique entropy guaranteed to satisfy a short-time H-theorem (6.3) and consistent with the general subdynamics in the sense that it satisfies Eq. (6.2). Thus $S(t)$ may, or may not, satisfy an H-theorem in some broad generality. It is, however, the only candidate entropy consistent with the demands outlined at the outset of this section.

VII. DISCUSSION

A. The late time evolution of S

Section I of this paper introduced a notion of particle entropy constructed explicitly as a measure of correlations which, at least in one special limit, is guaranteed to satisfy a short-time H-theorem inequality. Sections II and III then generalized that definition to a classical or quantum field, emphasizing in particular the question of "physical significance." The entropy so constructed is conserved absolutely in the absence of couplings between degrees of freedom, and, as such, there can be no entropy generation for a source-free linear field in Minkowski space. If, however, couplings are induced by sources, by nonlinearities, or by a nontrivial dynamical space-time, the entropy need no longer be conserved. And, as was illustrated in Sec. IV, one can prove a short-time H-theorem which guarantees that an initially uncorrelated state leads (at least) to an initial increase in entropy. Section V demonstrated that there also exists a direct connection between entropy and particle creation, proving that evolving correlations which induce a net entropy generation lead also to an enhancement in the rate at which quanta are produced. Finally, Sec. VI showed that, in a well defined mathematical sense, the entropy $S(t)$ defined in Secs. II and III is the only measure of entropy consistent with the notion of subdynamics considered in Sec. IV. $S(t)$ may—or may not—satisfy a sufficiently general H-theorem to be physically useful; it is, however, the only viable candidate for an entropy compatible with the underlying symmetries inherent in the subdynamics.

These are encouraging results, but it remains to consider in greater detail the more generic time dependence of S. After all, the principal utility of the intuitive notion of entropy is that S satisfies an H-theorem quite generally. It is, therefore, important to ask whether the S of Eq. (2.22) necessarily increases monotonically for all times and whether, or under what circumstances, there exists an approach of the field towards a unique equilibrium state.

The basic desired properties of S can be summarized by three conjectures, the validity of which will be motivated and discussed in the remainder of this subsection.

(1) Neglecting "initial transients," the entropy S is guaranteed to increase monotonically for all times.

(2) If there exists a well defined static equilibrium state $\mu_{\mathrm{eq}}\propto\exp(-\beta H)$, the increase in S coincides with an approach of μ towards its equilibrium form.

(3) The equilibrium μ_{eq} maximizes S, at least locally, with respect to variations $\delta\mu$ which satisfy appropriate constraints.

It is important to stress once again that S cannot increase monotonically for every possible system. If, e.g., the system evidences a periodic evolution, S must also be periodic; and thus, if the entropy increases at one point of time, it must decrease at another. A universal H-theorem can hold only for systems which are, in some suitable sense, "complex" or "ergodic." It is also important to emphasize that the neglect of initial transients is an essential caveat. In the absence of initial correlations, the immediate response of the system is to increase its entropy. But, as discussed in Sec. IV, if there do exist such initial correlations, they could in principle induce a net decrease in the entropy. Only after these nontrivial initial correlations have "died away" could one expect a universal H-theorem to hold.

The existence of such a universal H-theorem, and the existence of a systematic evolution, is especially difficult to address for a system in which no static end state can exist. Some concrete results are, however, known for the special case of systems characterized by a time-independent Hamiltonian. Most important, perhaps, is the fact that an arbitrary

initial $\mu_{\rm in}$ *cannot converge pointwise* towards the equilibrium $\mu_{\rm eq}$! It may well tend towards $\mu_{\rm eq}$ in some appropriate norm or in some suitable "time-averaged" sense, but a true pointwise convergence is simply impossible.

Indeed, that this is true is very easy to see. Given that H is independent of time, there exists a conserved energy and, as such, if some $\mu_{\rm in}$ were to converge towards a $\mu_{\rm eq}$, the final inverse temperature β would be determined uniquely by the demand that $\langle H\rangle_{\rm in} = \langle H\rangle_{\rm eq}$. The problem, however, is that this is not the only constraint that $\mu_{\rm eq}$ must satisfy. The linearity of the fundamental Liouville equation implies that *any* functional $\mathrm{Tr}\,\Lambda[\mu]$ must also be conserved, and it is easy to see that, in general, the unique $\mu_{\rm eq}$ determined by energy conservation cannot satisfy this infinite number of additional constraints. Only if one were to violate the Liouville equation by allowing interactions with some external "bath" could one hope to obtain a true pointwise convergence towards the canonical $\mu_{\rm eq}$. A completely isolated system cannot converge pointwise towards the true equilibrium.

However, the fact that μ cannot converge pointwise towards $\mu_{\rm eq}$ does not imply that an approximate H-theorem cannot exist. Indeed, a naive perturbation expansion appropriate in a "dilute gas" ($|H^I| \ll |H^0|$) approximation shows that an homogeneous system of N identical point masses will in fact evidence a systematic increase in entropy. Thus, in the absence of initial correlations, one concludes that the one-particle distribution function $f(i)$ will satisfy the Landau equation.[3,28] And it is well known that the Landau equation implies that $dS/dt \geqslant 0$, with equality holding if and only if $f(i) \propto \exp(-\beta H^0{}_i)$, where, in terms of the particle mass m, $H^0{}_i = \mathbf{p}_i{}^2/2m$.[3]

For these and related reasons, it is natural to ask whether the entropy (2.22) is in fact maximized by the equilibrium $g_{\rm eq}(A)$'s. More specifically, one would like to ascertain whether the equilibrium $S_{\rm eq}$ is a maximum with respect to perturbations $\delta\mu$ which preserve probability [so that $\mathrm{Tr}_A\, g(A) = 1$], energy $\langle H\rangle = \mathrm{Tr}\,\mu H$, and any other conserved quantities. This is difficult to determine in general because $\langle H\rangle$ involves not only the $g(A)$'s, but the higher-order correlations buried in such quantities as $\nu(A,B)$. This problem of entropy maximization would, however, seem tractable in a dilute gas approximation, where the total energy $\langle H\rangle$ can be approximated by the mean field value $\langle H\rangle^m \equiv \mathrm{Tr}\,\mu_R H$.

Indeed, by using Lagrange multiplier techniques, Lynden-Bell and Wood[29] have proved that a self-gravitating system of N identical particles confined within a sufficiently small spherical box will in fact maximize S at least locally with respect to infinitesmal changes in the f's which conserve probability and mean field energy. Their analysis does, however, reveal two interesting (and related?) points: (1) if the box is too large, S will still be *extremized* by the equilibrium $f_{\rm eq}(i)$'s, but it will not be a local *maximum*; and (2) although the $f_{\rm eq}(i)$'s can maximize S locally, they never constitute a global entropy maximum. These conclusions appear related to the fact that, for a self-gravitating system, the Hamiltonian is not bounded from below. For more well behaved interactions, one might expect that the equilibrium state $\mu_{\rm eq}$ does indeed correspond to a global entropy maximum.

As a concrete test of some interest, discussed in Ref. 6, one can examine the Hamiltonian system of Eq. (5.1) in the limit that the ω_A's and c_{AB}'s are independent of time, contrasting the equilibrium entropy $S_{\rm eq}$ with the entropy $S_{\rm in}$ associated with some plausible initial state of the same energy $\langle H\rangle$. Consider, for example, an initial state corresponding to a "pseudothermal" density matrix $\mu_{\rm in} \propto \exp(-\beta H^0)$. One concludes then that (i) if the $\omega_A{}^2$'s are positive and (ii) if the c_{AB}'s are sufficiently small, so that H is a positive quadratic form and $\mu_{\rm eq}$ is well defined, the entropy $S_{\rm eq}$ associated with $\mu_{\rm eq}$ will be greater than the initial $S_{\rm in}$. It is, however, also easy to show that $\mu_{\rm in}$ could not evolve pointwise towards the equilibrium $\mu_{\rm eq}$ since $\mathrm{Tr}\,\mu_{\rm in}\log\mu_{\rm in} \neq \mathrm{Tr}\,\mu_{\rm eq}\log\mu_{\rm eq}$.

B. Connection with black hole entropy

Superficially, at least, it might appear that the field entropy $S(t)$ defined in this paper has precious little to do with the entropy $S_{\rm BH}$ attributed to a black hole in general relativity. The former provides a measure of correlations induced by the evolving dynamics, whereas the latter, being defined in terms of an event horizon, would appear to be an intrinsically geometric quantity.[17]

There is, however, one significant feature which these two entropies share in common. The existence of a nontrivial entropy reflects in each instance the fact that one considers only an incomplete description of the system. If, as a practical matter, an observer is unable to measure complicated correlations between degrees of freedom, his knowledge of the system is only partial. And similarly, given that an observer cannot probe the insides of a black hole without being lost forever to the external Universe, the existence of such a black hole in his system imposes an inherent limitation upon what he can discover about its state.

If one starts with a field characterized initially by comparatively weak and insignificant correlations amongst degrees of freedom, a knowledge of the $g(A)$'s and the associated relevant μ_R constitutes a quite reasonable approximation to a complete characterization of the system, so that the field entropy S will be quite small. However, as the evolving dynamics leads to the generation of more significant and detailed correlations, the $g(A)$'s provide an increasingly incomplete characterization of the state of the system. One knows less and less about the true state of the system, and this loss of information is reflected by the fact that $S(t)$ will increase.

The scattering of a scalar field Φ by a Schwarzschild black hole admits to a similar interpretation. An initial state of the field, defined with support only outside the hole, can, at least in principle, be measured with arbitrary precision (modulo, perhaps, uncertainties in the form of the initial correlations). But there is no way that an observer can measure directly the complete final state: the portion of the field captured by the hole will remain, at least classically, forever inaccessible. This scattering experiment entails an intrinsic loss of information, and, as such, should entail also an increase in the entropy. Indeed, this is entirely consistent with the geometric notion of black hole entropy. The piece of the

field captured by the hole will presumably have a non-negative total energy, so that, as a consequence of the scattering process, the mass M of the hole will increase. This, however, corresponds to an increase in the area $A = 16\pi M^2$ of the hole and a concomitant increase in the entropy $S_{BH} = A/4$. Indeed, Zurek and Thorne[30] have very much adopted this point of view in their "derivation" of the generalized second law of thermodynamics for a system containing a black hole.

It is important to stress that, even in the presence of a black hole, the fundamental equations characterizing the evolution of the field remain deterministic. Given the specification of initial data, an observer could, at least in principle, predict uniquely the subsequent evolution of the system if he knew already what was inside the hole. Formally, at least, one can define a field theory in a Schwarzschild space-time, even inside the event horizon. It then makes sense to introduce a density matrix μ which will satisfy a linear Liouville equation, and, given this Liouville equation, one concludes that the "conventional" entropy $-\operatorname{Tr}\mu\log\mu$ will be a constant of the motion. The point, however, is that the total μ is not accessible to an observer outside the hole. What is relevant to him is instead the "piece" of the density matrix μ_r which could be measured directly by someone outside the hole. And, as discussed, e.g., by Sorkin and his co-workers,[31] the entropy $-\operatorname{Tr}\mu_r\log\mu_r$ need not be conserved.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge fruitful collaborations with Salman Habib and Bei Lok Hu. I am also grateful to Chris Stevens for finding the patience to explain the subtleties of quantum field theory in curved spaces to a rank novice.

Limited financial support was provided by the National Science Foundation and by the Center for Theoretical Physics at the University of Maryland.

[1]L. Boltzmann, Wien. Ber. **66**, 275 (1872).
[2]E. T. Jaynes, Am. J. Phys. **33**, 391 (1965).
[3]The notion of a "subdynamics," at least in its most elegant form, may be attributed primarily to the so-called Brussels school of statistical mechanics. The classic paper on the subject is R. Balescu and J. Wallenborn, Physica (Utrecht) **54**, 477 (1971). A more systematic and leisurely discourse may be found in R. Balescu, *Equilibrium and Non-Equilibrium Statistical Mechanics* (Wiley, New York, 1975). A more general, but less elegant, approach has been considered by such authors as C. R. Willis and R. H. Picard, Phys. Rev. A **9**, 1343 (1974) or H. E. Kandrup, Astrophys. J. **244**, 316 (1981).
[4]See, e.g., M. Haggerty and G. Severne, Adv. Chem. Phys. **35**, 119 (1977), and references contained therein.
[5]H. E. Kandrup, Class. Quantum Grav. **3**, L 55 (1986).
[6]B. L. Hu and H. E. Kandrup, Phys. Rev. D **35**, 1776 (1987).
[7]H. E. Kandrup, J. Math. Phys. **26**, 2850 (1985).
[8]Consistent with the point of view adopted by such authors as R. Hakim, J. Math. Phys. **8**, 1315 (1967) or W. Israel and H. E. Kandrup, Ann. Phys. (NY) **152**, 30 (1984), the point of view adopted in this paper is that one can define a classical theory of nonequilibrium statistical mechanics for any dynamical system with well-defined equations of motion, whether or not these equations derive from a Hamiltonian. The notion of an "equilibrium" theory will presumably require the existence of a time-independent H, but an analog of a Liouville equation exists quite generally, provided only that one can implement the notion of conservation of probability.
[9]S. Habib, preprint.
[10]W. Israel and H. E. Kandrup, Ann. Phys. (NY) **152**, 30 (1984).
[11]See, e.g., W. Israel, in *General Relativity, Papers in Honour of J. L. Synge*, edited by L. O'Raifertaigh (Clarendon, Oxford, 1972), pp. 201–241.
[12]The existence of a time-independent Hamiltonian is, however, necessary if one is to implement the especially elegant approach to nonequilibrium statistical mechanics developed by the Brussels school.
[13]D. W. Sciama, P. Candelas, and D. Deutsch, Adv. Phys. **30**, 327 (1981).
[14]See, e.g., the book by Balescu quoted in Ref. 3.
[15]S. R. de Groot, W. A. van Leeuwen, and Ch. G. van Weert, *Relativistic Kinetic Theory, Principles and Applications* (North-Holland, Amsterdam, 1980).
[16]E. A. Calzetta, S. Habib, and B. L. Hu, submitted to Phys. Rev. D.
[17]J. D. Bekenstein, Phys. Rev. D **7**, 2333 (1973); S. W. Hawking, Commun. Math. Phys. **43**, 199 (1975).
[18]N. D. Birrell and P. C. W. Davies, *Quantum Fields in Curved Spaces* (Cambridge U. P., Cambridge, 1982).
[19]H. E. Kandrup, J. Math. Phys. **25**, 3286 (1984).
[20]L. Parker, in *Asymptotic Structure of Spacetime*, edited by F. P. Esposito and L. Witten (Plenum, New York, 1977), pp. 107–226; Ya. B. Zel'dovich, in *Physics of the Expanding Universe*, edited by M. Demianski (Springer, New York, 1979), pp. 60–80.
[21]C. W. Misner, Phys. Rev. Lett. **22**, 1071 (1969).
[22]B. L. Hu, Phys. Rev. D **9**, 3263 (1974).
[23]The "spirit" of the exposition here is due to Balescu and his co-workers, although the actual manipulations described in this section parallel more closely the approach in the paper of Willis and Picard quoted in Ref. 3.
[24]The idea of using projection operators in statistical mechanics was first systematized by R. Zwanzig, in *Lectures in Theoretical Physics*, edited by W. E. Brittin, B. W. Downes, and J. Downes (Interscience, New York, 1961), pp. 106–141.
[25]As discussed, e.g., by the authors cited in Ref. 3, this third criterion is not just convenient practically, but in fact constitutes a deep and fundamental consistency check on the entire formalism. The conventional Brussels school would demand further that P be explicitly time independent, but, given that L itself may well be a function of t, this will not in general be possible. The application of these ideas to a time-dependent system has been considered in a Newtonian sense by H. E. Kandrup and S. Hill Kandrup, Astrophys. J. **277**, 1 (1984), where they were used to drive an exact closed equation for the evolution of the galaxy–galaxy pair correlation function. That successful application shows that these ideas can, at least in principle, be used to construct very different types of subdynamics.
[26]This is, e.g., the approach adopted by the Brussels school.
[27]See, e.g., R. P. Feynman, *Statistical Mechanics, A Set of Lectures* (Benjamin, Reading, MA, 1972).
[28]L. D. Landau, Phys. Z. Sowjetunion **10**, 154 (1936).
[29]D. Lynden-Bell and R. Wood, Mon. Not. R. Astron. Soc. **138**, 495 (1968).
[30]W. H. Zurek and K. S. Thorne, Phys. Rev. Lett. **54**, 2171 (1985).
[31]L. Bombelli, R. K. Koul, J. Lee, and R. D. Sorkin, Phys. Rev. D. **34**, 373 (1986).

A phase cell approach to Yang–Mills theory. II. Analysis of a mode

Paul Federbush
Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48109
Calvin Williamson
Department of Mathematics, University of Missouri—Columbia, Columbia, Missouri 65211

(Received 18 November 1986; accepted for publication 30 January 1987)

Properties of a mode announced in a previous paper are proved. This involves some complicated calculations in linear algebra, and observation of the structure of a function of several complex variables.

I. A_1' AND A_2', LINEAR ALGEBRA

$A_i'(p)$ is given by (3.12) of Ref. 1. We will consider a mode at level zero, specified by bond assignments at level zero, having only one bond assignment different from zero. We choose the bond to be the bond in the $+1$ direction from the origin, and the assignment to it to be unity. By invariance under certain interchanges of coordinate axes it is sufficient to find A_1' and A_2' to determine A_i'.

With i fixed, (3.12) of Ref. 1 is a matrix product, ABC, with A, B, C, respectively, a (1×6) matrix, (6×6) matrix, and (6×1) matrix. The coordinates of the six-dimensional space are labeled by oriented plaquette directions, for which we choose the ordering

$$(1,2),\ (2,3),\ (3,4),\ (1,3),\ (1,4),\ (2,4). \tag{1.1}$$

We find it convenient to introduce certain special notation and matrices. We let

$$f_i \equiv (e^{-ip_i} - 1) \tag{1.2}$$

and D be the 6×6 diagonal matrix with

$$D_{(i,j),(i,j)} \equiv f_i f_j \tag{1.3}$$

(here labeling rows and columns of D by the associated oriented plaquette directions). We let

$$\langle h(p)\rangle \equiv \sum_n \frac{1}{(p+2\pi n)^2} \prod_{i=1}^{4} \left| \frac{e^{ip_i}-1}{(p_i+2\pi n_i)} \right|^2 h(p+2\pi n), \tag{1.4}$$

where h is any function of $p=(p_1,p_2,p_3,p_4)$.

We now detail the matrices occurring in (3.12) of Ref. 1. We first view the (6×1) matrix

$$\begin{matrix} 1/p_1^2+1/p_2^2 \\ -1/p_2^2 \\ 0 \\ 1/p_1^2 \\ 1/p_1^2 \\ -1/p_2^2 \end{matrix} \quad \begin{matrix} -1/p_2^2 \\ 1/p_2^2+1/p_3^2 \\ -1/p_3^2 \\ 1/p_3^2 \\ 0 \\ 1/p_2^2 \end{matrix} \quad \begin{matrix} 0 \\ -1/p_3^2 \\ 1/p_3^2+1/p_4^2 \\ -1/p_3^2 \\ 1/p_4^2 \\ 1/p_4^2 \end{matrix} \quad \begin{matrix} 1/p_1^2 \\ 1/p_3^2 \\ -1/p_3^2 \\ 1/p_1^2+1/p_3^2 \\ 1/p_1^2 \\ -1/p_2^2 \end{matrix}$$

Here $\langle M_0\rangle$ is a positive Hermitian matrix, and w_1 is in its range. We deal with the limit

$$\lim_{\epsilon\to 0^+} (\langle M_0\rangle + \epsilon)^{-1} w_1. \tag{1.14}$$

$$\beta_\gamma e^{i\hat{\gamma}\cdot p} \equiv v_1 \equiv \bar{f}_1^{-1}\bar{D}w_1, \tag{1.5}$$

$$w_1 \equiv (1,0,0,1,1,0)^T, \tag{1.6}$$

where the bars indicate complex conjugates. We next view the left (1×6) matrix, two different vectors depending on whether one is studying A_1' or A_2'. With r_L defined by

$$r_L \equiv -\frac{i}{(2\pi)^2}\prod_k \left(\frac{e^{-ip_k}-1}{p_k}\right), \tag{1.7}$$

we have

$$\tilde{P}_1 = r_L(1/p_1)w_1^T D \tag{1.8}$$

and

$$\tilde{P}_2 = r_L(1/p_2)w_2^T D \tag{1.9}$$

with

$$w_2 \equiv (-1,1,0,0,0,1)^T. \tag{1.10}$$

At this point we may write (3.12) of Ref. 1 as

$$A_i'(p) = (1/p^2)\tilde{P}_i M^{-1} v_1. \tag{1.11}$$

Recall M is singular; and M^{-1} is defined on suitable vectors as $\lim_{\epsilon\to 0^+}(\epsilon + M)^{-1}$.

Substituting (1.5) and (1.8) or (1.9) into (1.11) we get

$$A_i' = (-1/p^2) r_L \bar{f}_1^{-1}(1/p_i) w_i^T \langle M_0\rangle^{-1} w_1, \quad i=1,2, \tag{1.12}$$

where

$$M = \bar{D}\langle M_0\rangle D \tag{1.13}$$

and the inverse of $\langle M_0\rangle$ taken in the same sense as the inverse of M. We now display M_0, itself, in its full glory:

$$\begin{matrix} 1/p_1^2 \\ 0 \\ 1/p_4^2 \\ 1/p_1^2 \\ 1/p_1^2+1/p_4^2 \\ 1/p_4^2 \end{matrix} \quad \begin{matrix} -1/p_2^2 \\ 1/p_2^2 \\ 1/p_4^2 \\ 0 \\ 1/p_2^2 \\ 1/p_2^2+1/p_4^2 \end{matrix} .$$

This limit may be taken by the following procedure. Let P be the (orthogonal) projection onto the range of $\langle M_0\rangle$, S. Let $m = P\langle M_0\rangle P$, a strictly positive Hermitian matrix. We invert m, in S, writing it as m^{-1}. One has for the result of (1.14)

 0022-2488/87/061416-04$02.50

$$m^{-1}w_1. \tag{1.15}$$

The observations of this paragraph reduce the computations of (1.12) to calculations in a three-dimensional vector space. We note that S is spanned by w_1, w_2, and $w_3 = (0,0,1,0,1,1)^T$.

We first present the results of the computation in (1.15):

$$\begin{aligned}\langle M_0\rangle^{-1}w_1 = (1/\mathscr{D})(&4bc + 4bd + 8cd, 4bd - 4cd,\\ &4bc - 4bd, 4bc + 8bd + 4cd,\\ &8bc + 4bd + 4cd, 4bc - 4cd)^T,\end{aligned} \tag{1.16}$$

where

$$\begin{aligned}a &= \langle 1/p_1^2\rangle, \quad b = \langle 1/p_2^2\rangle,\\ c &= \langle 1/p_3^2\rangle, \quad d = \langle 1/p_4^2\rangle,\end{aligned} \tag{1.17}$$

and

$$\mathscr{D} = 16\sum_k \prod_{j\neq k}\langle 1/p_j^2\rangle. \tag{1.18}$$

Here $\mathscr{D}$ has arisen as the determinant of the 3×3 matrix involved in the computation of m^{-1}.

Collecting our results and substituting into (1.12) we easily find

$$\begin{aligned}A_1' = \left(-r_L\bar{f}_1^{-1}\frac{1}{p^2}\right)\frac{1}{p_1}\frac{16}{\mathscr{D}}\Bigg(&\left\langle\frac{1}{p_2^2}\right\rangle\left\langle\frac{1}{p_3^2}\right\rangle\\ &+\left\langle\frac{1}{p_2^2}\right\rangle\left\langle\frac{1}{p_4^2}\right\rangle + \left\langle\frac{1}{p_3^2}\right\rangle\left\langle\frac{1}{p_4^2}\right\rangle\Bigg),\end{aligned} \tag{1.19}$$

$$A_2' = \left(-r_L\bar{f}_1^{-1}\frac{1}{p^2}\right)\frac{1}{p_2}\frac{16}{\mathscr{D}}\left(-\left\langle\frac{1}{p_3^2}\right\rangle\left\langle\frac{1}{p_4^2}\right\rangle\right). \tag{1.20}$$

II. A_i' AND $X(p)$

From (1.19) and (1.20), and the invariance under interchange of coordinate directions—except for the special one-direction—we find our expression for $A_i'(p)$,

$$A_i' = \left(-r_L\bar{f}_1^{-1}\frac{1}{p^2}\right)\frac{1}{p_i}\frac{16}{\mathscr{D}}l_i, \tag{2.1}$$

$$l_1 = \left\langle\frac{1}{p_2^2}\right\rangle\left\langle\frac{1}{p_3^2}\right\rangle + \left\langle\frac{1}{p_2^2}\right\rangle\left\langle\frac{1}{p_4^2}\right\rangle + \left\langle\frac{1}{p_3^2}\right\rangle\left\langle\frac{1}{p_4^2}\right\rangle, \tag{2.2}$$

$$l_i = -\prod_{j\neq 1,i}\left\langle\frac{1}{p_j^2}\right\rangle, \quad i\neq 1. \tag{2.3}$$

A little study of these expressions for the region near $p_1\sim p_2\sim p_3\sim p_4\sim 0$ shows that the A_i' have a singularity in the complex four-dimensional region on the surface $p^2 = 0$, of course hitting the real axis. We define $A_i^N(p)$, a gauge transformation of $A_i'(p)$, by

$$A_i^N(p) = A_i'(p) + p_iX(p), \tag{2.4}$$

with

$$X(p) = \left(-r_L\bar{f}_1^{-1}\frac{1}{(p^2)^2}\right)\frac{1}{\langle 1/p_1^2\rangle}\frac{1}{(1+p^2)^s}, \tag{2.5}$$

where s is an arbitrary integer, sufficiently large. The analytic properties of $A_i^N(p)$ will be studied in the next section.

III. ANALYTIC PROPERTIES OF $A_i^N(p)$

The analysis we follow herein is analogous to the similar analysis due to Gawedzki and Kupiainen in the appendix of Ref. 2. We take the analytic extensions of the expressions in Sec. II from real p to complex p. For example,

$$\bar{f}_i \to e^{ip_i} - 1$$

and

$$\left|\frac{e^{-ip_i}-1}{p_i+2\pi n_i}\right|^2 \to \frac{e^{-ip_i}-1}{p_i+2\pi n_i}\,\frac{e^{ip_i}-1}{p_i+2\pi n_i}.$$

We make the preliminary observation that f_i, D, $\mathscr{D}$, and $\langle\cdot\rangle$ are periodic functions of p, invariant under

$$p\to p + 2\pi n \tag{3.1}$$

(n is a four-vector with integer components). We note that r_L, p^2, and p_i are not periodic functions of p.

Because of the periodicity mentioned in the last paragraph it is natural to divide our results into the following three theorems.

Theorem 3.1 (Local analyticity): There is an $\epsilon_0 > 0$ such that in the domain, $\mathscr{D}_L$, specified by

$$-\pi \leqslant \mathrm{Re}\, p_j \leqslant \pi, \quad |\mathrm{Im}\, p_j| < \epsilon_0, \tag{3.2}$$

$A_i^N(p)$ is analytic.

Theorem 3.2 (Global analyticity): $A_i^N(p)$ is analytic in the domain, $\mathscr{D}_G$, specified by

$$|\mathrm{Im}\, p_j| < \epsilon_0. \tag{3.3}$$

Theorem 3.3 (Boundedness): Within the domain, $\mathscr{D}_B$, specified by

$$|\mathrm{Im}\, p_j| < \epsilon_0/2, \tag{3.4}$$

$A_i^N(p)$ satisfies bounds of the form

$$|A_i^N(p)| < c\prod_j \frac{1}{(|p_j|+1)}\,\frac{1}{|p^2|+1}. \tag{3.5}$$

In Theorems 3.2 and 3.3 the ϵ_0 is as defined in Theorem 3.1.

Theorems 3.2 and 3.3 are easy consequences of the form of $A_i^N(p)$ and Theorem 3.1. We will devote our attention entirely to the proof of Theorem 3.1, which is carried out in Sec. V.

IV. CONCLUSIONS

Equations (3.13)–(3.15) of Ref. 1 follow directly from Theorems 3.2 and 3.3 of the last section by standard techniques.

V. LOCAL ANALYTICITY

We must study $A_i^N(p)$ on $\mathscr{D}_L$ [of (3.2)]. We write $A_i^N(p)$ from (2.1)–(2.5),

$$A_i^N(p) = -r_L\bar{f}_1^{-1}\frac{1}{p^2}\left[\frac{1}{p_i}\frac{16}{\mathscr{D}}l_i + \frac{p_i}{p^2}\frac{1}{\langle 1/p_1^2\rangle}\frac{1}{(1+p^2)^s}\right]. \tag{5.1}$$

In studying analyticity we may remove the factor $-r_L$ and replace $\bar{f}_1^{-1}$ by p_1^{-1}. We thus study

$$a_i(p)=\frac{1}{p_1}\frac{1}{p^2}\left[\frac{1}{p_i}\frac{16}{\mathscr{D}}l_i+\frac{p_i}{p^2}\frac{1}{\langle 1/p_1^2\rangle}\frac{1}{(1+p^2)^s}\right]. \tag{5.2}$$

We introduce some notation

$$p_j^2\langle 1/p_j^2\rangle \equiv r_0(p)e_j(p)(1/p^2), \tag{5.3}$$

where $e_j(0)=1$ and

$$r_0(p)=\prod_{i=1}^{4}\left|\frac{e^{-ip_i}-1}{p_i}\right|^2. \tag{5.4}$$

It will be sufficient to study $a_1(p)$ and $a_2(p)$. With the notation of (5.3) and (5.4) we have

$$r_0(p)a_1(p)=\frac{1}{\Sigma_k p_k^2\Pi_{j\neq k}e_j}(p_4^2e_2e_3+p_3^2e_2e_4+p_2^2e_3e_4)+\frac{1}{p^2}\frac{p_1^2}{e_1}\frac{1}{(1+p^2)^s}, \tag{5.5}$$

$$r_0(p)a_2(p)=\frac{-p_2p_1}{\Sigma_k p_k^2\Pi_{j\neq k}e_j}e_3e_4+\frac{p_2p_1}{p^2}\frac{1}{e_1}\frac{1}{(1+p^2)^s}. \tag{5.6}$$

We now write $e_i(p)$ as

$$e_i(p)=1+\delta_i(p), \tag{5.7}$$

near $p=0$, δ_i is small. In (5.5) we substitute (5.7) and set $\delta_i=0$ to get

$$r_0(p)a_1^0(p)\equiv 1+\frac{p_1^2}{p^2}\left(\frac{1}{(1+p^2)^s}-1\right) \tag{5.8}$$

and

$$r_0(p)a_2^0(p)\equiv\frac{p_2p_1}{p^2}\left(\frac{1}{(1+p^2)^s}-1\right). \tag{5.9}$$

Writing

$$a_i(p)=a_i^0(p)+R_i(p), \tag{5.10}$$

we see from (5.8) and (5.9) that $a_i^0(p)$ are analytic in $\mathscr{D}_L$. This is the main "algebraic miracle" involved in showing local analyticity of $A_i^N(p)$. This algebraic miracle motivated, of course, our choice of $X_i(p)$, which, however, was far from unique. We turn our attention to $R_i(p)$. By showing the analyticity in $\mathscr{D}_L$ of R_i we will have completed the proof. (We have already used the analyticity in $\mathscr{D}_L$ of r_L, $p_1\bar{f}_1^{-1}$, and r_0^{-1}.)

All hinges on certain properties of e_i and δ_i in $\mathscr{D}_L$, which we now pursue.

VI. PROPERTIES OF THE $e_i(p)$ IN $\mathscr{D}_L$

(1) $e_i(p)$ is analytic, (6.1)

(2) $\delta_i(p)=p_i^2p^2h_i(p)$ with $h_i(p)$ analytic, (6.2)

(3) $e_i^{-1}(p)$ is analytic, (6.3)

(4) $\left[\frac{1}{\Sigma_k p_k^2\Pi_{j\neq k}e_j}-\frac{1}{p^2}\right]$ is analytic. (6.4)

It is quite immediate that (6.1)–(6.4) imply the analyticity of $R_i(p)$. We are faced with our final task, the proof of these properties. (1) and (2) follow from the form of $e_i(p)$ upon inspection. It is only (3) and (4) that must be studied.

We write $e_1(p)$ in elaborate fashion [$e_1(p)$ and $e_i(p)$, $i\neq 1$, have the same properties],

$$e_1(p)=1+p^2\sum_\rho\left[\chi_\rho(1)\prod_{i\notin\rho}p_i^2K_\rho(p)+\chi_{\rho c}(1)p_1^2\prod_{i\notin\rho}p_i^2K_\rho^1(p)\right], \tag{6.5}$$

$$K_\rho(p)=\sum_{n\sim\rho}\frac{1}{(p+2\pi n)^2}\prod_{i\notin\rho}\frac{1}{(p_i+2\pi n_i)^2}, \tag{6.6}$$

$$K_\rho^1(p)=\sum_{n\sim\rho}\frac{1}{(p+2\pi n)^2}\prod_{i\notin\rho}\frac{1}{(p_i+2\pi n_i)^2}\times\frac{1}{(p_1+2\pi n_1)^2}, \tag{6.7}$$

where

(a) ρ is a proper subset of (1,2,3,4);

(b) $\chi_\rho(1)=\begin{cases}1 & 1\in\rho,\\ 0 & 1\notin\rho,\end{cases}$

$\chi_{\rho c}(1)=\begin{cases}0 & 1\in\rho,\\ 1 & 1\notin\rho;\end{cases}$

(c) $n\sim\rho$ means that $\rho=\{i,n_i=0\}$.

We easily see, for p in $\mathscr{D}_L$, that

$$|K_\rho(p)|<m,\quad |K_\rho^1(p)|<m, \tag{6.8}$$

for some fixed m. And that

$$|\mathrm{Arg}\,K_\rho(p)|<\epsilon',\quad |\mathrm{Arg}\,K_\rho^1(p)|<\epsilon', \tag{6.9}$$

where ϵ' can be made arbitrarily small by choosing ϵ_0, of (3.2), suitably small. Again we have

$$\begin{aligned}&|\mathrm{Arg}\,p_i^2|<\epsilon'' \quad \text{if } |p_i|^2>\bar{\epsilon},\\ &|\mathrm{Arg}\,p^2|<\epsilon'' \quad \text{if } |p|^2>\bar{\epsilon},\end{aligned} \tag{6.10}$$

where ϵ'' and $\bar{\epsilon}$ can be fixed arbitrarily small, choosing ϵ_0 small enough. Picking ϵ', ϵ'', and $\bar{\epsilon}$ small enough, and using (6.8), we see $e_1(p)$ is invertible in $\mathscr{D}_L$ and so analytic implying (3), (6.3). [The terms on the right side of (6.5), individually, will either be small, or have small argument.]

We turn to the study of (4), (6.4). We write the brackets in (6.4) as

$$\frac{1}{p^2(1+g)}-\frac{1}{p^2}=\frac{1}{p^2}\frac{g}{1+g}, \tag{6.11}$$

this relation defining g. Property (4) will be proved by showing

(5) g/p^2 is analytic, (6.12)

(6) $(1+g)^{-1}$ is analytic. (6.13)

For g we find the formula

$$g=\sum_k\frac{p_k^2}{p^2}\left[\prod_{j\neq k}(1+\delta_j)-1\right]. \tag{6.14}$$

From (6.2) we see that (5) [(6.12)] holds. The proof that $(1+g)$ is invertible, and so analytic, follows quite immediately from the proof above that $e_1(p)$ is invertible.

ACKNOWLEDGMENTS

One of the authors (C.W.) was supported in part by the University of Missouri Research Council. This work was supported in part by the National Science Foundation under Grant No. PHY-85-02074.

[1]P. Federbush, "A phase cell approach to Yang–Mills theory. I. Modes, lattice-continuum duality," Commun. Math. Phys. **107**, 319 (1986).
[2]K. Gawedzki and A. Kupiainen," A rigorous block spin approach to massless lattice theories," Commun. Math. Phys. **77**, 31 (1980).

Quantum noncompact σ models[a)]

J. W. van Holten
NIKHEF-H, P.O. Box 41882, 1009 DB Amsterdam, The Netherlands

(Received 9 December 1986; accepted for publication 4 February 1987)

Two possible quantization procedures for field theories with noncompact symmetry are discussed: one in a positive-definite Hilbert space with negative-energy states and one in a Hilbert space with indefinite metric and positive-definite energy. The physical interpretation of these alternative procedures is explained in terms of the realization of the noncompact symmetry, which is broken in one case and not in the other. It is shown that in exactly solvable quantum σ models the noncompact symmetry is broken in the vacuum. Some arguments supporting the general validity of this result are given. It is concluded that in contrast to theories like quantum electrodynamics, the noncompact σ models are to be quantized in the positive-definite Hilbert space with negative-energy modes.

I. INTRODUCTION

Positivity of the energy of physical systems is one of the fundamental postulates of quantum theory. It is introduced to explain the existence of a stable ground state of the system. In addition, it is fundamental in establishing the second law of thermodynamics (i.e., the increase of entropy with energy). In particular, the kinetic energy of physical systems is usually assumed to be positive and therefore takes the form (after proper normalization of the variables):

$$T(\dot{q}_i) = \sum_i \frac{1}{2}\dot{q}_i^2 . \tag{1.1}$$

If the total energy is

$$H = T(\dot{q}_i) + \lambda V(q_i) , \tag{1.2}$$

and if $T(\dot{q}_i) \gg \lambda V(q_i)$, the system has an approximate $O(N)$ symmetry:

$$q_i' = R_{ij}q_j, \quad R_{ij} = R_{ji}^{-1} , \tag{1.3}$$

which becomes exact if either $V(q_i)$ is an invariant itself, $V(q_i) = f(q_i^2)$ (central forces), or if $\lambda \to 0$.

In systems with infinitely many degrees of freedom, like field theories, similar observations can be made. In particular, for fields ϕ_i described by a positive-definite Hamiltonian density

$$H = \frac{1}{2}\sum_i (\dot{\phi}_i^2 + (\nabla\phi_i)^2) + \lambda V(\phi_i) , \tag{1.4}$$

the kinetic term and gradient terms are $O(N)$ invariant, a consequence of the positivity of the theory. However, there are physically relevant theories that admit an (exact or approximate) noncompact symmetry, such as $O(N,M)$. Field theories of this type have Hamiltonian densities of the form

$$H = \frac{1}{2}\sum_i (\dot{\pi}_i^2 + (\nabla\pi_i)^2) - \frac{1}{2}\sum_r (\dot{\sigma}_r^2 + (\nabla\sigma_r)^2) + \lambda V(\sigma,\pi) , \tag{1.5}$$

where $\{\pi_i\}$ denotes a set of fields with positive energy and $\{\sigma_r\}$ a set of fields with negative energy. Of course, the appearance of degrees of freedom with negative energy contradicts the physical principles mentioned above. Although in some cases one may question the absolute validity of these principles, we study in this paper systems for which the degrees of freedom with negative energy are unphysical and can be eliminated by imposing suitable constraints, resulting in Hamiltonians which are bounded below. [For example, no instability arises in the absence of perturbations or interactions coupling to both the positive and the negative-energy modes; therefore, under special circumstances, it is possible to create systems with negative temperatures (Hamiltonians bounded above) in the laboratory.[1]]

Examples of theories of this kind are quite easy to find, and some are actually very familiar. As a first example, consider the $O(N,1)$ nonlinear σ model.[2-12] This model can be defined by the $O(N,1)$ invariant Lagrangian [our metric has signature $(+,+,+,-)$],

$$\mathscr{L} = -\frac{1}{2}\sum_{i=1}^{N}(\partial_\mu\pi_i)^2 + \frac{1}{2}(\partial_\mu\sigma)^2 , \tag{1.6}$$

with the additional constraint

$$\sigma^2 - \boldsymbol{\pi}^2 = 1/g . \tag{1.7}$$

To prove, that the energy of the system is bounded below, we solve the constraint in terms of N scalars φ_i:

$$\boldsymbol{\pi} = \frac{\boldsymbol{\varphi}}{\sqrt{1-g\boldsymbol{\varphi}^2}}, \quad \sigma = \frac{1}{\sqrt{g}}\frac{1}{\sqrt{1-g\boldsymbol{\varphi}^2}} . \tag{1.8}$$

Then the Lagrangian becomes

$$\mathscr{L} = -\frac{1}{2}\frac{(\partial_\mu\boldsymbol{\varphi})^2}{(1-g\boldsymbol{\varphi}^2)^2} . \tag{1.9}$$

The conjugate momentum to $\boldsymbol{\varphi}$ is then

$$\boldsymbol{\Pi}_\varphi = \frac{\dot{\boldsymbol{\varphi}}}{(1-g\boldsymbol{\varphi}^2)^2} , \tag{1.10}$$

and the corresponding Hamiltonian is

$$H = \frac{1}{2}(1-g\boldsymbol{\varphi}^2)^2\boldsymbol{\Pi}_\varphi^2 + \frac{1}{2}\frac{(\nabla\boldsymbol{\varphi})^2}{(1-g\boldsymbol{\varphi}^2)^2} . \tag{1.11}$$

[a)] This is a combined and revised version of two earlier papers, which appeared as preprints WUB 84-9 (unpublished) and NIKHEF 85-12 (unpublished).

 0022-2488/87/061420-10$02.50

It is manifestly positive definite. Note, that the Lagrangian (1.9) still possesses the noncompact $O(N,1)$ invariance, but now realized nonlinearly:

$$\delta\varphi_i = a_{ij}\varphi_j + (1/\sqrt{g})(b_i + gb_i\varphi^2 - 2g\mathbf{b}\cdot\boldsymbol{\varphi}\varphi_i)\,. \tag{1.12}$$

Hence, the physical scalars π_i or φ_i may be interpreted as Goldstone bosons describing the spontaneous breaking of an internal $O(N,1)$ symmetry to its compact subgroup $O(N)$ (Refs. 2,4,5). This example can easily be generalized to other nonlinear σ models on G/H, where G is noncompact and H its maximal compact subgroup. Such noncompact σ models form an intrinsic part of N-extended supergravity theories ($N \geqslant 4$), where they describe the scalar partners of the graviton.[13]

A second example of a constrained system with noncompact symmetry is provided by Maxwell's theory of electromagnetism. In this theory, the fundamental fields may be represented by a vector potential A_μ, from which one can compute electric and magnetic field strengths $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Since the physically observable field strengths are invariant under a gauge transformation $A_\mu \to A'_\mu = A_\mu + \partial_\mu \Lambda$, we can impose the Lorentz condition $\partial\cdot A = 0$. The Maxwell equations which describe the dynamics of the electromagnetic fields can then be obtained from the Lagrangian density

$$\mathscr{L} = -\tfrac{1}{2}(\partial_\mu \mathbf{A})^2 + \tfrac{1}{2}(\partial_\mu A_0)^2\,, \tag{1.13}$$

with the fields constrained by

$$\partial\cdot A = \nabla\cdot\mathbf{A} - \dot{A}_0 = 0\,. \tag{1.14}$$

This Lagrangian is manifestly invariant under $O(3,1)$ rotations on the fields $(\mathbf{A},A_0)$. Of course, these transformations are just the standard Lorentz transformations acting on the four-vector potential A_μ. The correct Hamiltonian density for the electric and magnetic fields is

$$H = \tfrac{1}{2}(\mathbf{E}^2 + \mathbf{B}^2)\,, \tag{1.15}$$

where $\mathbf{E} = \nabla A_0 - \dot{\mathbf{A}}$ and $\mathbf{B} = \nabla \times \mathbf{A}$, as usual. Again the Hamiltonian is positive definite.

A third example worth mentioning here is the theory of the relativistic string.[14] The free string may be described by the action[15]

$$S = \frac{1}{2}\int d^2\xi\, \sqrt{g} g^{ab}\, \partial_a X^\mu\, \partial_b X^\nu\, \eta_{\mu\nu}\,. \tag{1.16}$$

The variables X^μ represent the string coordinates, whilst g_{ab} is the metric of the two-dimensional world sheet of the string; $\eta_{\mu\nu}$ is the Lorentzian metric of the D-dimensional space-time in which the string moves. Clearly, the action is invariant under the (noncompact) D-dimensional Lorentz transformations. The 2-D metric g is not a dynamical degree of freedom, but may be interpreted as a kind of Lagrange multiplier imposing constraints

$$T_{ab} = \partial_a X^\mu\, \partial_b X_\mu - \tfrac{1}{2} g_{ab} g^{cd}\, \partial_c X^\mu\, \partial_d X_\mu = 0\,. \tag{1.17}$$

Again, these constraints guarantee the positivity of the energy in this theory.[16]

Thus we see that constrained systems with noncompact symmetries are quite common. Note, however, that there is an important difference between the first example and the other two examples: in the nonlinear σ model we could interpret the noncompact symmetry as a spontaneously broken internal symmetry, whereas the noncompact symmetry of the electromagnetic field and the relativistic string is Lorentz invariance (a space-time symmetry) which normally we do not expect to be broken.

In this paper we address the problem of quantizing systems with noncompact symmetries, and we show that the physical realization of the noncompact symmetry (spontaneously broken or not broken) determines the correct quantization scheme. We illustrate the general principles only for a few simple cases like the $O(N,1)$ model and Maxwell's theory, but the results carry over directly to more complicated models. Applications have been given elsewhere,[11] but a brief discussion of the consistency of the quantized theories is presented in the last section of this paper.

II. CANONICAL QUANTIZATION OF THE NEGATIVE-ENERGY MODES

Free fields may be considered as collections of harmonic oscillators. Therefore the study of harmonic oscillators is directly applicable to the case of relativistic fields. In this section we discuss two different quantization procedures for oscillators with negative kinetic energy.[17–20] Consider a classical harmonic oscillator with negative energy and Lagrangian

$$L = -\tfrac{1}{2}\dot{q}^2 + \tfrac{1}{2}\omega^2 q^2\,. \tag{2.1}$$

Note, that owing to the time-reversal invariance of L, the action $\int L\,dt$ is obtained from that of the positive-energy oscillator by reversing time: $t \to -t$. The first method to construct a quantized version of this system uses the standard correspondence between Poisson brackets and commutators. For the momentum and Hamiltonian of the classical theory defined by (2.1) we obtain

$$p = \frac{\partial L}{\partial \dot{q}} = -\dot{q}, \quad H = p\dot{q} - L = -\frac{1}{2}(p^2 + \omega^2 q^2)\,. \tag{2.2}$$

Clearly the classical Hamiltonian is negative definite (hence bounded above). The Hamiltonian equations of motion are

$$\dot{q} = \{H,q\}, \quad \dot{p} = \{H,p\}\,, \tag{2.3}$$

where the Poisson brackets are defined by

$$\{A,B\} = \frac{\partial A}{\partial p}\frac{\partial B}{\partial q} - \frac{\partial B}{\partial p}\frac{\partial A}{\partial q}\,. \tag{2.4}$$

It follows that

$$\{p,q\} = 1. \tag{2.5}$$

According to the correspondence principle, one obtains a quantum theory by replacing the coordinates and momenta by operators $\hat{p}$ and $\hat{q}$ and the Poisson bracket $\{\,,\}$ by a commutator $(i/\hbar)\,[\,,]$. Thus

$$[\hat{p},\hat{q}] = -i\hbar\,. \tag{2.6}$$

As for the standard harmonic oscillator, we introduce the creation and annihilation operators defined by

$$\hat{a} = (1/\sqrt{2\hbar\omega})(\omega\hat{q} + i\hat{p}), \quad \hat{a}^\dagger = (1/\sqrt{2\hbar\omega})(\omega\hat{q} - i\hat{p})\,. \tag{2.7}$$

They satisfy the commutation relation

$$[\hat{a},\hat{a}^\dagger] = 1. \tag{2.8}$$

With these operators one can construct a complete Hilbert space of orthonormal states for the negative-energy oscillator, starting from the ground state defined by

$$\hat{a}|0\rangle = 0, \quad \langle 0|0\rangle = 1 . \tag{2.9}$$

This state is well defined, as may be checked in the coordinate representation, in which

$$|0\rangle \to \psi_0 = ce^{-(\omega/2\hbar)q^2} . \tag{2.10}$$

The excited states are

$$|n\rangle = ((\hat{a}^\dagger)^n/\sqrt{n!})|0\rangle, \quad \langle n|m\rangle = \delta_{n,m} . \tag{2.11}$$

Thus all states have positive norm. The Hamiltonian reads

$$\hat{H} = -\hbar\omega(\hat{a}^\dagger\hat{a} + \tfrac{1}{2}) . \tag{2.12}$$

Its eigenvalues are negative definite:

$$E_n = -(n + \tfrac{1}{2})\hbar\omega, \quad n = 0,1,2,\ldots . \tag{2.13}$$

The time dependence of the operators in the Heisenberg representation follows from the Schrödinger equation

$$i\hbar\frac{d\hat{a}}{dt} = [\hat{a},\hat{H}] . \tag{2.14}$$

This gives

$$\hat{a}(t) = \hat{a}(0)e^{i\omega t}, \quad \hat{a}^\dagger(t) = \hat{a}^\dagger(0)e^{-i\omega t} . \tag{2.15}$$

Since ω is taken positive, the annihilation operator $\hat{a}$ here has the same time dependence as the creation operator $\hat{a}^\dagger$ in the case of the positive-energy oscillator, and vice versa. From Eqs. (2.7) and (2.15) one obtains

$$\hat{q}(t) = (\hbar/2\omega)^{1/2}(\hat{a}(0)e^{i\omega t} + \hat{a}^\dagger(0)e^{-i\omega t}) , \tag{2.16}$$

and for the Green's function:

$$\begin{aligned}\langle 0|T\hat{q}(t)\hat{q}(0)|0\rangle &= \frac{\hbar}{2\omega}(\theta(t)e^{i\omega t} + \theta(-t)e^{-i\omega t}) \\ &= -\frac{i\hbar}{2\pi}\int_{-\infty}^{\infty} dk\,\frac{e^{-ikt}}{k^2 - \omega^2 - i\epsilon} .\end{aligned} \tag{2.17}$$

This Green's function differs from that of the standard positive-energy oscillator in two respects: (1) the $i\epsilon$ prescription has the opposite sign; and (2) the residue has an extra minus sign as well.

The above results are easily understandable from the point of view of time reversal, since negative energies and positive time direction are equivalent to positive energies and negative time direction. The notions of advanced and retarded Green's functions have therefore been interchanged. We now contrast this treatment with an alternative method of quantization involving negative-norm states. This method results in positive energy expectation values. As noted already below Eq. (2.15) positive energies require interchanging the role of creation and annihilation operators.[17] Thus if we define

$$\begin{aligned}\hat{b}(t) &= \hat{b}(0)e^{-i\omega t} = \hat{a}^\dagger(t), \\ \hat{b}^\dagger(t) &= \hat{b}^\dagger(0)e^{i\omega t} = \hat{a}(t) ,\end{aligned} \tag{2.18}$$

then

$$\hat{q}(t) = (\hbar/2\omega)^{1/2}(\hat{b}(0)e^{-i\omega t} + \hat{b}^\dagger(0)e^{i\omega t}) . \tag{2.19}$$

The Hamiltonian becomes

$$\hat{H} = \hbar\omega(-\hat{b}^\dagger\hat{b} + \tfrac{1}{2}) . \tag{2.20}$$

In spite of its somewhat unusual appearance, this represents an oscillator with positive energy. The difference with the ordinary harmonic oscillator is that $\hat{b},\hat{b}^\dagger$ satisfy unphysical equal-time commutation relations[17,18]:

$$[\hat{b},\hat{b}^\dagger] = -1 . \tag{2.21}$$

If we define the energy eigenstates by

$$\hat{b}\,|\underline{0}\rangle = 0, \quad \langle\underline{0}|\underline{0}\rangle = 1 , \tag{2.22}$$

for the ground state, and

$$|\underline{n}\rangle = ((b^\dagger)^n/\sqrt{n!})|\underline{0}\rangle , \tag{2.23}$$

for the excited states, then the energy expectation values are $\langle E_n\rangle = \hbar\omega(n + \frac{1}{2})$, while the normalization of the states is

$$\langle\underline{n}|\underline{m}\rangle = (-)^n\delta_{n,m} . \tag{2.24}$$

Thus all odd eigenstates have negative norm. Using (2.19), (2.21), and (2.22), we can compute the two-point function:

$$\begin{aligned}\langle\underline{0}|T\hat{q}(t)\hat{q}(0)|\underline{0}\rangle &= -(\hbar/2\omega)(\theta(t)e^{-i\omega t} + \theta(-t)e^{i\omega t}) \\ &= -\frac{i\hbar}{2\pi}\int_{-\infty}^{\infty} dk\,\frac{e^{-ikt}}{k^2 - \omega^2 + i\epsilon} .\end{aligned} \tag{2.25}$$

Again, the residue is negative, but the $i\epsilon$ prescription is now the same as for the positive-energy harmonic oscillator.

Summarizing and comparing the two different quantum theories of the negative-energy oscillator, we note that they are related formally by the correspondence (2.18). However they differ in the choice of groundstates used; namely, the state $|0\rangle$ is annihilated by the operator $\hat{a}$, while the state $|\underline{0}\rangle$ is annihilated by $\hat{b} = \hat{a}^\dagger$. Because $|0\rangle \neq |\underline{0}\rangle$, the two procedures correspond to different dynamical realizations of the system. In this respect one may think of the relation (2.18) as a kind of Bogoliubov transformation between the groundstates $|0\rangle,|\underline{0}\rangle$. An order parameter can be introduced as $\langle\hat{b}^\dagger\hat{b}\rangle$, which is positive in the state $|0\rangle$, and vanishes in the other one, $|\underline{0}\rangle$.

III. REALIZATION OF NONCOMPACT SYMMETRIES IN QUANTUM SYSTEMS

In order to study the interpretation of the two quantization schemes in the context of systems with noncompact symmetry, it suffices to consider the simple example of two oscillators with opposite energy and an $O(1,1)$ symmetry:

$$L = \tfrac{1}{2}(\dot{x}_1^2 - \dot{x}_2^2) - \tfrac{1}{2}\omega^2(x_1^2 - x_2^2) . \tag{3.1}$$

This is just the linear $O(1,1)$ sigma model in $(0+1)$ dimensions, invariant under the infinitesimal $O(1,1)$ transformations

$$\delta x_1 = \alpha x_2, \quad \delta x_2 = \alpha x_1 . \tag{3.2}$$

If the quantum system is defined using the positive-definite Hilbert space with negative-energy states, the operators describing the two degrees of freedom of the system are

$$\begin{aligned}\hat{x}_1(t) &= (\hbar/2\omega)^{1/2}(\hat{a}_1 e^{-i\omega t} + \hat{a}_1^\dagger e^{i\omega t}) , \\ \hat{x}_2(t) &= (\hbar/2\omega)^{1/2}(\hat{a}_2 e^{i\omega t} + \hat{a}_2^\dagger e^{-i\omega t}) .\end{aligned} \tag{3.3}$$

The raising and lowering operators $\hat{a}_i^\dagger$, $\hat{a}_i$ satisfy the equal-time commutation relations

$$[\hat{a}_i,\hat{a}_j^\dagger] = \delta_{ij}, \quad (i,j) = (1,2) . \tag{3.4}$$

The Hamiltonian is

$$\hat{H} = \hbar\omega(\hat{a}_1^\dagger\hat{a}_1 - \hat{a}_2^\dagger\hat{a}_2) . \tag{3.5}$$

The noncompact symmetry is realized on the creation and annihilation operators by

$$\delta\hat{a}_1 = \alpha\hat{a}_2^\dagger, \quad \delta\hat{a}_2 = \alpha\hat{a}_1^\dagger . \tag{3.6}$$

One can show without difficulty, that this leaves the commutator (3.4) invariant. The transformations (3.6) are generated by a charge

$$\hat{Q} = \hat{a}_1\hat{a}_2 - \hat{a}_1^\dagger\hat{a}_2^\dagger = -\hat{Q}^\dagger . \tag{3.7}$$

Its commutators with the $\hat{a}_i$, $\hat{a}_i^\dagger$ are

$$[\hat{Q},\hat{a}_1] = \hat{a}_2^\dagger, \quad [\hat{Q},\hat{a}_2] = \hat{a}_1^\dagger . \tag{3.8}$$

The $O(1,1)$ charge $\hat{Q}$ commutes with the Hamiltonian:

$$[\hat{Q},\hat{H}] = 0 . \tag{3.9}$$

Hence the Hamiltonian is invariant and the charge conserved.

We define the groundstate as the state with occupation numbers (0,0) (it is more appropriate to call this a vacuum state, since it is not the state of lowest energy):

$$\hat{a}_i|0\rangle = 0, \quad \langle 0|0\rangle = 1. \tag{3.10}$$

Since the two oscillators do not interact, no energy can be exchanged between the positive and negative-energy subsystems. Therefore the vacuum (3.10) is stable under perturbations that couple only to the positive-energy degree of freedom (as one expects when the negative-energy degree of freedom is unphysical and can be eliminated by imposing suitable constraints).

It is immediately evident, that the vacuum (3.10) is not $O(1,1)$ invariant:

$$\hat{Q}|0\rangle = -|1;1\rangle . \tag{3.11}$$

Here $|n;m\rangle$ denotes the state with occupation numbers (n,m) for the oscillators (1,2). By continuing to apply $\hat{Q}$, we can actually construct an infinite multiplet of states with the same (vanishing) energy. This is expected from the theory of unitary representations of noncompact groups.[19,20] It is easily explained here by observing that the commutator of $\hat{Q}$ with the occupation number operators $\hat{n}_i = \hat{a}_i^\dagger\hat{a}_i$ is nonvanishing, even when acting on the vacuum. We wish to stress that the noninvariance of the vacuum (3.10) is not a result of diagonalizing a wrong set of observables in searching for a ground state. Attempts to construct an $O(1,1)$ invariant state which does not have well-defined occupation numbers,

$$\Psi = \sum_{n,m} c_{n,m}|n;m\rangle \quad \text{with} \quad \hat{Q}\Psi = 0 , \tag{3.12}$$

fail, because such states turn out to have infinite norm. Thus we have a conserved charge $\hat{Q}$ (commuting with the Hamiltonian), but no invariant states. In this sense the noncompact symmetry is broken in this realization of the quantum $O(1,1)$ model.

The observations made for this simple system can be extended to the case of noncompact field theories as well. Thus a quantum noncompact field theory with a positive-definite Hilbert space does not have an invariant vacuum, although the charges connected with the noncompact symmetry may be conserved.

When the other quantization procedure is chosen, using a Hilbert space with indefinite metric, the dynamical variables of the theory can be expanded as

$$\begin{aligned} \hat{x}_1(t) &= (\hbar/2\omega)^{1/2}(\hat{a}_1 e^{-i\omega t} + \hat{a}_1^\dagger e^{i\omega t}) , \\ \hat{x}_2(t) &= (\hbar/2\omega)^{1/2}(\hat{a}_2 e^{-i\omega t} + \hat{a}_2^\dagger e^{i\omega t}) , \end{aligned} \tag{3.13}$$

where the previous $\hat{b}$ operators are now denoted by $\hat{a}_2,\hat{a}_2^\dagger$, and the fundamental commutators are

$$[\hat{a}_i,\hat{a}_j^\dagger] = (\tau_3)_{ij} . \tag{3.14}$$

The Hamiltonian is

$$\hat{H} = \hbar\omega(\hat{a}_1^\dagger\hat{a}_1 - \hat{a}_2^\dagger\hat{a}_2 + 1) . \tag{3.15}$$

In this case the noncompact $O(1,1)$ symmetry is realized by

$$\delta\hat{a}_1 = \alpha\hat{a}_2, \quad \delta\hat{a}_2 = \alpha\hat{a}_1 . \tag{3.16}$$

Again, the commutator (3.14) is invariant. The generator of the transformations is

$$\hat{Q} = \hat{a}_1^\dagger\hat{a}_2 - \hat{a}_2^\dagger\hat{a}_1 . \tag{3.17}$$

It commutes with the occupation number operators only on states with occupation numbers (0,0). There is a unique state with this property; thus the ground state defined by

$$\hat{a}_i|\underline{0}\rangle = 0, \quad \langle\underline{0}|\underline{0}\rangle = 1, \tag{3.18}$$

is $O(1,1)$ invariant: $\hat{Q}|\underline{0}\rangle = 0$.

This result holds more generally for systems with noncompact symmetries quantized in a Hilbert space with indefinite metric, which has only positive-energy states.[18]

Summarizing, the two quantization schemes for systems with noncompact symmetry correspond to a realization of the system in a positive-definite Hilbert space with negative-energy states and spontaneously broken noncompact symmetry, in the sense defined below Eq. (3.11); or alternatively a realization of the system in a Hilbert space with indefinite metric, but with strictly positive energy and an invariant vacuum.

IV. PATH INTEGRAL QUANTIZATION

As might be expected, the two different quantum realizations of theories with negative-energy modes also have different path-integral descriptions. The correct form of the path integral for the two quantum theories discussed in Secs. II and III can be established in various ways, for example, by discretizing time, from the holomorphic representation, or using the completeness relation for the eigenfunctions to compute the integral kernel for the Schrödinger equation. Here we do not give a derivation from first principles, but merely state the results. However, we compute the two-point functions and show that they reproduce those of Eqs. (2.17) and (2.25). For the case of noninteracting oscillators this amounts in fact to a complete proof of the results.

For the case of the negative-energy harmonic oscillator quantized in a positive definite Hilbert space the correct form of the path integral is

$$Z[j] = \langle 0,T\to\infty \,|\, 0,T\to -\infty\rangle_j$$
$$= \lim_{\epsilon\downarrow 0} \int Dq(t)\, \exp\frac{i}{\hbar}\int_{-\infty}^{\infty}\left(L + \frac{i}{2}\epsilon q^2 + jq\right)dt \tag{4.1}$$

with L as in Eq. (2.1). The integration over the $q(t)$ runs from $-\infty$ to $+\infty$ along the real axis. Strictly speaking, the $i\epsilon$ term is not necessary for convergence of the integral, but it is included to indicate the proper analytic continuation of the frequency ω to complex values. Such a continuation allows a smooth limit to the case of zero mass ($\mathrm{Re}\,\omega = 0$), which needs extra regularization (the determinant develops a zero mode). In fact the integral is well defined for all complex values of ω such that $\mathrm{Im}\,\omega \geqslant 0$; however it does not exist for $\mathrm{Im}\,\omega < 0$. [Of course, the determinant does exist for $\mathrm{Im}\,\omega < 0$; however, for such values it cannot be represented by the integral (4.1).] The analytic continuation made in (4.1) also defines the correct time ordering, as may be seen by computing the two-point function and comparing it to (2.17). This computation is facilitated by the inclusion of the external source term $j(t)q(t)$: the propagator is now obtained in standard fashion from the quadratic term in the expansion of $\log Z[j]$ in terms of the sources $j(t)$:

$$\langle q(t)q(0)\rangle = \frac{-\hbar^2}{Z[0]}\frac{\delta^2 Z[j]}{\delta j(t)\delta j(0)}\bigg|_{j=0}$$
$$= \frac{-i\hbar}{2\pi}\int_{-\infty}^{\infty} dk\, \frac{e^{-ikt}}{k^2 - \omega^2 - i\epsilon}. \tag{4.2}$$

This is indeed the same as in Eq. (2.17), including the $i\epsilon$ prescription. From the derivation of (2.17) it can be seen that the $i\epsilon$ prescription determines the $\theta(t)$ functions necessary for obtaining the correct time ordering of the product $\hat{q}(t)\hat{q}(0)$.

In contradistinction to this, the description of the system in a negative-norm Hilbert space is recovered from the path integral

$$\underline{Z}[j] = \langle \underline{0},T\to\infty \,|\, \underline{0},T\to -\infty\rangle$$
$$= \lim_{\epsilon\downarrow 0} \int Dq(t)\, \exp\frac{i}{\hbar}\int_{-\infty}^{\infty}\left(L - \frac{i\epsilon}{2}q^2 + jq\right)dt, \tag{4.3}$$

where the domain of integration of $q(t)$ is now from $-i\infty$ to $+i\infty$. The reason for this is, that in the negative-norm space the operators $\hat{q}$ have imaginary eigenvalues.[21] Here the $i\epsilon$ prescription indicates convergence of the integrals for $\mathrm{Im}\,\omega \leqslant 0$. It agrees with the time ordering in the canonical treatment (2.25), as follows from the two-point function:

$$\langle q(t)q(0)\rangle = -\frac{\hbar^2}{\underline{Z}[0]}\frac{\delta^2 \underline{Z}[j]}{\delta j(t)\delta j(0)}\bigg|_{j=0}$$
$$= -\frac{i\hbar}{2\pi}\int_{-\infty}^{\infty} dk\, \frac{e^{-ikt}}{k^2 - \omega^2 + i\epsilon}. \tag{4.4}$$

Comparison with the expression (2.25) shows this to be correct indeed.

V. NONCOMPACT σ MODELS

In the previous sections we have shown that there are two alternative methods of quantizing theories with noncompact symmetries, differing in the way the noncompact symmetry is realized in the spectrum. Now we must determine which quantization procedure to follow in actual physical theories. For Maxwell's theory of the electromagnetic field or the theory of the relativistic string the noncompact symmetry involved is Lorentz invariance. Because this symmetry is not broken, these theories must be quantized in a Hilbert space with indefinite metric. This leads to the standard Gupta–Bleuler formulation of quantum electrodynamics.[22]

The situation is different in the case of noncompact σ models. Below we present a number of arguments showing that in these theories the noncompact symmetry is broken (in the sense of Sec. III). Therefore we are forced to conclude that the quantum description requires one to work with an extended Hilbert space having positive-definite metric and negative-energy states.

An explicit proof of this statement can be given for a number of exactly solvable models. The first of these is the nonlinear $O(1,1)$ model. This model is a nonlinear counterpart of the model that was briefly discussed in its $(0+1)$-dimensional version in Sec. III. In line with the general discussion of nonlinear $O(N,1)$ models in Eqs. (1.6)–(1.12) it is defined by the Lagrangian

$$\mathscr{L} = \tfrac{1}{2}(\partial_\mu \sigma)^2 - \tfrac{1}{2}(\partial_\mu \pi)^2, \tag{5.1}$$

with the constraint

$$\sigma^2 - \pi^2 = 1/g > 0. \tag{5.2}$$

One may now proceed to solve the constraint as in Eq. (1.8), in terms of a single scalar field φ:

$$\pi = \frac{\varphi}{\sqrt{1-g\varphi^2}}, \quad \sigma = \frac{1}{\sqrt{g}}\frac{1}{\sqrt{1-g\varphi^2}},$$
$$\mathscr{L} = -\tfrac{1}{2}(\partial_\mu\varphi)^2/(1-g\varphi^2)^2. \tag{5.3}$$

This is invariant under the nonlinear Abelian transformations

$$\delta\varphi = (1/\sqrt{g})b(1-g\varphi^2). \tag{5.4}$$

However, in this case there exists a more convenient parametrization of the model, using a scalar field ϑ defined by

$$\pi = (1/\sqrt{g})\sinh(\sqrt{g}\vartheta), \quad \sigma = (1/\sqrt{g})\cosh(\sqrt{g}\vartheta). \tag{5.5}$$

In terms of this field ϑ the Lagrangian becomes

$$\mathscr{L} = -\tfrac{1}{2}(\partial_\mu\vartheta)^2. \tag{5.6}$$

Thus the theory describes a single real, massless scalar field, and is invariant under nonlinear $O(1,1)$ transformations of the form

$$\vartheta \to \vartheta' = \vartheta + \eta, \tag{5.7}$$

where η is a constant.

The importance of this result is, that the spectrum of a massless field is well known to be continuous. In particular, there is no normalizable ground state. This may be checked directly, by enclosing the system in a box of finite dimension L and decomposing the field into plane waves:

$$\vartheta(x) = \sum_{\mathbf{k}\in\mathbb{Z}^d} \vartheta_{\mathbf{k}}(t)e^{2\pi i\mathbf{k}\cdot\mathbf{x}/L}, \quad \vartheta^*_{\mathbf{k}} = \vartheta_{-\mathbf{k}}. \tag{5.8}$$

The Hamiltonian separates into a direct sum of Hamilto-

nians for the modes (5.8). Correspondingly, the wave function factorizes into an infinite product of wave functions, one for each mode. For all the modes with space momentum $\mathbf{k}\neq 0$, the Hamiltonian is that of a harmonic oscillator with frequency $\omega_k = (2\pi/L)|\mathbf{k}|$. Thus these modes have normalizable wave functions $\psi_{\mathbf{k}}$. For the zero mode ϑ_0 the Hamiltonian is that of a free, nonrelativistic particle in one dimension. Its eigenstates are plane waves, which are not normalizable. To get normalizable states, we have to form wave packets by linear superposition of the zero modes. Therefore the wave function for the free, massless scalar field has the form

$$\Psi[\vartheta_k] = \int_{-\infty}^{\infty} d\alpha\, c(\alpha) e^{i\alpha\vartheta_0} \prod_{\mathbf{k}\neq 0} \psi_{\mathbf{k}}[\vartheta_{\mathbf{k}}], \tag{5.9}$$

where

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} d\alpha |c(\alpha)|^2 = 1 .$$

The nonlinear $O(1,1)$ transformation (5.7) shifts the zero mode by η, but does not affect the other modes:

$$\delta\vartheta_0 = \eta; \quad \delta\vartheta_{\mathbf{k}} = 0, \quad \mathbf{k}\neq 0 . \tag{5.10}$$

The result of this transformation on the wave function (5.9) is that all $c(\alpha)$ are multiplied by a phase factor $\exp(i\alpha\eta)$. This is a unitary transformation, preserving the normalization, but changing the form of the superposition and hence the state of the system. In fact, the plane wave basis for the zero modes is an infinite-dimensional unitary representation of the Abelian $O(1,1)$ group. Therefore we have no normalizable $O(1,1)$-invariant ground state for the nonlinear $O(1,1)$ model.

The conclusion arrived at for the Abelian $O(1,1)$ model also holds for the non-Abelian $O(N,1)$ models. This has been proved in detail for the $O(3,1)$ model in $(0+1)$ dimensions by Velo and Wess,[6] and for the supersymmetric $O(2,1)$ model in $(0+1)$ dimensions by Davis *et al.*[3] In this last model there are in fact normalizable ground states with energy zero, separated by a finite energy gap from the rest of the spectrum, which is purely continuous. However, these zero energy states form two infinite-dimensional discrete series, which are unitary representations of $O(2,1)$ (Refs. 23, 24). Therefore there is still no $O(2,1)$ invariant vacuum. Such discrete series arise in fact for all supersymmetric $O(N,1)$ models with N even.[4] Therefore none of these models possesses an $O(N,1)$ invariant vacuum. That this situation generalizes to the case of the supersymmetric $O(N,1)$ field theories in $(d+1)$ dimensions ($d\geqslant 1$) is shown by using Wittens index theorem.[23]

Before proceeding to discuss more general arguments indicating the breaking of the noncompact symmetry in the nonlinear $O(N,1)$ models, we would like to return once more to the $O(1,1)$ model and discuss the construction of its Hilbert space of states in the formalism with constraints, as in (5.1) and (5.2). We consider the model in $(0+1)$ dimensions for simplicity, but this restriction is not essential. Before imposing the constraint, the wave functions of the system in the extended Hilbert space with positive-definite metric (and negative energies) are

$$\psi(\sigma,\pi) = \int_{-\infty}^{\infty}\frac{d\alpha}{2\pi}\int_{-\infty}^{\infty}\frac{d\eta}{2\pi} c(\alpha,\eta) e^{i(\alpha\pi-\eta\sigma)} . \tag{5.11}$$

We can now impose the constraint (5.2) as a constraint on the wave functions

$$(\sigma^2 - \pi^2 - (1/g))\psi(\sigma,\pi) = 0 . \tag{5.12}$$

This constraint is satisfied if the Fourier coefficients $c(\alpha,\eta)$ satisfy the $(1+1)$-dimensional Klein–Gordon equation:

$$(-\partial_\eta^2 + \partial_\alpha^2 - 1/g)c(\alpha,\eta) = 0 . \tag{5.13}$$

The solutions of this equation are of the form

$$c(\alpha,\eta) = \int_{-\infty}^{\infty} d\pi \int_{-\infty}^{\infty} d\sigma\, \delta(\sigma^2 - \pi^2 - (1/g)) \times \varphi(\sigma,\pi) e^{-i(\alpha\pi-\eta\sigma)} . \tag{5.14}$$

Taking $\eta = 0$ this reduces to the plane wave basis (5.9). Thus we can indeed obtain the correct physical states by imposing the constraint (5.2) on the states of the extended Hilbert space with positive-definite metric and negative-energy modes.

With the nonlinear transformation law of the physical fields under the $O(N,1)$ transformations (1.12), it is actually not surprising, that the $O(N,1)$ symmetry is broken to its compact $O(N)$ subgroup in the quantum nonlinear σ models. Assuming the quantum theory to exist, it follows from the Goldstone theorem. Denoting the compact $O(N)$ generators by $\hat{R}$, and the remaining noncompact ones by $\hat{N}$, we can show that the vacuum expectation value of the massless physical fields $\langle\varphi_i\rangle$ vanishes because of the $O(N)$ invariance of the vacuum. Consider an infinitesimal $O(N)$ transformation of $\langle\varphi_i\rangle$:

$$\delta(R)\langle\varphi_i\rangle = a_{ij}\langle\varphi_j\rangle = \langle 0|[\hat{R}(a),\varphi_i]|0\rangle = 0. \tag{5.15}$$

The vanishing of this expression follows, since the compact $O(N)$ symmetry of the vacuum may safely be assumed in the absence of a symmetry breaking potential. However, for a transformation in a noncompact direction we obtain

$$\delta(N)\langle\varphi_i\rangle = \langle 0|[\hat{N}(b),\varphi_i]|0\rangle = (1/\sqrt{g})b_j(\delta_{ij} + g[\delta_{ij}\langle\varphi_k^2\rangle - 2\langle\varphi_i\varphi_j\rangle]) . \tag{5.16}$$

For $N\geqslant 2$ the right-hand side cannot vanish, proving that the vacuum cannot be invariant under these transformations. For $N=1$ we have already proved the nonexistence of an invariant vacuum. We conclude that, provided the quantum theory is well defined, the vacuum of the nonlinear $O(N,1)$ σ model is not invariant under the noncompact symmetries. Finally we briefly consider the constraint (1.7). Taking vacuum expectation values on both sides of the equation, we get after proper renormalization:

$$\langle\sigma^2\rangle_{\text{ren}} - \langle\boldsymbol{\pi}^2\rangle_{\text{ren}} = 1/g_{\text{ren}} > 0 . \tag{5.17}$$

If the renormalized coupling constant is positive (and finite), we are forced to conclude that

$$\langle\sigma^2\rangle_{\text{ren}} \neq \langle\boldsymbol{\pi}^2\rangle_{\text{ren}} \quad \text{and} \quad \langle\sigma^2\rangle_{\text{ren}} > 0 . \tag{5.18}$$

The first condition can be avoided only if both vacuum expectation values become infinite:

$$(\langle\sigma^2\rangle_{\text{ren}}, \langle\boldsymbol{\pi}^2\rangle_{\text{ren}}) \to \infty .$$

This possibility is presumably realized in the $(1+1)$-dimensional models,[4,10] since for $d=1$ no properly massless scalars exist.[24] However, for $d\geqslant 2$, model calculations and the analysis of the $O(1,1)$ model above indicate that the vacuum expectation values are finite[4,9,11] and the conditions (5.18) must be satisfied. Again we are led to conclude that the $O(N,1)$ symmetry is broken.

The arguments presented above indicate, that for the noncompact nonlinear σ models the quantization in an extended Hilbert space for constrained fields (σ,π) should proceed with positive norm, but negative-energy states for the unphysical degrees of freedom σ. The energy of the full theory is nevertheless bounded below, because the constraint (1.7) [(5.2)] guarantees, that the excitation of negative-energy states can only take place, if simultaneously sufficiently many positive-energy modes of the π fields are excited as well.

VI. A MODEL COMPUTATION

The conclusion of Sec. V, that in the formulation with constraints noncompact σ models are to be quantized using a positive-definite Hilbert space with negative-energy states for the unphysical degrees of freedom, implies that the free propagators of the fields (σ,π) used in perturbation theory take the form

$$\begin{aligned}\langle\sigma(x)\sigma(0)\rangle &= -\int\frac{d^nk}{(2\pi)^n i}\,\frac{e^{ik\cdot x}}{k^2+m^2+i\epsilon},\\ \langle\pi_i(x)\pi(0)\rangle &= \delta_{ij}\int\frac{d^nk}{(2\pi)^n i}\,\frac{e^{ik\cdot x}}{k^2+m^2-i\epsilon},\end{aligned}\tag{6.1}$$

where n denotes the dimensionality of space-time; cf. Eqs. (2.17) and (4.2). In this section we show that this prescription leads to correct results in a simple toy-model computation. Although this model is strongly simplified, we believe it has all the necessary features to test the analytic continuation of m^2 implicit in Eqs. (6.1), as it appears for example in path integrals of the type (4.1) and (4.3). Instead of these infinite-dimensional integrals, we consider the ordinary double integral

$$\begin{aligned}Z_-(m_1^2,m_2^2) &= \int_{-\infty}^{\infty}d\pi\int_{-\infty}^{\infty}d\sigma\,\delta\!\left(\pi^2-\sigma^2+\frac{1}{g}\right)\\ &\quad\times e^{-(i/2)(m_1^2\pi^2-m_2^2\sigma^2)},\end{aligned}\tag{6.2}$$

where for definiteness we choose $m_2^2\geqslant m_1^2\geqslant 0$ and $g>0$. This is a zero-dimensional version of the nonlinear $O(1,1)$ model, with an additional $O(1,1)$ breaking for $m_1^2\neq m_2^2$. Inspection of the integral shows it to be well defined for $\operatorname{Im} m_1^2\leqslant\operatorname{Im} m_2^2$. The integral can be evaluated directly by first carrying out the integration over σ; this eliminates the δ function:

$$Z_-(m_1^2,m_2^2) = 2\sqrt{g}e^{im_2^2/2g}\int_0^{\infty}d\pi\,\frac{e^{(i/2)(m_2^2-m_1^2)\pi^2}}{\sqrt{1+g\pi^2}}.\tag{6.3}$$

It is convenient to redefine the parameters as follows:

$$\begin{aligned}&\mu^2=(1/4g)(m_1^2+m_2^2),\\ &\omega=(1/4g)(m_2^2-m_1^2),\quad s=1+2g\pi^2.\end{aligned}\tag{6.4}$$

Then

$$Z_-(m_1^2,m_2^2) = e^{i\mu^2}\int_1^{\infty}ds\,\frac{e^{i\omega s}}{\sqrt{s^2-1}}.\tag{6.5}$$

This integral is defined for $\operatorname{Im}\omega\geqslant 0$ $(\omega\neq 0)$ and all complex values of μ^2, in agreement with the complex domains of m_1^2,m_2^2. The result of (6.5) is a Hankel function of the first kind:

$$Z_-(m_1^2,m_2^2) = (i\pi/2)e^{i\mu^2}H_0^{(1)}(\omega).\tag{6.6}$$

Of course, the Hankel function can be defined in the whole complex ω plane, but the integral representation (6.6) exists only for $\operatorname{Im}\omega\geqslant 0$ (Ref. 24).

It is instructive to compare this result with the compact $O(2)$ version of this integral:

$$\begin{aligned}&Z_+(m_1^2,m_2^2)\\ &\quad=\int_{-\infty}^{\infty}d\pi\int_{-\infty}^{\infty}d\sigma\,\delta\!\left(\sigma^2+\pi^2-\frac{1}{g}\right)\\ &\qquad\times\exp\!\left(-\frac{i}{2}(m_1^2\pi^2+m_2^2\sigma^2)\right)\\ &\quad=2\sqrt{g}e^{-im_2^2/2g}\int_0^{1/\sqrt{g}}d\pi\,\frac{\exp[(i/2)(m_2^2-m_1^2)\pi^2]}{\sqrt{1-g\pi^2}},\end{aligned}\tag{6.7}$$

which exists for all complex values of m_1^2,m_2^2. With μ^2 and ω as in (6.4) and taking

$$s=-1+2g\pi^2,$$

the integral (6.7) becomes

$$Z_+(m_1^2,m_2^2) = e^{-i\mu^2}\int_{-1}^{+1}ds\,\frac{e^{i\omega s}}{\sqrt{1-s^2}}.\tag{6.8}$$

In contrast to (6.5), this integral is indeed well defined for all complex values of ω. It is an integral representation of the Bessel function of order zero:

$$Z_+(m_1^2,m_2^2) = \pi e^{-i\mu^2}J_0(\omega).\tag{6.9}$$

In actual quantum-field theoretical models, the path integrals usually cannot be evaluated directly and one must take recourse to perturbation theory. The constraint for (σ,π) is then imposed using a Lagrange multiplier α. For the integral (6.2) such a formulation is obtained by substitution of the Fourier representation of the δ function:

$$\begin{aligned}Z_-(m_1^2,m_2^2) &= \int_{-\infty}^{\infty}d\pi\int_{-\infty}^{\infty}d\sigma\int_{-\infty}^{\infty}\frac{d\alpha}{4\pi}\exp\!\left\{-\frac{i}{2}\left[m_1^2\pi^2-m_2^2\sigma^2+\alpha\left(\pi^2-\sigma^2+\frac{1}{g}\right)\right]\right\}\\ &= \int_{-\infty}^{\infty}d\pi\int_{-\infty}^{\infty}d\sigma\int_{-\infty}^{\infty}\frac{d\alpha}{4\pi}\exp\!\left\{-\frac{i}{2}\left[(m_1^2+\alpha)\pi^2-(m_2^2+\alpha)\sigma^2+\frac{\alpha}{g}\right]\right\}.\end{aligned}\tag{6.10}$$

The perturbation theory treatment is obtained by interchanging the order of integration, allowing the integrations over (σ,π)

to be performed first and that over α afterwards. For the integral (6.10) this interchange is allowed for $\operatorname{Im} m_1^2 \leqslant 0$ and $\operatorname{Im} m_2^2 \geqslant 0$, and gives

$$Z_-(m_1^2,m_2^2) = \int_{-\infty}^{\infty} \frac{d\alpha}{4\pi} \exp\left(-\frac{i\alpha}{2g}\right)\left\{\int_{-\infty}^{\infty} d\pi \int_{-\infty}^{\infty} d\sigma \exp\left(-\frac{i}{2}[(m_1^2+\alpha)\pi^2-(m_2^2+\alpha)\sigma^2]\right)\right\}$$
$$= \frac{1}{2}\int_{-\infty}^{\infty} d\alpha \frac{\exp(-i\alpha/2g)}{\sqrt{(m_1^2+\alpha)(m_2^2+\alpha)}}. \tag{6.11}$$

Now define

$$t = -((\alpha/2g)+\mu^2).$$

Then

$$Z_-(m_1^2,m_2^2) = \frac{1}{2} e^{i\mu^2} \int_{-\infty}^{\infty} dt \frac{e^{it}}{\sqrt{t^2-\omega^2}}. \tag{6.12}$$

A similar treatment of the compact model (6.7) requires $\operatorname{Im} m_1^2 \leqslant 0$, $\operatorname{Im} m_2^2 \leqslant 0$. In order to show that this reduces to the form (6.5), we must take care of the multivaluedness of the square root, and deal with the singularities at $t = \pm\omega$. The square root becomes single valued if we make cuts along the half-lines $[\omega,\infty)$ and $(-\infty,-\omega]$. For $\operatorname{Im}\omega > 0$ the integration does not encounter the cuts (Fig. 1). In the limit $\operatorname{Im}\omega \to 0$, the path of integration passes along the upper edge of the cut $(-\infty,-\omega]$ and the lower edge of the cut $[\omega,\infty)$. Now consider the contour C of Fig. 2. Since there are no singularities inside the contour, we have

$$\oint_C dz \frac{e^{iz}}{\sqrt{z^2-\omega^2}} = 0. \tag{6.13}$$

In the limit $R \to \infty$ the part of the integral along the semicircular arc vanishes. Therefore

$$\int_{-\infty}^{\operatorname{Re}\omega} dt \frac{e^{it}}{\sqrt{t^2-\omega^2}} = -\int_{\operatorname{Re}\omega = \omega - i\operatorname{Im}\omega}^{\omega + i\operatorname{Im}\omega} dt \frac{e^{it}}{\sqrt{t^2-\omega^2}} - \int_{\omega + i\operatorname{Im}\omega}^{\infty + 2i\operatorname{Im}\omega} dt \frac{e^{it}}{\sqrt{t^2-\omega^2}}. \tag{6.14}$$

Hence the contour of Fig. 1 may be deformed to pass along the upper and lower edges of the cut $[\omega,\infty)$, as in Fig. 3. Finally, the two parts above and below the cut contribute equal amounts, because the change in sign of the square root is balanced by the reversal of the direction of the path. The final result is therefore

$$Z_-(m_1^2,m_2^2) = e^{i\mu^2} \int_{\omega}^{\infty} dt \frac{e^{it}}{\sqrt{t^2-\omega^2}} = e^{i\mu^2} \int_1^{\infty} ds \frac{e^{i\omega s}}{\sqrt{s^2-1}}. \tag{6.15}$$

Thus the result (6.5) is reproduced and the representation of the integral by (6.11) is correct with $\operatorname{Im} m_1^2 \leqslant 0$, $\operatorname{Im} m_2^2 \geqslant 0$. Writing explicitly the real and imaginary parts as $m_{1,2}^2 \to m_{1,2}^2 \pm i\epsilon_{1,2}$, we can replace (6.11) in the limit $\epsilon_{1,2} \downarrow 0$ (such that the logarithm is single valued) by

$$Z_-(m_1^2,m_2^2) = \int_{-\infty}^{\infty} \frac{d\alpha}{4\pi} \exp\left(-\frac{i\alpha}{2g}\right) \times \left\{\exp\left(-\frac{1}{2}\ln(m_1^2+\alpha-i\epsilon_1) - \frac{1}{2}\ln(m_2^2+\alpha+i\epsilon_2)\right)\right\}. \tag{6.16}$$

This is the form that, generalized to higher-dimensional models, is the starting point for a perturbative analysis. It is now straightforward to establish the results

$$\langle\sigma^2\rangle = -2i\frac{\partial}{\partial m_2^2} Z_-(m_1^2,m_2^2) = \left\langle \frac{i}{m_2^2+\alpha+i\epsilon_2} \right\rangle_\alpha,$$
$$\langle\pi^2\rangle = 2i\frac{\partial}{\partial m_1^2} Z_-(m_1^2,m_2^2) = \left\langle \frac{-i}{m_1^2+\alpha-i\epsilon_1} \right\rangle_\alpha, \tag{6.17}$$

which are identical to Eqs. (6.1) for $n = 0$. It is obvious that the replacement of (π,σ) by fields $(\pi(x),\sigma(x))$ and $m_{1,2}^2$ by differential operators $\partial_\mu^2 + m_{1,2}^2$ does not alter these results. Hence the analytic continuation chosen for the propagators in Eq. (6.1) is consistent for the noncompact models and leads to a correct perturbative treatment of the quantum field theory,[11] provided that this theory is well defined.

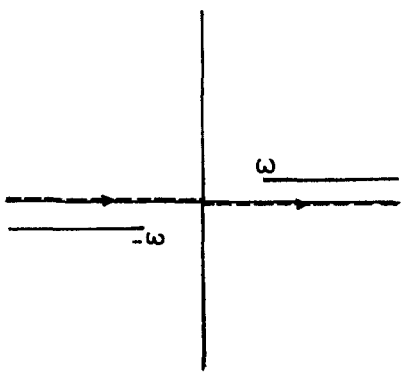

FIG. 1. Path of integration in complex t plane for Eq. (6.12).

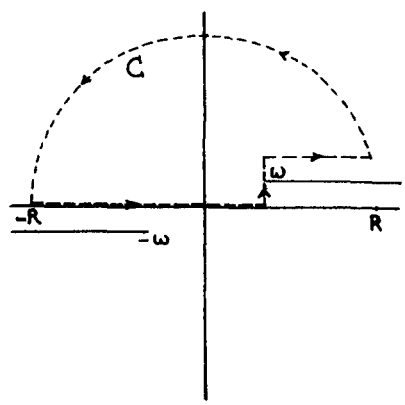

FIG. 2. Closed contour in z plane, Eq. (6.13).

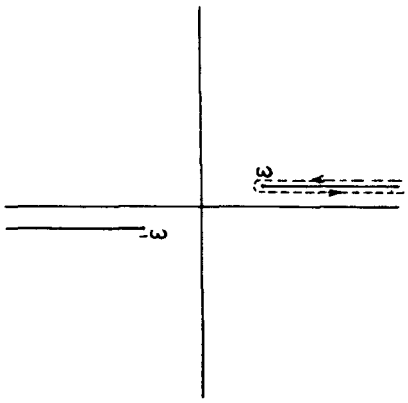

FIG. 3. Equivalent path of integration in t plane for $Z_-(m_1^2,m_2^2)$.

VII. DISCUSSION

The argument presented in this paper may be summarized as follows. There exist two different quantization schemes for field theories with noncompact symmetry of the kinetic energy operator: one with the noncompact symmetry broken, but a positive-definite Hilbert space and one with the noncompact symmetry not broken, but being realized in a Hilbert space of indefinite metric. It is argued from the examples of some explicitly solvable models that in nonlinear σ models with noncompact symmetry the noncompact symmetry is broken in the sense that there is no invariant vacuum state, even though the charges of the noncompact symmetry are conserved. The general validity of this result in nonlinear σ models is supported by other, more model-independent arguments. From this it is inferred that the quantization of the linearized formulation of such theories in an extended Hilbert space with constraints requires the use of a positive-definite Hilbert space with negative-energy modes for the unphysical degrees of freedom.

The question of whether such a quantization is consistent with causality and stability is far from obvious. As we have argued in Secs. I and V, the constraint (1.7) [(5.2)] is crucial in establishing the boundedness of the Hamiltonian and therefore it must necessarily be conserved in the full quantum theory. This requires the quantum theory including the constraint to be renormalizable, since otherwise counterterms might modify the constraint in such a way that the negative-energy modes can no longer be removed from the theory. Obviously, the renormalization must be carried out very carefully; for example, mixing bare and renormalized quantities in the analysis may easily lead to inconsistencies.[12]

Most studies have been done on $(0+1)$- and $(1+1)$-dimensional models. The $d=0$ models can be solved exactly and confirm the analysis presented here.[3,4,6] In $d=1$ exact results can be obtained in the large-N limit of $O(N,1)$ and noncompact $CP(N,1)$ models.[4,5,7-12] However, in this case the precise quantization of the negative-energy modes is not so important, because most of its contributions disappear in the limit $N\to\infty$. Nevertheless, negative-metric quantization has been shown to give rise to certain inconsistencies.[12] Also, as explained in Sec. V, the behavior of the noncompact models in $d=1$ is somewhat peculiar[4,10] because of the problems with infrared divergences for massless scalars.[25]

In $(3+1)$ dimensions a study has been made of the $SU(1,1)/U(1)$ nonlinear σ model,[11] which describes the scalar sector of $N=4$ supergravity.[13] The model has been analyzed at the one-loop level using the positive metric quantization. At this level it was shown to be renormalizable, with the energy bounded below.[11] These results can in fact be extended to the case of $SU(N,N)/SU(N)\times SU(N)\times U(1)$ models. However, it is not known what happens when one includes higher loop contributions. Some arguments against consistency of the models at higher loops were coined by the authors of Ref. 12, but we believe their arguments to be inconclusive (Ref. 26). In particular, the derivation of cutting rules and the conditions of unitarity are modified by the use of negative-energy modes, since their propagator has a nonstandard analytic continuation. Also, in the context of N-extended supergravity, it is known that the models are finite to two loops in perturbation theory.[27] Thus it seems that any inconsistencies are suppressed, at least till the next order of computation. This may be particularly relevant for gauged $N=8$ supergravity,[28] where the dynamical generation of composite $SU(8)$ gauge bosons may be possible by the mechanism of Ref. 11. This would greatly improve the phenomenological prospects for this theory as a unified model of electro-weak, strong, and gravitational interactions.

[1]E. M. Purcell and R. V. Pound, Phys. Rev. **81**, 279 (1951).
[2]K. Cahill, Phys. Rev. D **18**, 2930 (1978); B. Julia and J. F. Luciani, Phys. Lett. B **90**, 270 (1980).
[3]A. C. Davis, A. J. Macfarlane, and J. W. van Holten, Nucl. Phys. B **216**, 493 (1983).
[4]A. C. Davis, A. J. Macfarlane, and J. W. van Holten, Nucl. Phys. B **232**, 473 (1984); J. W. van Holten, Z. Phys. C **27**, 57 (1985).
[5]Y. Cohen and E. Rabinovici, Phys. Lett. B **124**, 371 (1983); Y. Ohnuki, preprint DPNU-86-21.
[6]G. Velo and J. Wess, Nuovo Cimento **1**, 177 (1971).
[7]M. Gomes and Y. K. Ha, Phys. Lett. B **145**, 235 (1984); Y. K. Ha B **256**, 687 (1985).
[8]N. Ohta, Phys. Lett. B **134**, 75 (1984).
[9]A. C. Davis, A. J. Macfarlane and J. W. van Holten, Phys. Lett. B **125**, 151 (1983).
[10]D. J. Amit and A. C. Davis, Nucl. Phys. B **225** (FS9), 221 (1983).
[11]J. W. van Holten, Nucl. Phys. B **242**, 307 (1984); Phys. Lett. B **135**, 427 (1984).
[12]A. C. Davis, M. D. Freeman, and A. J. Macfarlane, Nucl. Phys. B **258**, 393 (1985); T. Morozumi and S. Nojiri, Prog. Theor. Phys. **75** (1985).
[13]E. Cremmer, S. Ferrara, and J. Scherk, Phys. Lett. B **74**, 61 (1978); E. Cremmer and B. Julia, Nucl. Phys. B **159**, 141 (1979).
[14]Y. Nambu, *Proceedings of the International Conference on Symmetries and Quark Models* (Wayne State Univ., Detroit, MI, 1969); H. B. Nielsen, *15th Conference on High Energy Physics* (Naukova Dumka, Kiev, 1970); L. Susskind, Nuovo Cimento **69**, 457 (1970).
[15]L. Brink, P. Di Vecchia, and P. S. Howe, Phys. Lett. B **65**, 471 (1976); S. Deser and B. Zumino, Phys. Lett. B **65**, 369 (1976).
[16]R. C. Brower, Phys. Rev. D **6**, 1655 (1972); P. Goddard and C. B. Thorn, Phys. Lett. B **40**, 235 (1972).
[17]W. Heisenberg, Z. Phys. **90**, 209 (1934); J. R. Oppenheimer and W. Furry, Phys. Rev. **45**, 245 (1934).
[18]P. A. M. Dirac, Proc. R. Soc. London Ser. A **180**, 1 (1942); W. Pauli, Rev. Mod. Phys. **15**, 175 (1943).
[19]V. Bargmann, Ann. Math. **48**, 568 (1947); A. O. Barut and C. Fronsdal, Proc. R. Soc. London Ser. A: **287**, 532 (1965); W. J. Holman and L. C. Biedenharn, Ann. Phys. (NY) **39**, 1 (1966).

[20]B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974), Chaps. 17 and 18.
[21]H. Arisue, T. Fujiwara, T. Inoue, and K. Ogawa, J. Math. Phys. **22**, 2055 (1981).
[22]S. N. Gupta, Proc. Phys. Soc. **63**, 68 (1950); K. Bleuler, Helv. Phys. Acta **23**, 567 (1950).
[23]E. Witten, Nucl. Phys. B **202**, 253 (1982).
[24]R. Courant and D. Hilbert, *Methoden der Mathematischen Physik* I (Springer, Berlin, 1968), Chap. VII.
[25]S. Coleman, Commun. Math. Phys. **31**, 264 (1973).
[26]I thank T. Morozumi and S. Nojiri for correspondence concerning this point.
[27]M. T. Grisaru, Phys. Lett. B **66**, 75 (1977); M. T. Grisaru and P. van Nieuwenhuizen, *Deeper Pathways in High Energy Physics*, edited by A. Perlmutter and L. F. Scott (Plenum, New York, 1977).
[28]B. de Wit and H. Nicolai, Nucl. Phys. B **208**, 323 (1982).

Definition and general properties of a class of non-Markovian collision processes

M. Moreau
Laboratoire de Physique Theorique des Liquides, Université Pierre et Marie Curie, 4, place Jussieu, 75252-Paris Cedex 05, France

B. Gaveau
UER de Mathématiques, Université Pierre et Marie Curie, 4, place Jussieu, 75252-Paris Cedex 05, France

(Received 16 July 1986; accepted for publication 31 December 1986)

A class of non-Markovian stochastic processes is defined; it generalizes several anterior models, and allows a trajectorial description of collisions in dense fluids when the system loses all memory at each collision. The abstract formalism is studied here: an integral evolution equation is derived, as well as an integrodifferential equation that generalizes the master equation; some asymptotic properties of these processes are established. Applications to specific models will be treated in other articles.

I. INTRODUCTION

Several authors[1,2] have recently developed stochastic models to describe the molecular relaxation and chemical reactions in a fluid. In most of these models the intermolecular potentials are supposed to have a very short range, compared to the average distance between two molecules (long range potentials may be simulated, in many cases, by mean field potential[2,3] and are not considered explicitly here). Then the interaction of a molecule, or of a reacting complex, with the other particles of the fluid may be represented as a collision localized in time and space; between two collisions, the molecule undergoes a determinist evolution under the influence of the external field, or effective external field.

Naturally, an exact probabilistic description of the fluid is obtained formally by applying the Liouville theorem to the complete density function, and writing the well-known hierarchy of equations for the reduced probability densities. However, this method does not yield explicit results easily, except in the case of the Boltzmann approximation, where multiple collisions are neglected.

Another method is to consider the evolution of a test molecule as a perturbation of its deterministic evolution under the action of collisions with other particles, represented as stochastic events; if the distribution laws in time and phase space of these collisions may be determined *a priori* then it is generally not difficult to study the stochastic evolution of the molecule. Although this approach is less satisfying from a theoretical point of view, since it implies a heuristic evaluation of the collision laws, it yields a simple description of the fluid and permits us to integrate experimental data or semiempirical deductions; furthermore, it appears to be very convenient for numerical simulation,[3,4] and has given valuable results in the theory of chemical reactivity in liquid phase. Thus it seems interesting to extend the validity of this method. In the works published on the subject,[3,5] three main assumptions are generally made: the collisions are instantaneous; the collision times of the test molecule are exponentially distributed: the collision rate at any time does not depend on the past collisions, nor on the state of the molecule; and in a collision the transition probabilities only depend on the initial and final states.

Of course these assumptions are related to the fundamental hypothesis that the evolution of the test molecule may be represented by a Markov process. However, it is clear that they can only be justified approximately and in special conditions; in particular, they are not appropriate in the case of a liquid medium: then, and generally, the evolution of the molecule is non-Markovian and cannot be treated by a standard formalism.

The difficulty is to define non-Markovian processes which are powerful enough to represent the physical phenomena, at least schematically, and not too general in order to be of practical use. For this purpose it is often assumed that the processes obey a generalized Langevin equation,[6,7] as it may be shown in the linear response approximation: but the memory function is hardly obtained from the first principles, and it must generally be approximated by a truncated continued fraction expansion, where the few generalized friction coefficients are chosen in order to fit experimental data, which is not very satisfying from a conceptual point of view. Among the various non-Markovian processes introduced by previous authors one may quote the semi-Markov processes defined by Feller[8] or the generalized random-walk processes used by Kenkre, Montroll, and Shlesinger[9]; the most interesting for physical purposes seems to be the composite stochastic processes studied by van Kampen[10]; however, their definition is not wide enough to be applied to the modelization of chemical reactions. Thus, using the mathematical formalism of stochastic processes, we introduced[11] the so-called collision processes which give a convenient generalization of the anterior works. Here we present this class of processes in a more physical way and study their main properties. Applications to reactional microdynamics are given in another paper.[12]

II. NON-MARKOVIAN COLLISION PROCESSES

A. The model

We consider a system, which may be a test molecule or group of molecules (a reacting complex, for instance) evolving under the action of internal and external forces. We assume that this evolution suffers different phases, labeled by

 0022-2488/87/061430-10$02.50

the index ϵ running from 0 to m, which will be called "regimes"; for instance, regime 0 may be the free evolution of the system, or the evolution in an external field between two collisions; regime 1 may be the evolution during a collision with a particle of kind 1, etc. More generally ϵ could be an element of a countable set Σ (or even of an uncountable set; but we will not consider such a general case in order to avoid technical difficulties which are unnecessary for application to concrete models).

During regime ϵ, the instantaneous state of the system is represented by an element of a probability space E_ϵ which may depend on ϵ. Generally E_ϵ is a n-dimensional Euclidean space R^n or a discrete set, or a manifold, and its dimension or its structure may change with ϵ; then the system is described at time t by a n-dimensional random vector X, and its probability law is defined by its probability density $p_\epsilon(x,t)$, $x \in R^n$, or more generally by the probability measure $P_\epsilon(A,t)$ which gives the probability to find the system at time t in some subset A of E_ϵ. We have given in Ref. 12 three examples modelizing chemical reactions in the presence of a solvent. In one of these examples, the state space E_ϵ can be either one point space (for example a bound state), or a spatial interval with two velocities to take into account the diffusion on the top of a barrier potential.

Whatever may be the physical interpretation of regimes and state spaces, our fundamental hypothesis will be that *the process keeps no memory of the events prior to the beginning of a regime*, specified by its initial time and initial state. This is of course a kind of partial Markov property and does not correspond to general non-Markov processes; however, it allows for nonexponential waiting times and is sufficient in many problems. The same hypothesis was made by van Kampen in defining the composite stochastic process,[10] with more restrictive conditions. Now it is necessary to specify the description of this class of processes.

B. Evolution during a regime

Under regime ϵ the evolution of the system is a Markov process with an infinitesimal operator $L_\epsilon(x)$,

$$\frac{\partial}{\partial t} p_\epsilon(x,t) = L_\epsilon(x)\, p_\epsilon(x,t)\,, \tag{1}$$

where $L_\epsilon(x)$ is a linear operator acting on the x variables only, satisfying the condition

$$\int dx\, L_\epsilon(x)\, p_\epsilon(x,t) = 0\,, \tag{2}$$

which can be written in the integral form

$$L_\epsilon(x)\, p_\epsilon(x,t) = \int dx'\, K_\epsilon(x|x';t)\, p_\epsilon(x',t)\,, \tag{3}$$

with

$$\int dx\, K_\epsilon(x|x';t) = 0\,. \tag{4}$$

For a stationary continuous process with transition rate $W_\epsilon(x'|x)$ from state x to state x',

$$L_\epsilon(x)\, p_\epsilon(x,t) = \int dx'\, W_\epsilon(x|x')\, p_\epsilon(x',t) - p_\epsilon(x,t) \int dx'\, W_\epsilon(x'|x)\,, \tag{5}$$

which can be written in the integral form (3) with the kernel

$$K_\epsilon(x|x') = W_\epsilon(x|x') - \delta(x-x') \int dx''\, W_\epsilon(x''|x')\,. \tag{6}$$

In the case of a deterministic process with conjugated coordinates $x = \{q_i, p_i\}$ and Hamiltonian $H_\epsilon(x)$, $L_\epsilon(x)$ is of course the Lagrangian operator

$$L_\epsilon(x)\, p_\epsilon(x,t) = \sum_i \left(\frac{\partial H_\epsilon}{\partial q_i} \frac{\partial}{\partial p_i} - \frac{\partial H_\epsilon}{\partial p_i} \frac{\partial}{\partial q_i} \right) p_\epsilon(x,t) \tag{7}$$

corresponding to the integral kernel,

$$K_\epsilon(x|x') = \sum_i \left(\frac{\partial H_\epsilon}{\partial p_i} \delta(p_i - p_i') \frac{\partial}{\partial q_i'} \delta(q_i - q_i') - \frac{\partial H_\epsilon}{\partial q_i} \delta(q_i - q_i') \frac{\partial}{\partial p_i'} \delta(p_i - p_i') \right). \tag{8}$$

C. Evolution of the regime ϵ

This evolution contains all the (possibly) non-Markovian behavior of the system. It is well known[12] that when the successive regimes $\epsilon_0, \epsilon_1, \ldots$ constitute a Markov chain with an exponentially distributed pausing time between two changes, the process $\epsilon(t)$ is Markovian, since for $t > t_0$, $\epsilon(t)$ is independent of the events preceding t once $\epsilon(t_0)$ is known. Conformally with the fundamental hypothesis, we assume that once the state x_0 is known at the beginning t_0 of a regime ϵ_0, the process $\{\epsilon(t), X(t)\}$ at further times $t > t_0$ is independent of the events preceding t_0: this is not true if t_0 is not the beginning of the regime ϵ_0 (unless the pausing time in ϵ_0 is exponentially distributed) so that the process is generally non-Markovian.

In our formalism the changes of regime are instantaneous; let $C(\epsilon_0)$ be the first regime following ϵ_0, and $T_{\epsilon_1\epsilon_0}$ the duration of regime ϵ_0 when $C(\epsilon_0) = \epsilon_1$: if ϵ_0 begins at time t_0 and is changed into ϵ_1 at time t_1, then $T_{\epsilon_1\epsilon_0} = t_1 - t_0$; then a stochastic variable $T_{\epsilon\epsilon_0}$ may be defined for any regime ϵ by taking $T_{\epsilon\epsilon_0} = \infty$ if $\epsilon \neq C(\epsilon_0) = \epsilon_1$. We leave the possibility of having $C(\epsilon_0) = \epsilon_0$, since the transition $\epsilon_0 \to \epsilon_0$ may be defined (for instance it may account for instantaneous elastic collisions); in other cases we take $T_{\epsilon_0\epsilon_0} = \infty$. Since the evolution of $X(t)$ under regime ϵ_0 is specified by the infinitesimal operator L_{ϵ_0}, the whole process $\{\epsilon(t), X(t)\}$ is completely determined after the beginning t_0 of regime ϵ_0 by giving the laws of $T_{\epsilon\epsilon_0}$ for any ϵ, and the law of change of the state in the transition $\epsilon_0 \to \epsilon_1$. Of course, these laws are conditioned by the initial values $\epsilon(t_0) = \epsilon_0, x(t_0) = x_0$, and can be interdependent.

We now assume the following.

(i) The laws of $T_{\epsilon\epsilon_0}$ and $C(\epsilon_0)$ only depend on ϵ, on ϵ_0, and on the invariants of the evolution under the action of $L_{\epsilon 0}$ [the fact that these laws depend only on these quantities will

be explicitly used in the derivation of (Eq. 29)]. These laws are specified by

$$\Lambda_{\epsilon_0}(t\,|x_0,t_0) = \mathrm{Prob}(\inf_{\epsilon} T_{\epsilon\epsilon_0} > t - t_0 | x_0,t_0,\epsilon_0) \tag{9}$$

(which is the conditional probability of having no transition in the time interval $]t_0,t]$), and by

$$\Gamma_{\epsilon_1\epsilon_0}(\,]t,t'\,]\,|x_0,t_0) = \mathrm{Prob}(T_{\epsilon_1\epsilon_0} = t_1 - t_0 \in]t-t_0, t'-t_0],\ C(\epsilon_0) = \epsilon_1 | x_0,t_0,\epsilon_0) \tag{10}$$

(which is the conditional probability that the first transition after t_0 occurs in the time internal $]t,t']$ $(t_0 \leqslant t < t')$ and yields the regime ϵ_1). We shall also use the corresponding density $\gamma_{\epsilon_1\epsilon_0}(t\,|x_0,t_0)$,

$$\Gamma_{\epsilon_1\epsilon_0}(dt\,|x_0,t_0) \equiv \Gamma_{\epsilon_1\epsilon_0}(\,]t,t+dt\,]\,|x_0,t) = \gamma_{\epsilon_1\epsilon_0}(t\,|x_0,t_0)dt\,. \tag{10'}$$

(ii) If the state of the system is $x_1^- \in E_{\epsilon_0}$ at the end t_1 of regime ϵ_0 $[X(t_1-0) = x_1^-]$, the state $X(t_1)$ at the beginning of the next regime ϵ_1 is an element of E_{ϵ_1} which is independent of all other events before t_1; thus the effect of the transition is specified by the transition probability

$$Y_{\epsilon_1\epsilon_0}(A\,|x_1^-,t_1) = \mathrm{Prob}\,(X(t_1) = x_1 \in A\,|X(t_1-0) = x_1^-;\ \text{a transition } \epsilon_0 \to \epsilon_1 \text{ occurs at } t_1) \tag{11}$$

(where A is a subset of the phase space E_{ϵ_1}) or by the corresponding density $y_{\epsilon_1\epsilon_0}(x_1|x_1^-,t_1)$,

$$Y_{\epsilon_1\epsilon_0}(dx_1|x_1^-,t_1) = y_{\epsilon_1\epsilon_0}(x_1|x_1^-,t_1)dx_1\,. \tag{11'}$$

In some cases $X(t_1)$ may be determined unambiguously from x_1^- by a mapping f of E_{ϵ_0} into E_{ϵ_1},

$$X(t_1) = f(x_1^-) \text{ with probability 1;}$$

then $y_{\epsilon_1\epsilon_0}$ is obviously a δ function. In particular E_{ϵ_1} may be identical to E_{ϵ_0}, with $X(t_1) = x_1^-$ if the transition does not change the state of the system.

D. Compound transition probabilities

Our main purpose is to find the conditional probability $P_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)$ that the system be at time t under regime ϵ and in some state x of the subset A of the phase space E_ϵ, knowing that a transition has occurred at time t_0, yielding regime ϵ_0 and state $x_0 \in E_{\epsilon_0}$,

$$P_{\epsilon\epsilon_0}(A,t\,|x_0,t_0) = \mathrm{Prob}(X(t) = x \in A,\ \epsilon(t) = \epsilon | X(t_0) = x_0\,,\ \epsilon(t_0) = \epsilon_0;\ \text{a transition occurs at } t_0)\,. \tag{12}$$

Here t is any time posterior to t_0 and not necessarily a collision time.

This conditional probability may be computed from the following two basic quantities, which we call *compound transition probabilities.*

(a) The first is the conditional probability $\Theta_{\epsilon_0}(A,t\,|x_0,t_0)$ that no transition occurs in the time interval $]t_0,t]$ and that the state $X(t) = x$ at time t belongs to the subset A of the phase space E_{ϵ_0},

$$\Theta_{\epsilon_0}(A,t\,|x_0,t_0) = \mathrm{Prob}(x(t) = x \in A,\ \inf_{\epsilon} T_{\epsilon\epsilon_0} > t - t_0 | x_0,t_0,\epsilon_0)\,. \tag{13}$$

(b) The other is the conditional probability $\Phi_{\epsilon_1\epsilon_0}(A,\,]t,t']\,|x_0,t_0)$ that the first transition after t_0 occurs at some epoch t_1 of the time interval $]t,t']$ $(t_0 \leqslant t_1 < t')$, leading to the regime ϵ, and to some state x_1 belonging to the subset A of the phase space E_{ϵ_1},

$$\Phi_{\epsilon_1\epsilon_0}(A,\,]t,t']\,|x_0,t_0) = \mathrm{Prob}(X(t_1) = x_1 \in A\,;\ T_{\epsilon_1\epsilon_0} = t_1 - t_0 \in]t-t_0, t'-t_0];\ C(\epsilon_0) = \epsilon_1 | x_0,t_0,\epsilon_0)\,. \tag{14}$$

Although these two quantities Θ_{ϵ_0} and $\Phi_{\epsilon_1\epsilon_0}$ may possibly be defined in more general situations, we shall restrict ourselves to the particular case described by the hypotheses (i) and (ii) settled in 2.3: then the pausing time in regime ϵ_0 does not depend on the state evolution during ϵ_0, and we have

$$\Theta_{\epsilon_0}(A,t\,|x_0,t_0) = \Lambda_{\epsilon_0}(t\,|x_0,t_0)P_{\epsilon_0}(A,t\,|x_0,t_0)\,, \tag{15}$$

where $\Lambda_{\epsilon_0}(t\,|x_0,t)$, defined by (9), specifies the pausing time in ϵ_0 and $P_{\epsilon_0}(A,t\,|x_0,t_0)$ is the transition probability from state x_0 at time t_0 under regime ϵ_0. Furthermore,

$$\Phi_{\epsilon_1\epsilon_0}(A,dt_1|x_0,t_0) = \int Y_{\epsilon_1\epsilon_0}(A\,|x_1^-,t_1)\Gamma(dt_1|x_0,t_0)P_{\epsilon_0}(dx_1^-,t_1|x_0,t_0)\,, \tag{16}$$

where $\Phi_{\epsilon_1\epsilon_0}$ and $Y_{\epsilon_1\epsilon_0}$ are defined by (10) and (11).

We shall also use the densities θ_{ϵ_0} and $\varphi_{\epsilon_1\epsilon_0}$ corresponding to Θ_{ϵ_0} and $\Phi_{\epsilon_1\epsilon_0}$ when they exist,

$$\Theta_{\epsilon_0}(x,t\,|x_0,t_0) = \Lambda_{\epsilon_0}(t\,|x_0,t_0)\,p_{\epsilon_0}(x,t\,|x_0,t_0)\,, \tag{17}$$

$$\varphi_{\epsilon_1\epsilon_0}(x_1,t_1|x_0,t_0) = \int dx_1^-\, y_{\epsilon_1\epsilon_0}(x_1|x_1^-,t_1)\gamma_{\epsilon_1\epsilon_0}(t_1|x_0,t_0) \times p_{\epsilon_0}(x_1^-,t_1|x_0,t_0)\,. \tag{18}$$

Clearly the quantities Θ and Φ are not independent. As a matter of fact it results from definitions (13) and (14) that

$$\Theta_{\epsilon_0}(E_{\epsilon_0},t\,|x_0,t_0) + \sum_{\epsilon_1} \Phi_{\epsilon_1\epsilon_0}(E_{\epsilon_1},\,]t_0,t\,]\,|x_0,t_0) = 1\,, \tag{19}$$

since the second term on the left-hand side is the conditional probability that a first transition after t_0 occurs before or at t. With the expressions (15) and (16) for Θ and Φ, this relation becomes

$$\Lambda_{\epsilon_0}(t\,|x_0,t_0) + \sum_{\epsilon_1} \Gamma_{\epsilon_1\epsilon_0}(\,]t_0,t\,]\,|x_0,t_0) = 1 \tag{19'}$$

with the same interpretation.

Clearly $\Theta(E_{\epsilon_0},t\,|x_0,t_0)$ or $\Lambda_{\epsilon_0}(t\,|x_0,t_0)$ cannot increase with t. From now on we shall assume that these quantities tend to 0 when $t \to \infty$, that is to say, *for any initial regime and state, at least one transition will almost surely occur;* in this case

$$\Theta_{\epsilon_0}(E,\infty\,|x_0,t_0) = \Lambda_{\epsilon_0}(\,\infty\,|x_0,t_0) = 0\,. \tag{20}$$

Although the statement of the previous definitions and hypotheses has been somewhat lengthy, it is now easy to obtain

the evolution equation of the conditional probability $P_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)$.

III. EVOLUTION EQUATIONS OF THE SYSTEM

A. Fundamental evolution equation

We want to obtain an evolution equation for the probabilities $P_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)$. We divide the set of trajectories starting from x_0 in regime ϵ_0 at time t_0 (just after a collision has occurred) and arriving in the subset A of E_ϵ in regime ϵ at time t, into two subclasses.

(1) The first subset is the set of trajectories that do not suffer any collision and thus do not change the regime in the time interval $]t_0,t]$ so that in particular $\epsilon(t)=\epsilon_0$; this means that $\mathrm{Inf}_\epsilon\, T_{\epsilon\epsilon_0} > t - t_0$ and by Ref. 9, this subset of trajectories has the probability

$$\Theta_{\epsilon_0}(A,t\,|x_0,t_0)\delta_{\epsilon\epsilon_0}\,. \tag{21}$$

(2) The second subset is the set of trajectories suffering at least one collision in the time interval $]t_0,t]$; if $t_1\in]t_0,t]$ denotes the first collision time occurring in this interval, then the probability that the trajectories change their regime from ϵ_0 to ϵ_1 at a first collision in the time interval $]t_1,t_1+dt_1]$, and are found in the volume element dx_1 in E_{ϵ_1} just after the collision, is $\Phi_{\epsilon_1\epsilon_0}(dx_1,dt_1|x_0,t_0)$. After that collision, the process follows independently and has probability $P_{\epsilon\epsilon_1}(A,t\,|x_1,t_1)$ to go from (x_1,t_1) to (A,t) (because, precisely, t_1 is a collision time). The probability of this subset is then the sum over all these possibilities,

$$\int_{t_1\in]t_0,t]}\sum_{\epsilon_1}\int_{x_1^-\in E_{\epsilon_1}} P_{\epsilon\epsilon_1}(A,t\,|x_1,t_1)\Phi_{\epsilon_1\epsilon_0}(dx_1,dt_1|x_0,t_0)\,. \tag{21'}$$

Thus P satisfies the integral equation obtained by adding the contribution (21) and (21'),

$$\begin{aligned}P_{\epsilon\epsilon_0}(A,t\,|x_0,t_0) &= \Theta_{\epsilon_0}(A,t\,|x_0,t_0)\delta_{\epsilon\epsilon_0}\\ &\quad+\sum_{\epsilon_1}\int_{t_1\in]t_0,t]}\int_{x_1\in E_{\epsilon_1}} P_{\epsilon\epsilon_1}(A,t\,|x_1,t_1)\\ &\quad\times\Phi_{\epsilon_1\epsilon_0}(dx_1,dt_1|x_0,t_0)\,,\end{aligned} \tag{22}$$

and the conditional density p satisfies

$$\begin{aligned}&p_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)\\ &\quad=\theta_{\epsilon_0}(x,t\,|x_0,t_0)\delta_{\epsilon\epsilon_0}+\sum_{\epsilon_1}\int_{t_0}^{t}dt_1\int dx_1\\ &\qquad\times p_{\epsilon\epsilon_1}(x,t\,|x_1,t_1)\,\varphi_{\epsilon_1\epsilon_0}(x_1,t_1|x_0,t_0)\,.\end{aligned} \tag{23}$$

Equations (22) and (23) are the fundamental evolution equations under their integral form. It is seen that they are "backward" equations, since the operator of the right-hand side acts on the initial values ϵ_0, x_0, t_0 only.

Remark: This is a generalized version of a renewal equation [see (18)] but with memory and nonidentical laws.

B. Integrodifferential evolution equation

In order to compare the evolution equation of our processes with the usual master equation of Markov processes, we derive Eq. (23) with respect to the initial time t_0, and obtain

$$\begin{aligned}&\frac{\partial}{\partial t_0}p_{\epsilon\epsilon_0}(x,t\,|x_0,t_0)\\ &\quad=\frac{\partial}{\partial t_0}\theta_{\epsilon_0}(x,t\,|x_0,t_0)\delta_{\epsilon\epsilon_0}+\sum_{\epsilon_1}\int_{t_0}^{t}dt_1\int dx_1\\ &\qquad\times p_{\epsilon\epsilon_1}(x,t\,|x_1,t_1)\,\frac{\partial}{\partial t_0}\varphi_{\epsilon_1\epsilon_0}(x_1,t_1|x_0,t_0)\\ &\qquad-\sum_{\epsilon_1}\int dx_1\,p_{\epsilon\epsilon_1}(x,t\,|x_1,t_0)\varphi_{\epsilon_1\epsilon_0}(x_1,t_0|x_0,t_0)\,.\end{aligned} \tag{24}$$

Now we use the expressions (17) and (18) of θ and φ,

$$\theta_{\epsilon_0}(x,t\,|x_0,t_0)=\Lambda_{\epsilon_0}(t\,|x_0,t_0)\,p_{\epsilon_0}(x,t\,|x_0,t_0)\,, \tag{25}$$

$$\begin{aligned}\varphi_{\epsilon_1\epsilon_0}(x_1,t_1|x_0,t_0) &= \int dx_1^-\,y_{\epsilon_1\epsilon_0}(x_1|x_1^-,t_1)\delta_{\epsilon_1\epsilon_0}\\ &\quad\times(t_1|x_0,t_0)\,p_{\epsilon_0}(x_1^-,t_1|x_0,t_0)\,.\end{aligned} \tag{26}$$

We notice that under regime ϵ_0, the evolution equation (1) implies the forward equation,

$$\frac{\partial}{\partial t}p_{\epsilon_0}(x,t\,|x_0,t_0)=L_{\epsilon_0}(x)\,p_{\epsilon_0}(x,t\,|x_0,t_0)\,, \tag{27}$$

whereas the backward equation (which is more convenient mathematically, see Ref. 13) reads

$$\frac{\partial}{\partial t_0}p_{\epsilon_0}(x,t\,|x_0,t_0)=p_{\epsilon_0}(x,t\,|x_0,t_0)\overline{L}_{\epsilon_0}(x_0)\,, \tag{27'}$$

where $\overline{L}_{\epsilon_0}(x_0)$ operates on the left, on variable x_0; $\overline{L}_{\epsilon_0}$ is the opposite of the adjoint $L_{\epsilon_0}^*$ of L_{ϵ_0},

$$\overline{L}_{\epsilon_0}=-L_{\epsilon_0}^*\,. \tag{28}$$

Since $\Lambda_0(t\,|x_0,t_0)$ (the probability that no transition occurs before t) and $\gamma_{\epsilon_1\epsilon_0}(t_1|x_0,t_0)dt_1$ (the probability that the first transition occurs in dt_1 and leads to ϵ_1) only depend on x_0 through the evolution constants, we have, for instance,

$$(\theta_{\epsilon_0}p_{\epsilon_0})\overline{L}_{\epsilon_0}=\theta_{\epsilon_0}(p_{\epsilon_0}\overline{L}_{\epsilon_0})\,.$$

Then Eq. (24) becomes

$$\begin{aligned}\frac{\partial}{\partial t_0}p_{\epsilon\epsilon_0}(x,t\,|x_0,t_0) &= \Lambda_{\epsilon_0}(t\,|x_0,t_0)\,p_{\epsilon_0}(x,t\,|x_0,t_0)\overline{L}_{\epsilon_0}(x_0)\delta_{\epsilon\epsilon_0}+\frac{\partial}{\partial t_0}\Lambda_{\epsilon_0}(t\,|x_0,t_0)\,p_{\epsilon_0}(x,t\,|x_0,t_0)\delta_{\epsilon\epsilon_0}\\ &\quad+\left\{\sum_{\epsilon_1}\int_{t_0}^{t}dt_1\int dx_1\,dx_1^-\,p_{\epsilon\epsilon_1}(x,t\,|x_1,t_1)\,y_{\epsilon_1\epsilon_0}(x_1|x_1^-,t_1)\delta_{\epsilon_1\epsilon_0}(t_1|x_0,t_0)\,p_{\epsilon_0}(x_1^-,t_1|x_0t_0)\right\}\overline{L}_{\epsilon_0}(x_0)\\ &\quad+\sum_{\epsilon_1}\int_{t_0}^{t}dt_1\int dx_1\,dx_1^-\,p_{\epsilon\epsilon_1}(x,t\,|x_1,t_1)\,y_{\epsilon_1\epsilon_0}(x_1|x_1^-,t_1)\,\frac{\partial}{\partial t_0}\gamma_{\epsilon_1\epsilon_0}(t_1|x_0,t_0)\,p_{\epsilon_0}(x_1^-,t_1|x_0,t_0)\\ &\quad-\sum_{\epsilon_1}\int dx_1\,p_{\epsilon\epsilon_1}(x,t\,|x_1,t_1)\,\varphi_{\epsilon_1\epsilon_0}(x_1,t_0|x_0,t_0)\,.\end{aligned} \tag{29}$$

Using the integral equation (23), it is seen that the first and third term on the right-hand side sum up to give $p_{\epsilon\epsilon_0}\overline{L}_{\epsilon_0}$. Furthermore, the second term may be expressed in function of $p_{\epsilon\epsilon_0}$ by using (23) again, if $\Lambda_{\epsilon_0}^{-1}(\partial/\partial t)\Lambda_{\epsilon_0}$ is finite. Then (29) can be written

$$\frac{\partial}{\partial t_0}p_{\epsilon\epsilon_0}(x,t\,|x_0,t_0) = p_{\epsilon\epsilon_0}(x,t\,|x_0,t_0)\overline{L}_{\epsilon_0}(x_0) - p_{\epsilon\epsilon_0}(x,t\,|x_0,t_0)\Lambda_{\epsilon_0}^{-1}\frac{\partial}{\partial t_0}\Lambda_{\epsilon_0}(x,t\,|x_0,t_0)$$
$$-\sum_{\epsilon_1}\int dx_1\, p_{\epsilon\epsilon_1}(x,t\,|x_1,t_0)\, y_{\epsilon_1\epsilon_0}(x_1|x_0,t_0)\gamma_{\epsilon_1\epsilon_0}(t_0|x_0,t_0) + \sum_{\epsilon_1}\int_{t_0}^{t} dt_1\int dx_1\, dx_1^-\, p_{\epsilon\epsilon_1}(x,t\,|x_1,t_1)$$
$$\times\, y_{\epsilon_1\epsilon_0}(x_1|x_1^-,t_1)N_{\epsilon_1\epsilon_0}(t,t_1|x_0,t_0)\, p_{\epsilon_0}(x_1^-,t_1|x_0,t_0)\,, \tag{30}$$

with

$$N_{\epsilon_1\epsilon_0}(t,t_1|x_0,t_0) = \frac{\partial}{\partial t_0}\gamma_{\epsilon_1\epsilon_0}(t_1|x_0,t_0) - \gamma_{\epsilon_1\epsilon_0}(t_1|x_0,t_0)\Lambda_{\epsilon_0}^{-1}$$
$$\times\frac{\partial}{\partial t_0}\Lambda_{\epsilon_0}(t\,|x_0,t_0)\,. \tag{31}$$

This is the generalized evolution equation of the process. It is not a master equation in the usual sense since the $p_{\epsilon\epsilon_0}(x,t\,|x_0,t_0)$ are not transition probabilities (remember that t_0 is a time of "collision," or of a change of regime, and not any time). Only the fourth term on the right-hand side of (30) contains a memory kernel, because of the memory factor $N_{\epsilon_1,\epsilon_0}$. This memory vanishes if

$$-\gamma_{\epsilon_1\epsilon_0}^{-1}\frac{\partial}{\partial t_0}\gamma_{\epsilon_1\epsilon_0}(t_1|x_0,t_0)$$
$$= -\Lambda_{\epsilon_0}^{-1}\frac{\partial}{\partial t_0}\Lambda_{\epsilon_0}(t\,|x_0,t_0) \equiv \lambda_{\epsilon_0}(x_0,t_0)\,, \tag{32}$$

which implies that this common positive value $\lambda_{\epsilon_0}(x_0,t_0)$ is independent of ϵ_1, t_1, and t. Then we see from (32) that

$$\Lambda_{\epsilon_0}(t\,|x_0,t_0) = \exp-\int_{t_0}^{t} dt'\,\lambda_{\epsilon_0}(x_0,t')\,, \tag{33}$$

$$\gamma_{\epsilon_1\epsilon_0}(t_1|x_0,t_0) = \gamma_{\epsilon_1\epsilon_0}(t_1|x_0,t_1)\exp-\int_{t_0}^{t_1} dt'\,\lambda_{\epsilon_0}(x_0,t')\,. \tag{33'}$$

It results from these relations that Eq. (30) reduces to an ordinary evolution equation without memory if

$$\Lambda_{\epsilon_0}^{-1}(t\,|x_0,t_0)\frac{\partial}{\partial t_0}\Lambda_{\epsilon_0}(t\,|x_0,t_0) = -\lambda_{\epsilon_0}(x_0,t)\,, \tag{34}$$

$$\Lambda_{\epsilon_0}^{-1}(t\,|x_0,t_0)\,\gamma_{\epsilon_1\epsilon_0}(t\,|x_0,t_0) = \gamma_{\epsilon_1\epsilon_0}(t\,|x_0,t)\,. \tag{34'}$$

These quantities are, respectively, the conditional probability rate of having any transition at time t, and of having a transition towards regime ϵ_1 at time t, knowing that there has been no transition from t_0 to t. Equation (34) and (34') show that if the memory vanishes they may depend on t but not on the time t_0 of the previous transition. These conditions are necessary to obtain a Markov process.

Remark: In the case of Markov processes, the transition probabilities specify the Markov process in an unique way (i.e., up to an isomorphism of probability spaces). In our case, we can study by the generalized evolution equation only some special transition probabilities. In general, they will not be sufficient to determine the process completely, but for physical applications they will give information on the two-time correlation functions, the equilibrium value distribution, and the rate constants (see Ref. 12 and subsequent publications).

On the other hand, we have described in Ref. 11 the trajectories of our processes; this work stresses the analytical aspect and we refer to[11] for a detailed description of the trajectories.

C. Time-homogeneous processes

It is conceptually interesting to distinguish the initial and final times in the conditional probabilities, as has been done previously, in order to point out the backward nature of the equations. However, many of the physical processes are time homogeneous: the waiting time $T_{\epsilon_1\epsilon_0}$ of a transition $\epsilon_0\to\epsilon_1$ does not depend on the beginning t_0 of regime ϵ_0, and all two-time quantities only depend on the difference between these times; then the quantities λ_{ϵ_0} and $\gamma_{\epsilon_1\epsilon_0}$ defined by (34) and (34') are independent of t, as well as the transition rates $y_{\epsilon_1\epsilon_0}(x_1|x_1^-)$. In this case we write

$$p_{\epsilon\epsilon_0}(x,t\,|x_0,t_0) = p_{\epsilon\epsilon_0}(x,t-t_0|x_0)\,,$$

and the generalized master equation becomes

$$\frac{\partial}{\partial t}p_{\epsilon\epsilon_0}(x,t\,|x_0)$$
$$= -p_{\epsilon\epsilon_0}(x,t\,|x_0)\overline{L}_{\epsilon_0}(x_0) - p_{\epsilon\epsilon_0}(x,t\,|x_0)\Lambda_{\epsilon_0}^{-1}\frac{\partial\Lambda}{\partial t}(x,t\,|x_0) + \sum_{\epsilon_1}\int dx_1\, p_{\epsilon\epsilon_1}(x,t\,|x_1)y_{\epsilon_1\epsilon_0}(x_1|x_0)\gamma_{\epsilon_1\epsilon_0}(x_0)$$
$$+\sum_{\epsilon_1}\int_{t_0}^{t} d\tau\int dx_1\, dx_1^-\, p_{\epsilon\epsilon_1}(x,t-\tau|x_1)y_{\epsilon_1\epsilon_0}(x_1|x_1^-)N_{\epsilon_1\epsilon_0}(t,\tau|x_0)p_{\epsilon_0}(x_1^-,\tau|x_0)\,. \tag{35}$$

The memory term $N_{\epsilon_1\epsilon_0}$ is easily deduced from (31),

$$N_{\epsilon_1\epsilon_0}(t,\tau|x_0) = \frac{\partial}{\partial\tau}\gamma_{\epsilon_1\epsilon_0}(\tau|x_0) - \gamma_{\epsilon_1\epsilon_0}(\tau|x_0)\Lambda_{\epsilon_0}^{-1}(t|x_0)\frac{\partial}{\partial t}\Lambda_{\epsilon_0}(t|x_0) . \tag{36}$$

From (33) and (33'), $N_{\epsilon_1\epsilon_0}$ vanishes if

$$\Lambda_{\epsilon_0}(t|x_0) = \exp(-\lambda_{\epsilon_0}(x_0)t) , \tag{37}$$

$$\gamma_{\epsilon_1\epsilon_0}(t|x_0) = \gamma_{\epsilon_1\epsilon_0}(x_0)\exp(-\lambda_{\epsilon_0}(x_0)t) , \tag{37'}$$

thus exponential waiting times are necessary and sufficient conditions to recover a process without memory, as was expected; in this case (35) is exactly the backward master equation.

From now on we will treat time-homogeneous processes, unless otherwise specified.

D. A particular case

It may occur that the conditional probability of having a transition $\epsilon_0 \to \epsilon_1$ between t and $t + dt$, knowing that there is a transition from ϵ_0 in this time interval, is independent of the pausing time in ϵ_0; then one may write

$$\frac{\gamma_{\epsilon_1\epsilon_0}(t|x_0)}{-(\partial/\partial t)\Lambda_{\epsilon_0}(t|x_0)} = a_{\epsilon_1\epsilon_0}(x_0) = a_{\epsilon_1\epsilon_0}(x_1^-) \tag{38}$$

(since the dependence on x_0 is only through constant of motions, which are the same in state x_1^- at the end of regime ϵ_0).

Physically, this property can be true if all transitions have the same cause which determines the end of ϵ_0, the choice of the next regime ϵ_1 being independent of the pausing time in ϵ_0. In this case the generalized master equation may be written in a simple form by defining the matrices

$$\begin{aligned} &\mathbf{p} = (p_{\epsilon\epsilon_0}) , \quad \mathbf{p}^0 = (\delta_{\epsilon\epsilon_0}p_{\epsilon_0}) , \\ &\mathbf{\Lambda} = (\delta_{\epsilon\epsilon_0}\Lambda_{\epsilon_0}) , \quad \boldsymbol{\mu} = (y_{\epsilon\epsilon_0}a_{\epsilon\epsilon_0}) , \\ &\mathbf{N}(t,\tau|x_0) = \frac{\partial^2}{\partial\tau^2}\mathbf{\Lambda}(\tau|x_0) - \frac{\partial}{\partial\tau}\mathbf{\Lambda}(\tau|x_0)\mathbf{\Lambda}^{-1}(t|x_0)\frac{\partial}{\partial t}\mathbf{\Lambda}(\tau|x_0) , \end{aligned} \tag{39}$$

and the (diagonal) matrix of operators,

$$\overline{\mathbf{L}}(x_0) = (\delta_{\epsilon\epsilon_0}\overline{L}_{\epsilon_0}(x_0)) .$$

Then, using matrices products, Eq. (35) becomes

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{p}(x,t|x_0) &= -\mathbf{p}(xt|x_0)\overline{\mathbf{L}}(x_0) + \mathbf{p}(xt|x_0)\mathbf{\Lambda}^{-1}\frac{\partial}{\partial t}\mathbf{\Lambda}(t|x_0) \\ &\quad - \int dx_1\, \mathbf{p}(xt|x_1)\boldsymbol{\mu}(x_1|x_0)\frac{\partial}{\partial t}\mathbf{\Lambda}(0|x_0) \\ &\quad - \int_0^t d\tau \int dx_1\, dx_1^-\, \mathbf{p}(x,t-\tau|x_1) \\ &\quad \times \boldsymbol{\mu}(x_1|x_1^-)\,\mathbf{p}^0(x_1^-,\tau|x_0)\mathbf{N}(t,\tau|x_0) , \end{aligned} \tag{40}$$

which is the final equation given in Ref. 11.

It is clear from (37) and (37') that Markov processes must be of this kind. In this case we can write

$$\Lambda(t|x_0) = \exp(-\lambda t) , \tag{41}$$

and Eq. (40) becomes a standard master equation,

$$\frac{\partial}{\partial t}\mathbf{p}(xt|x_0) = -\mathbf{p}(xt|x_0)\overline{\mathbf{L}}(x_0) + \int dx_1\{\mathbf{p}(xt|x_1) - \mathbf{p}(xt|x_0)\}\boldsymbol{\mu}(x_1|x_0)\lambda . \tag{42}$$

If the Markov process during regime ϵ is a jump process with an exponential pausing time with time constant η_ϵ and transition rate $\nu_\epsilon(x'|x)$ from x to x', it is seen that the total Markov process is a jump process with time constant $\lambda_\epsilon + \eta_\epsilon$ and transition rates from ϵ to ϵ',

$$\frac{\mu_{\epsilon'\epsilon}(x'|x)\lambda_\epsilon + \delta_{\epsilon'\epsilon}\nu_\epsilon(x'|x)\eta_\epsilon}{\lambda_\epsilon + \eta_\epsilon} .$$

E. A remark on non-Markovian regimes

It can be interesting for some applications to study the more general situation where the evolution in the regimes is not Markovian. Then it can be seen that, if quantities θ and ϕ defined by (13) and (14) can be computed, the integral equation (22) applies, but the integrodifferential equations such as (29) obviously do not make sense.

IV. GENERAL PROPERTIES

A. Condensed evolution equation

Much works on non-Markov processes use a generalized Langevin equation: for this reason we have carefully studied how such an equation may be deduced from the fundamental evolution equation (23). However, the general properties of the process are more easily derived from the integral equation (22) for the conditional probability $P_{\epsilon\epsilon_0}(A,t|x_0,t_0)$ to be at time t in regime ϵ and in subset A of the state space, knowing that the system was in regime ϵ_0 and state x_0 at time t_0, just after a transition.

We again use the compound transition probabilities Θ_{ϵ_0} and $\Phi_{\epsilon\epsilon_0}$ defined in Sec. II D with the same notations.

In order to write the evolution equations concisely we introduce the matrices

$$\begin{aligned} &\mathbf{P} = (P_{\epsilon\epsilon_0}) , \\ &\mathbf{\Theta} = (\Theta_{\epsilon\epsilon_0}) \equiv (\delta_{\epsilon\epsilon},\Theta_{\epsilon_0}) , \\ &\boldsymbol{\phi} = (\Phi_{\epsilon\epsilon_0}) , \end{aligned} \tag{43}$$

and we define the * product as the matrical product followed by the integrations on x_1 and on t_1. Then Eq. (22) may be written

$$\mathbf{P}(A,t|x_0,t_0) = (\mathbf{\Theta} + \mathbf{P}*\boldsymbol{\phi})(A,t|x_0,t_0) . \tag{44}$$

Here A is a subset of the Cartesian product E of all the state spaces E_ϵ,

$$E = \prod_\epsilon E_\epsilon .$$

In these condensed notations, relation (19) between Θ and ϕ becomes

$$\mathbf{1} = (\mathbf{U} + \mathbf{1}*\Phi)(E,t\,|x_0,t_0)\,, \tag{45}$$

where $\mathbf{1}$ denotes the $(m+1)$ line vector $(1,1,\ldots,1)$ and $\mathbf{U}$ the $(m+1)$ line vector defined by

$$\mathbf{U} = (\Theta_{\epsilon_0}),\quad \epsilon_0 = 0,1,2,\ldots\,. \tag{46}$$

This condition is clearly necessary in order that the solution of (22) be normalized, since it results from (22) by writing $\mathbf{P}(E,t\,|x_0,t_0) = 1$. We shall see that it is not always sufficient.

Before studying the solution of Eqs. (22) and (23), we notice that since $\mathbf{U}(E,t\,|x_0,t_0)\to 0$ as $t\to\infty$, Eq. (45) shows that Φ is a solution of the equation

$$\mathbf{v} = (\mathbf{v}*\Phi)(E,\infty\,|x_0,t_0)\,, \tag{47}$$

which reads explicitly

$$v_{\epsilon_0}(x_0,t_0) = \sum_{\epsilon_1}\int_{t_1>t_0}\int_{x_1\in E_1} v_{\epsilon_1}(x_1,t_1) \times \Phi_{\epsilon_1\epsilon_0}(dx_1,dt_1\,|x_0,t_0)\,, \tag{48}$$

where $v_{\epsilon_0}(E,\infty\,|x_0,t_0)\equiv v_{\epsilon_0}(x_0,t_0)$ is a function ϵ_0,x_0,t_0.

On the other hand, if we assume that $P_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)$ has a limit $Q_{\epsilon\epsilon_0}(A\,|x_0,t_0) = P_{\epsilon\epsilon_0}(A,\infty\,|x_0,t_0)$ when $t\to\infty$, Eq. (22) shows that the $(p+1)$ vector $\mathbf{Q}_\epsilon = (Q_{\epsilon\epsilon_0})$, $\epsilon_0 = 0,1,2,\ldots$, is a solution of (47) for any ϵ or A. Thus if this equation has a unique solution (up to a multiplicative constant) we conclude that

$$\mathbf{Q}_\epsilon \propto \mathbf{1}\,,$$

which means that $P_{\epsilon\epsilon_0}(A,\infty\,|x_0,t_0)$ do not depend on ϵ_0,x_0,t_0.

The previous assumptions must be verified on particular models; nevertheless it can be said that the asymptotic probability law, if defined, is generally independent of the initial conditions, as could be expected.

B. Formal solution of the evolution equation

Iterating Eq. (22) yields the formal solution of Eqs. (22) and (23),

$$\mathbf{P} = \Theta + \Theta*\Phi + \Theta*\Phi*\Phi + \cdots + \Theta*\Phi^{*n} + \cdots\,, \tag{49}$$

where the right-hand side displays the successive contributions of the evolutions with no change of regime between t_0 and t, with one change, two changes, etc.

This is the equation written by van Kampen for the composite stochastic processes.[10] In Appendix A, it is shown that the formal series that appears in Eq. (49) is indeed convergent.

In most practical cases its sum is normalized (unless an infinite number of transitions can occur during a finite time interval) and then it yields the unique physically significative solution of the evolution equation. However, this solution, under the form (49), is of little practical use; for this reason we now turn to the solution of the evolution equation by Laplace transforms.

V. STUDY BY LAPLACE TRANSFORM

A. General solution

The Laplace transform of a matrix function $\mathbf{f}(x,t\,|x_0)$ will be denoted $\hat{\mathbf{f}}(x,s|x_0)$,

$$\hat{\mathbf{f}}(x,s|x_0) = \int_0^\infty dt\, e^{-st}\mathbf{f}(x,t\,|x_0)\,. \tag{50}$$

We now explicitly restrict to time homogeneous processes, represented by their probability density matrix $\mathbf{p}(x,t\,|x_0,0) \equiv \mathbf{p}(x,t\,|x_0)$. Then the fundamental equation (23) for $\mathbf{p}$ yields

$$\hat{\mathbf{p}}(x,s|x_0) = \hat{\theta}(x,s|x_0) + \int dx_1\, \hat{\mathbf{p}}(x,s|x_1)\hat{\varphi}(x_1,s|x_0)\,, \tag{51}$$

or in condensed notations

$$\hat{\mathbf{p}} = \hat{\theta} + \hat{\mathbf{p}}\hat{\varphi}\,, \tag{52}$$

where the product $\hat{\mathbf{p}}\hat{\varphi}$ implies the matricial product and the summation on the intermediate coordinate x_1.

By iteration it follows from (52) that

$$\hat{\mathbf{p}} = \hat{\theta}(\hat{\mathbf{I}} + \hat{\varphi} + \hat{\varphi}\hat{\varphi} + \cdots + \hat{\varphi}^n + \cdots)\,, \tag{53}$$

where $\hat{\mathbf{I}} = (\hat{I}_{\epsilon\epsilon_0})$ and

$$\hat{I}_{\epsilon\epsilon_0}(x|x_0) = \delta_{\epsilon\epsilon_0}\delta(x-x_0)\,.$$

The results of Sec. IV B show that the series (52) is convergent [or at least, the analogous series for $\hat{P}(A,s|x_0)$, which is obtained by integrating (52) on the subset A of the x space].

If the matrix $\hat{\mathbf{I}}$-$\hat{\varphi}$ is invertible in the sense of the product used in (52), the solution of (52) may also be written

$$\hat{\mathbf{p}} = \hat{\theta}(\hat{\mathbf{I}} - \hat{\varphi})^{-1}\,. \tag{54}$$

In general no condensed expression of $(\hat{\mathbf{I}} - \hat{\varphi})^{-1}$ is known, and the expansion (53) must be used. However, in the next subsection we consider a particular case where this expansion may be avoided.

B. System "without phase space"

Under this name we consider systems the description of which includes no state x (or a single, definite x for each regime ϵ). Then the product used in (52) is only a matricial product and the solution (54) may be easily studied.

1. Solution

We notice that the conservation relation (45),

$$1 = \theta_{\epsilon_0}(t) + \sum_\epsilon \int_0^t dt'\,\varphi_{\epsilon\epsilon_0}(t')$$

implies

$$\hat{\theta} = s^{-1}(\hat{\mathbf{I}} - \hat{\mathbf{F}}(s)) \tag{55}$$

with

$$\hat{F}_{\epsilon\epsilon_0}(s) = \delta_{\epsilon\epsilon_0}\sum_{\epsilon'}\hat{\varphi}_{\epsilon'\epsilon_0}(s)\,. \tag{56}$$

Except for special values of s, the matrices $\mathbf{A}(s) = \hat{\mathbf{I}} - \hat{\varphi}(s)$ and $\hat{\mathbf{I}} - \hat{\mathbf{F}}(s)$ can be inverted, so that the solution (54) is

$$\hat{\mathbf{P}}(s) = \hat{\boldsymbol{\theta}}(s)\hat{\mathbf{A}}(s)$$
$$= s^{-1}(\hat{\mathbf{I}} - \hat{\mathbf{F}}(s))(\hat{\mathbf{I}} - \hat{\boldsymbol{\varphi}}(s)) = s^{-1}\mathbf{B}^{-1}. \quad (57)$$

It is clear from (55) and (56) that the vector $\mathbf{1} = (1,1,...,1)$ is a left eigenvector of $\mathbf{B} = (\hat{\mathbf{I}} - \hat{\boldsymbol{\varphi}})\ (\hat{\mathbf{I}} - \hat{\mathbf{F}})^{-1}$, with eigenvalue 1; thus the same property is true for $\mathbf{B}^{-1}$, and it results from (57) that

$$\sum_{\epsilon} \hat{p}_{\epsilon\epsilon_0}(s) = s^{-1},$$

which implies that the conditional probabilities $p_{\epsilon\epsilon_0}(t)$ are indeed normalized.

2. Stationary probability

The asymptotic value $\mathbf{p}(\infty)$ of $\mathbf{p}(t)$ when $t \to \infty$ is obtained by taking the limit of $s\hat{\mathbf{p}}(s)$ when $s \to 0$: $s\hat{\mathbf{p}}(s) \underset{s\to 0}{\to} \mathbf{p}(\infty)$, and it is shown in Appendix B that

$$p_\epsilon(\infty) = \lim_{s\to 0} s\hat{p}_\epsilon(s) = \frac{\bar{t}_\epsilon q_\epsilon}{\Sigma_{\epsilon'} \bar{t}_{\epsilon'} q_{\epsilon'}}, \quad (58)$$

where

$$\bar{t}_\epsilon = \int_0^\infty dt\, t \sum_{\epsilon'} \varphi_{\epsilon'\epsilon}(t) \quad (59)$$

is the mean waiting time in regime ϵ, and $q = (q_\epsilon)$ is the only right eigenvector of $\hat{A}(0)$ corresponding to the eigenvalue 0.

Thus the stationary probability $p_\epsilon(\infty)$ is independent of the initial regime, as it should be.

3. Relaxation towards the stationary state

If the required conditions are fulfilled[13] the conditional probability density $p(t)$ can be obtained from its Laplace transform $\tilde{\mathbf{p}}(s)$ by using the Laplace inversion theorem,[14]

$$\mathbf{p}(t) = \lim_{s\to 0} s\hat{\mathbf{p}}(s) + \sum_i \operatorname*{res}_{s_i} (e^{st}\hat{\mathbf{p}}(s)), \quad (60)$$

where $\{s_i\}$ denotes the poles of $e^{st}\mathbf{p}(s)$, other than $s = 0$. These poles are the zeros of $\det \mathbf{A}(s) = \det(\mathbf{I} - \boldsymbol{\varphi}(s))$, except in exceptional cases, which will be discarded in this general discussion.

It is shown in Appendix C that $\det A(s)$ has no zero for $R_e s > 0$. Thus the poles of $\tilde{\mathbf{p}}(s)$ have negative real parts, and the corresponding time exponentials of $p(t) - p(\infty)$ are time decreasing, which shows that $p(t)$ indeed tends to $p(\infty)$.

VI. CONCLUSION

It has been seen that the present model of non-Markovian processes leads to a rather simple formalism, which may be treated by generalized master equations, or by integral equations; in some cases it is possible to obtain the most important properties of the process analytically.

Many points remain to be studied, such as the possible approximation schemes [the simplest one being implied by the iterative solution (49)] or, more fundamentally, the connection of the conditional probabilities used here with the theory of fluctuations in thermodynamic equilibrium. We postpone such discussions to further papers. Applications of this formalism are given in other articles.[12]

APPENDIX A: CONVERGENCE OF THE FORMAL SOLUTION OF THE EVOLUTION EQUATION

Using the notation of Sec. IV let us consider the formal solution (49) of the evolution equation (22),

$$\mathbf{P} = \boldsymbol{\Theta} + \boldsymbol{\Theta}*\boldsymbol{\Phi} + \boldsymbol{\Theta}*\boldsymbol{\Phi}*\boldsymbol{\Phi} + \cdots + \boldsymbol{\Theta}*\boldsymbol{\Phi}^{*n} + \cdots. \quad (A1)$$

We shall see that this formal series is indeed convergent.

As a matter of fact let us define the conditional probability $\mathbf{P}^{\bar{n}}$,

$$\mathbf{P}^{\bar{0}} = \boldsymbol{\Theta}, \quad (A2)$$

$$\mathbf{P}^{\bar{n}} = \boldsymbol{\Theta}* \sum_{k=0}^{n} \boldsymbol{\phi}^{*n}. \quad (A3)$$

Here $P^{\bar{n}}_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)$ is the conditional probability of being in regime ϵ and subset A of E_ϵ at time t after at most n transitions since time t_0. Clearly

$$\mathbf{P}^{\overline{n+1}} = \boldsymbol{\Theta} + \mathbf{P}^{\bar{n}}*\boldsymbol{\phi} = \mathbf{P}^{\bar{n}} + \boldsymbol{\Theta}*\boldsymbol{\phi}^{*(n+1)} \quad (A4)$$

and $P^{\bar{n}}_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)$ increases with n.

On the other hand,

$$P^{\bar{n}}_{\epsilon\epsilon_0}(A,t\,|x_0,t_0) \leqslant P^{\bar{n}}_{\epsilon\epsilon_0}(E_\epsilon,t\,|x_0,t_0).$$

But if

$$\sum_{\epsilon} P^{\bar{n}}_{\epsilon\epsilon_0}(E_\epsilon,t\,|x_0,t_0) \leqslant 1, \quad (A5)$$

then by (A4) and (A5),

$$\sum_{\epsilon} P^{\overline{n+1}}_{\epsilon\epsilon_0}(E_\epsilon,t\,|x_0,t_0) \leqslant (\mathbf{U} + \mathbf{1}*\boldsymbol{\Phi})(E_\epsilon,t\,|x_0,t_0) = 1,$$

so that (A5) is satisfied for all n, since it is true for $n = 0$. Thus $P^{\bar{n}}_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)$ converges to a limit as $n \to \infty$,

$$P^{\bar{n}}_{\epsilon\epsilon_0}(A,t\,|x_0,t_0) \to P_{\epsilon\epsilon_0}(A,t\,|x_0,t_0) \leqslant P_{\epsilon\epsilon_0}(E_\epsilon,t\,|x_0,t_0) \leqslant 1 \quad (A6)$$

and by (A4) the limit matrix $\mathbf{P}(A,t\,|x_0,t_0)$ satisfies the fundamental equation (44).

However, it may happen that the limit $\Sigma_\epsilon P_{\epsilon\epsilon_0}(E_\epsilon,t\,|x_0,t_0)$ is inferior to 1. But from definition (A6) it is the probability that a finite number of transitions occurs between t and t', including 0 transition; thus $1 - \Sigma P_{\epsilon\epsilon_0}(E_\epsilon,t\,|x_0,t_0)$ is the probability of having an infinite number of transitions in this time interval: it may differ from 0 in some special cases which will be excluded from this study.

In particular, if $\Theta_{\epsilon_0}(E_{\epsilon_0},t\,|x_0,t_0)$ is superior to some decreasing positive function of t, $1 - a(t)$, independent of the initial conditions,

$$\Theta_{\epsilon_0}(E_{\epsilon_0},t\,|x_0,t_0) \geqslant 1 - a(t) > 0, \quad (A7)$$

then it is easily shown by recurrence that

$$P^{\bar{n}}_{\epsilon_0}(t\,|x_0,t_0) \equiv \sum_{\epsilon} P^{\bar{n}}_{\epsilon\epsilon_0}(E_\epsilon,t\,|x_0,t_0) \geqslant 1 - (a(t))^{n+1}, \quad (A8)$$

which implies

$$P^{\bar{n}}_{\epsilon_0}(t\,|x,t_0) \to 1 \quad \text{as } n \to \infty.$$

Condition (A7) is satisfied if, for instance, the pausing times are exponential,

$$\Theta_{\epsilon_0}(E_{\epsilon_0},t\,|x_0,t_0) \propto \exp(-\lambda_{\epsilon_0}(x_0)(t-t_0))$$

with $\lambda_{\epsilon_0}(x_0)$ inferior to some constant λ independent of ϵ_0 and x_0.

On the contrary (A7) is not true in the case of periodic changes of regime with a definite period τ (which is considered lower) since $\Theta_{\epsilon_0}=0$ for $t>\tau$. But in this case $P^{\bar{n}}_{\epsilon_0}(t\,|x_0,t_0)$, which is the probability of having at most n transitions during $]t_0,t]$, is obviously 1 for $n>t/\tau$, so that P, given by formula (A6), is again normalized.

Finally, it should be observed that if (A6) is normalized it is the only acceptable solution of Eq. (22), since another acceptable solution should satisfy, for any subset A of $E=\Pi_\epsilon E_\epsilon$,

$$Q_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)\geqslant\delta_{\epsilon\epsilon_0}\Theta_{\epsilon_0}(A,t\,|x_0,t_0)\,.$$

In the same way

$$\mathbf{Q}=\mathbf{\Theta}+\mathbf{Q}*\mathbf{\Phi}\geqslant\mathbf{\Theta}+\mathbf{\Theta}*\mathbf{\Phi}=\mathbf{P}^{\bar{2}}\,,$$

and by recurrence

$$\mathbf{Q}\geqslant\mathbf{P}^{\bar{n}}\,,$$

which implies

$$Q_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)\geqslant P_{\epsilon\epsilon_0}(A,t\,|x_0,t_0)\,, \tag{A9}$$

$$\sum_\epsilon Q_{\epsilon\epsilon_0}(E_\epsilon,t\,|x_0,t_0)\geqslant\sum_\epsilon P_{\epsilon\epsilon_0}(E_\epsilon,t\,|x_0,t_0)=1\,. \tag{A10}$$

Thus $\mathbf{Q}$ cannot be normalized to 1, unless the equality holds in (A10), so that it also holds in (A9) for any and any A.

APPENDIX B: STATIONARY SOLUTION OF THE EVOLUTION EQUATION

According to Sec. V B, the conditional probabilities $P_{\epsilon\epsilon_0}(t)$ in systems without phase space are determined by their Laplace transforms under matricial form

$$\hat{\mathbf{P}}(s)=\hat{\boldsymbol{\theta}}(s)(\hat{\mathbf{I}}-\hat{\boldsymbol{\varphi}}(s))^{-1}=s^{-1}\mathbf{B}^{-1}(s)\,, \tag{B1}$$

and the asymptotic value $\mathbf{P}(\infty)$ of $\mathbf{P}(t)$ is the limit of $s\mathbf{P}(s)$ when $s\to 0$.

Although the matrix $\mathbf{A}(s)=\hat{\mathbf{I}}-\hat{\boldsymbol{\varphi}}(s)$ is singular for $s=0$ since

$$\sum_\epsilon\varphi_{\epsilon\epsilon_0}(0)=\sum_\epsilon\int_0^\infty dt\,\varphi_{\epsilon\epsilon_0}(t)=1\,, \tag{B2}$$

$\mathbf{B}^{-1}(s)$ generally has a finite limit as $s\to 0$. As a matter of fact, if $\bar{A}_{\epsilon'\epsilon}(s)$ is the minor of the element $A_{\epsilon'\epsilon}(s)=\delta_{\epsilon'\epsilon}-\hat{\varphi}_{\epsilon'\epsilon}(s)$ of $\mathbf{A}(s)$, the determinant of $\mathbf{A}(s)$ is, for any ϵ,

$$\det\mathbf{A}(s)=\sum_{\epsilon_0}\Big(\sum_{\epsilon'}A_{\epsilon'\epsilon_0}(s)\Big)\bar{A}_{\epsilon\epsilon_0}(s)\,, \tag{B3}$$

whereas

$$A^{-1}_{\epsilon\epsilon_0}(s)=\bar{A}_{\epsilon_0\epsilon}(s)(\det\mathbf{A}(s))^{-1}\,. \tag{B4}$$

Generally $\bar{A}_{\epsilon\epsilon_0}(s)$ tends to a finite limit $A_{\epsilon\epsilon_0}(0)$ when $s\to 0$. Furthermore,

$$s\hat{\theta}_{\epsilon_0}(s)=\sum_\epsilon A_{\epsilon\epsilon_0}(s)=s\bar{t}_{\epsilon_0}+O(s)\,, \tag{B5}$$

where

$$\bar{t}_{\epsilon_0}=\int_0^\infty dt\,t\sum_\epsilon\varphi_{\epsilon\epsilon_0}(t) \tag{B6}$$

is the mean waiting time in regime ϵ_0. Then it results from (B5) that if $s\to 0$,

$$s\hat{P}_{\epsilon\epsilon_0}(s)=s\hat{\theta}_\epsilon(s)A^{-1}_{\epsilon\epsilon_0}(s)\to\frac{\bar{t}_\epsilon\bar{A}_{\epsilon_0\epsilon}(0)}{\Sigma_{\epsilon'}\bar{t}_{\epsilon'}\bar{A}_{\epsilon_0\epsilon'}(0)}=P_{\epsilon\epsilon_0}(\infty)\,. \tag{B7}$$

Now for any ϵ_0,

$$\sum_{\epsilon_1}A_{\epsilon_0\epsilon_1}(0)\bar{A}_{\epsilon_0\epsilon_1}(0)=\det\mathbf{A}(0)=0\,. \tag{B8}$$

But the matrix $\hat{\boldsymbol{\varphi}}(0)$ is a stochastic matrix: its elements are non-negative and by (B2) they add up to one in each column; thus by the theorem of Frobenius one is a simple eigenvalue of $\hat{A}(0)$, which corresponds to a unique right eigenvector $\mathbf{q}=(q_{\epsilon_0})$ with non-negative components which add up to 1: it is the stationary probability of a Markov chain on the regimes ϵ, with transition probability from ϵ_0 to ϵ given by $\hat{\varphi}_{\epsilon\epsilon_0}(0)=\int_0^t dt\,\varphi_{\epsilon\epsilon_0}(t)$. This Markov chain is constructed from the primitive process by deleting the influence of time.

Thus (B8) shows that there exists constant λ_{ϵ_0} such that

$$\bar{A}_{\epsilon_0\epsilon_1}(0)=\lambda_{\epsilon_0}q_{\epsilon_1}$$

and (B7) yields the stationary probability,

$$p_{\epsilon\epsilon_0}(\infty)=\frac{\bar{t}_\epsilon q_\epsilon}{\Sigma_{\epsilon'}\bar{t}_{\epsilon'}q_{\epsilon'}}\,, \tag{B9}$$

which is independent of the initial regime.

APPENDIX C: ASYMPTOTIC BEHAVIOR OF SYSTEMS WITHOUT PHASE SPACE

It is known[15] that if a function $f(t)$ satisfies

$$f(t)\propto e^{-\lambda t}(1+\alpha(t))\,,$$

where $\alpha(t)$ is bounded and tends to 0 as $t\to\infty$, then its Laplace transform $\hat{f}(s)$ admits a pole of first order for $s=-\lambda$, and no other pole in the region $R_e s\geqslant-\lambda$. Thus the asymptotical behavior of the conditional probability $\mathbf{P}(t)$ is found by calculating the poles of $\hat{\mathbf{P}}(s)$, which are the zeros of $\det A(s)$ (excepted in particular cases); but $\det A(s)$ has no zero for $R_e s\geqslant 0$. As a matter of fact,

$$\det\mathbf{A}(s)=\det(\hat{\mathbf{I}}-\hat{\boldsymbol{\phi}}(s))=\prod_i(1-\lambda_i(s))\,, \tag{C1}$$

if $\{\lambda_i(s)\}$ is the spectrum of matrix $\hat{\boldsymbol{\Phi}}(s)$.

We may write

$$\max_i|\lambda_i(s)|\leqslant\max_\epsilon\sum_{\epsilon'}|\hat{\varphi}_{\epsilon'\epsilon}(s)|\leqslant\max_\epsilon\sum_{\epsilon'}\int_0^\infty dt\,e^{-(R_e t)}\varphi_{\epsilon'\epsilon}(t)\,, \tag{C2}$$

and thus, if $\hat{\varphi}_\epsilon(s)=\Sigma_{\epsilon'}\hat{\varphi}_{\epsilon'\epsilon}(s)$,

$$\max_i|\lambda_i(s)|\leqslant\max_\epsilon\hat{\varphi}_\epsilon(R_e s)\,. \tag{C3}$$

(i) When $R_e s > 0$, $\hat{\varphi}_\epsilon(R_e s) < \hat{\varphi}_\epsilon(0) = 1$: by (C1) it is seen that det $\mathbf{A}(s)$ cannot vanish.

(ii) When $R_e s = 0$ and $s = iy \neq 0$, in (C2) the equality holds if and only if the waiting time $T_{\epsilon'\epsilon}$ of a transition $\epsilon \to \epsilon'$ takes discrete values $\tau_{\epsilon'\epsilon} + 2k\pi/y$, $k = 0,1,2,\ldots$. Then it may be shown that det $\mathbf{A}(iy)$ cannot vanish, unless the different regimes are initiated at discrete, periodic times, with period $2\pi/y$. Such a special case, which should be treated as a discrete time process, is out of the scope of the present work.

Proof of (ii): Let $p_{\epsilon'\epsilon}$ be the probability that the first transition from ϵ leads to ϵ',

$$p_{\epsilon\epsilon'} = \int_0^\infty dt\, \varphi_{\epsilon'\epsilon}(t) = \hat{\varphi}_{\epsilon'\epsilon}(0) . \tag{C4}$$

Thus $\varphi_{\epsilon'\epsilon}(t)/p_{\epsilon'\epsilon}$ is a probability density on positive times, and the complex number

$$\int_0^\infty dt\, e^{-iyt} \varphi_{\epsilon'\epsilon}(t)/p_{\epsilon'\epsilon} = z_{\epsilon'\epsilon} \tag{C5}$$

necessarily lies *inside* the complex circle $|z| < 1$, unless the probability density $\varphi_{\epsilon'\epsilon}(t)/p_{\epsilon'\epsilon}$ is concentrated on times for which e^{-iyt} has the same value; this means that there exists a deterministic time $\tau_{\epsilon'\epsilon} > 0$ such that

$$\text{Prob}(e^{-iyT_{\epsilon'\epsilon}} = e^{-iy\tau_{\epsilon'\epsilon}}) = 1 ;$$

in this case the only possible values of $T_{\epsilon'\epsilon}$ are

$$t^k_{\epsilon'\epsilon} = \tau_{\epsilon'\epsilon} + 2k\pi/y, \quad k = 0,1,2,\ldots . \tag{C6}$$

As a result we may write

$$|\hat{\varphi}_{\epsilon'\epsilon}(iy)| = p_{\epsilon'\epsilon}|z_{\epsilon'\epsilon}| \leqslant p_{\epsilon'\epsilon} ,$$

$$\max_\epsilon \sum_{\epsilon'} |\hat{\varphi}_{\epsilon'\epsilon}(iy)| \leqslant 1 ,$$

the equality being possible only if $|z_{\epsilon'\epsilon}| = 1$, which implies (C6) for at least one ϵ, as shown previously.

More precisely it is seen from (C1) that det $\mathbf{A}(iy)$ vanishes only if 1 is an eigenvalue of $\hat{\varphi}(iy)$, which implies the existence of a left eigenvector $\mathbf{u} = \{\tilde{u}_{\epsilon'}\}$ such that

$$u_\epsilon = \sum_{\epsilon'} u_{\epsilon'} \varphi_{\epsilon'\epsilon}(iy) . \tag{C7}$$

Let the maximum value of $|u_\epsilon|$ be realized for some regime ϵ_0; then using definitions (C5) and (C7) for $\epsilon = \epsilon_0$,

$$1 = \sum_{\epsilon'} \frac{u_{\epsilon'}}{u_{\epsilon_0}} z_{\epsilon'\epsilon_0} p_{\epsilon'\epsilon_0} . \tag{C8}$$

Now it should be noticed that, since $|z_{\epsilon'\epsilon_0}| < 1$,

$$|(u_{\epsilon'}/u_{\epsilon_0}) z_{\epsilon'\epsilon_0}| \leqslant 1 .$$

Thus the equality (C8) can hold only if

$$(u_{\epsilon'}/u_{\epsilon_0}) z_{\epsilon'\epsilon_0} = 1 \tag{C9}$$

for every regime ϵ' such that $p_{\epsilon'\epsilon_0} \neq 0$. Equation (C9) implies $|u_{\epsilon'}/u_{\epsilon_0}| = 1$ and $|z_{\epsilon'\epsilon_0}| = 1$, which in turn implies condition (C6), and by (C5)

$$z_{\epsilon'\epsilon_0} = e^{-iy\tau_{\epsilon'\epsilon_0}} . \tag{C10}$$

Defining τ_ϵ up to an additive constant by

$$u_\epsilon = |u_\epsilon| e^{-iy\tau_\epsilon} , \tag{C11}$$

it results from (C9)–(C11) that

$$\tau_{\epsilon'\epsilon_0} = \tau_{\epsilon'} - \tau_{\epsilon_0} + 2k\pi/y . \tag{C12}$$

But (C8) may be used in the same way when ϵ_0 is replaced by any regime ϵ accessible from ϵ_0 ($p_{\epsilon\epsilon_0} \neq 0$) since then $|u_\epsilon| = |u_{\epsilon_0}|$. In the simplest case $p_{\epsilon'\epsilon} \neq 0$ for any two regimes ϵ and ϵ'; as a consequence the waiting time $T_{\epsilon'\epsilon}$ for a first transition $\epsilon \to \epsilon'$ can only take the values

$$\tau_{\epsilon'} - \tau_\epsilon + 2k\pi/y, \quad k = 0,1,2,\ldots,$$

which implies the assertions (ii).

[1] R. Kapral, Adv. Chem. Phys. **48**, 71 (1981).
[2] D. Chandler and L. R. Pratt, J. Chem. Phys. **65**, 2925 (1976).
[3] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).
[4] J. A. Montgomery, Jr., D. Chandler, and B. J. Berne, J. Chem. Phys. **70**, 4056 (1979).
[5] S. Adelman, J. Chem. Phys. **73**, 3145 (1980).
[6] B. Carmeli and A. Nitzan, J. Chem. Phys. **80**, 3596 (1984).
[7] S. H. Northrup and J. T. Hynes, J. Chem. Phys. **71**, 871, 884 (1979).
[8] W. Feller, Proc. Natl. Acad. Sci. USA **51**, 653 (1964).
[9] W. M. Kenkre, E. W. Montroll, and M. F. Shlesinger, J. Stat. Phys. **9**, 45 (1979).
[10] N. G. Van Kampen, Physica A **96**, 435 (1979).
[11] B. Gaveau and M. Moreau, Lett. Math. Phys. **9**, 213 (1985).
[12] D. Borgis, B. Gaveau, and M. Moreau, "A class of collision processes with memory and application to simple chemical reactions in a solvent," to be published in J. Stat. Phys.
[13] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1971), Vol. II.
[14] M. R. Spiegel, *Laplace Transforms*, Schaum series (McGraw-Hill, New York, 1973).
[15] J. L. Doob, *Stochastic Processes* (Wiley, New York, 1967).